master ATiAM

Acoustique
Traitement du signal
Informatique
Appliqués à la
Musique

# A survey of Machine Learning techniques

Arshia Cont
MuTant Team,
IRCAM.
cont@ircam.fr

ATIAM 2013-14

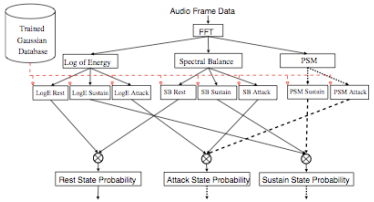
# Plan

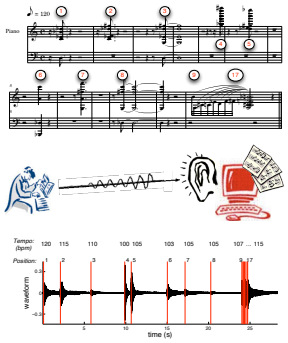
- Last time we saw:
  - Bayesian Decision Theory
  - Maximum Likelihood Parameter Estimations
  - Bayesian Parameter Estimation
- Today, we will look at:
  - Kernel Based Parameter Estimation
  - Mixture Models and EM Algorithm
  - Some Non-parametric methods
  - Sequential Learning
    - HMMs
    - Kalman Filters



# Some musical examples

- The score following problem:
  - Need *observation models* at the front-end.
  - From audio frames to low-level state probs:



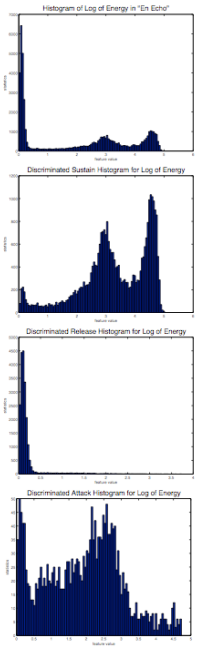



  - This was the model for *suivi* object (now defunct!)
- Problem 1: is how to train the generative probability models which give informative probabilities on Rest/Attack/Sustain

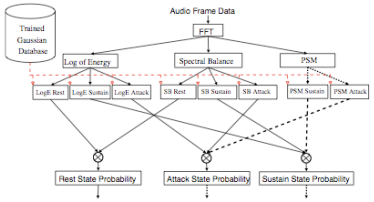


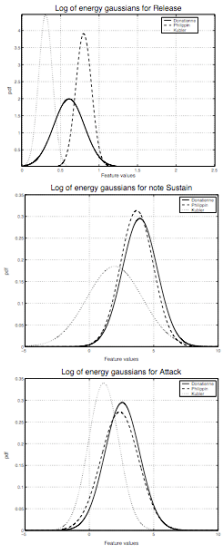
# Example

- Problem 2: These models will be probably different from performance to performance / musician to musician!
- Design cycle: Learning from rehearsal recordings
  - Gather segmented data
  - Design the generative models for each attribute
  - Train models from labeled database
  - Test!
  - Incorporate them in the realtime system for the performance.




Learned Gaussians for different performers>>>




<<<Observation Histograms



# Kernel-based and non-parametric methods



## Plan

- We have already looked at Bayesian Decision Rules, and how to optimize them through Maximum-Likelihood (ML) or Bayesian Parameter Estimation... .
  - In all these formulations, we assume that X is generated by a probability density P(X)
- Practical densities do not approximate well using simple probability density families!
- We now look at ways to approach P(X) when the data is non-trivial or more complicated than a known and simple probability family... .
  - So far, we have considered *parametric* density estimations...
  - Today, we consider *non-parametric* density estimates...



## Non-parametric density estimates

- Given iid training set $\mathcal{D} = \{\mathbf{x}_1, \ldots \mathbf{x}_n\}$, the goal is to estimate

$$P_{\mathbf{X}}(\mathbf{x})$$

- Consider a region $\mathcal{R}$, and define

$$P = P_{\mathbf{X}}[\mathbf{x} \in \mathcal{R}] = \int_{\mathcal{R}} P_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}.$$

and define

$$K = \sharp\{\mathbf{x}_i \in \mathcal{D} | \mathbf{x}_i \in \mathcal{R}\}.$$

- This is a binomial distribution of paramter $P$

$$\begin{aligned} P_K(k) &= \mathcal{B}(n, P) \\ &= \binom{n}{k} P^k (1-P)^{n-k} \end{aligned}$$

# Binomial random variable

- ML estimate of P

$$\hat{P} = \frac{k}{n}.$$

and statistiscs

$$\begin{aligned} E[\hat{P}] &= \frac{1}{n}E[k] = \frac{1}{n}nP = P \\ var[\hat{P}] &= \frac{1}{n^2}var[k] = \frac{P(1-P)}{n}. \end{aligned}$$

- Note that $var[\hat{P}] \leq 1/4n$ goes to zero very quickly, i.e.

$$\hat{P} \to P.$$

| N | 10 | 100 | 1,000 | … |
|---|---|---|---|---|
| Var[P] < | 0.025 | 0.0025 | 0.00025 | |

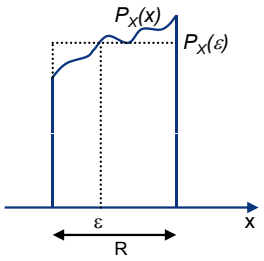


# Histogram

- this means that *k/n* is a very good estimate of *P*
- on the other hand, from the mean value theorem, if $P_X(x)$ is continuous $\exists \epsilon \in \mathcal{R}$ such that

$$P = \int_{\mathcal{R}} P_{\mathbf{X}}(\mathbf{x})d\mathbf{x} = P_{\mathbf{X}}(\epsilon)\int_{\mathcal{R}} d\mathbf{x} = P_{\mathbf{X}}(\epsilon)V(\mathcal{R}).$$

- this is easiest to see in 1D
  - can always find a box such that the integral of the function is equal to that of the box
  - since $P_X(x)$ is continuous there must be a $\varepsilon$ such that $P_X(\varepsilon)$ is the box height






# Histogram

- hence

$$P_{\mathbf{X}}(\epsilon) = \frac{P}{V(\mathcal{R})} \approx \frac{\hat{P}}{V(\mathcal{R})} = \frac{k}{nV(\mathcal{R})}$$

- using continuity of $P_X(x)$ again and assuming *R* is small

$$P_{\mathbf{X}}(\mathbf{x}) \approx \frac{k}{nV(\mathcal{R})}, \ \forall \mathbf{x} \in V(\mathcal{R})$$

- this is the histogram
- it is the simplest possible non-parametric estimator
- can be generalized into kernel-based density estimator



# Kernel density estimates

- assume $\mathcal{R}$ is the $d$-dimensional cube of side $h$

$$V = h^d$$

and define *indicator* function of the unit hypercube

$$\phi(\mathbf{u}) = \begin{cases} 1, & \text{if } |u_i| < 1/2 \\ 0, & \text{otherwise.} \end{cases}$$

hence

$$\phi\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right) = 1$$

iif $\mathbf{x}_i \in$ hypercube of volume $V$ centered at $\mathbf{x}$.

- the number of sample points in the hypercube is

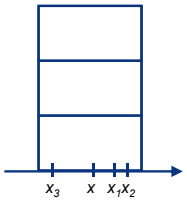
$$k_n = \sum_{i=1}^{n} \phi\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right)$$

# Kernel density estimates

- this means that the histogram can be written as

$$P_{\mathbf{X}}(\mathbf{x}) = \frac{1}{nh^d}\sum_{i=1}^{n}\phi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right)$$

- which is equivalent to:
  - "put a box around $X$ for each $X_i$ that lands on the hypercube"
  - can be seen as a very crude form of interpolation
  - better interpolation if contribution of $X_i$ decreases with distance to $X$
- consider other windows $\phi(x)$






# Windows

- what sort of functions are valid windows?
- note that $P_X(x)$ is a pdf if and only if

$$P_{\mathbf{X}}(\mathbf{x}) \geq 0, \forall \mathbf{x} \text{ and } \int P_{\mathbf{X}}(\mathbf{x})d\mathbf{x} = 1$$

- since

$$\begin{aligned}\int P_{\mathbf{X}}(\mathbf{x})d\mathbf{x} &= \frac{1}{nh^d}\sum_{i=1}^{n}\int\phi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right)d\mathbf{x}\\ &= \frac{1}{nh^d}\sum_{i=1}^{n}\int\phi(\mathbf{y})\,h^d d\mathbf{y}\\ &= \frac{1}{n}\sum_{i=1}^{n}\int\phi(\mathbf{y})\,d\mathbf{y}\end{aligned}$$

- these conditions hold if $\phi(x)$ is itself a pdf

$$\phi(\mathbf{x}) \geq 0, \forall \mathbf{x} \text{ and } \int \phi(\mathbf{x})d\mathbf{x} = 1$$



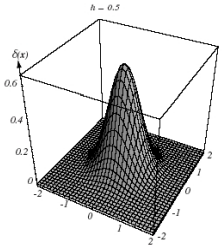
# Gaussian kernel

- probably the most popular in practice

$$\phi(\mathbf{x}) = \frac{1}{\sqrt{2\pi}^d}e^{-\frac{1}{2}\mathbf{x}^T\mathbf{x}}$$

- note that $P_X(x)$ can also be seen as a sum of pdfs centered on the $X_i$ when $\phi(x)$ is symmetric in $X$ and $X_i$

$$P_{\mathbf{X}}(\mathbf{x}) = \frac{1}{nh^d}\sum_{i=1}^{n}\phi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right)$$

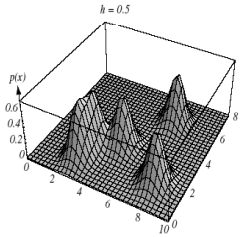




# Gaussian kernel

- Gaussian case can be interpreted as
  - sum of $n$ Gaussians centered at the $X_i$ with covariance $hI$
  - more generally, we can have a full covariance

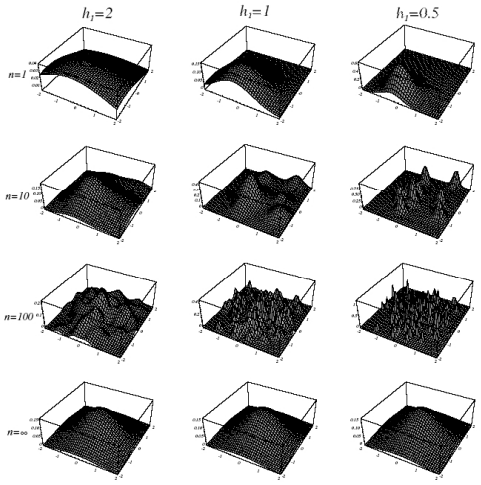
$$P_{\mathbf{X}}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\sqrt{(2\pi)^d|\Sigma|}}e^{-\frac{1}{2}(\mathbf{x}-\mathbf{x}_i)^T\Sigma^{-1}(\mathbf{x}-\mathbf{x}_i)}$$



- sum of $n$ Gaussians centered at the $X_i$ with covariance $\Sigma$
- Gaussian kernel density estimate: *"approximate the pdf of X with a sum of Gaussian bumps"*

# Example

- example: fit to N(0,I) using h = $h_1/n^{1/2}$
- small h: spiky
- need a lot of points to converge (variance)
- large h: approximate N(0,I) with a sum of Gaussians of larger covariance
- will never have zero error (bias)





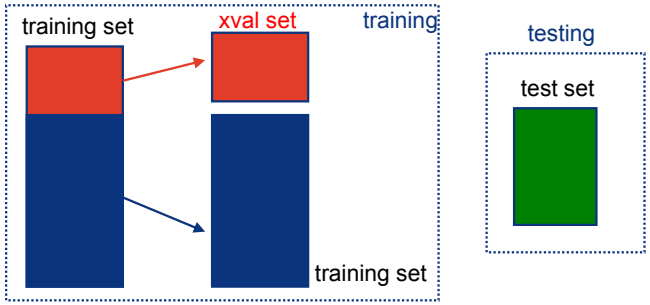

# Optimal bandwidth

- in practice this has limitations
  - does not say anything about the finite data case (the one we care about)
  - still have to find the best k
- usually we end up using trial and error or techniques like cross-validation



# Cross-validation

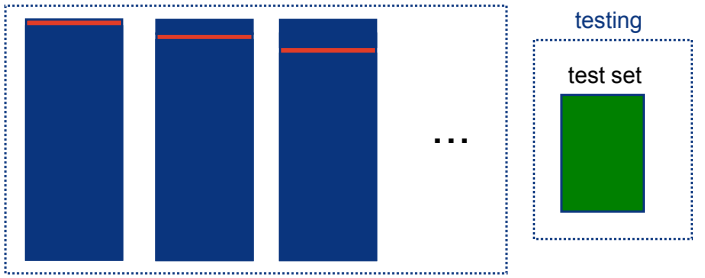
- basic idea:
  - leave some data out of your training set (cross validation set)
  - train with different parameters
  - evaluate performance on cross validation set
  - pick best parameter configuration






# Leave-one-out cross-validation

- many variations
- leave-one-out CV:
  - compute n estimators of $P_X(x)$ by leaving one $X_i$ out at a time
  - for each $P_X(x)$ evaluate $P_X(X_i)$ on the point that was left out
  - pick $P_X(x)$ that maximizes this likelihood



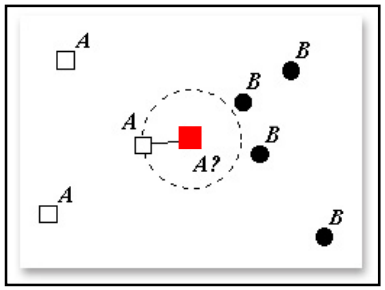
# Non-parametric classifiers

- given kernel density estimates for all classes we can compute the BDR
- since the estimators are non-parametric the resulting classifier will also be non-parametric
- this term is general and applies to any learning algorithm
- a very simple example is the nearest neighbor classifier



# Nearest neighbor classifier

- is the simplest possible classifier that one could think of:
  - it literally consists of assigning to the vector to classify the label of the closest vector in the training set
  - to classify the red point:
    - measure the distance to all other points
    - if the closest point is a square, assign to "square" class
    - otherwise assign to "circle" class



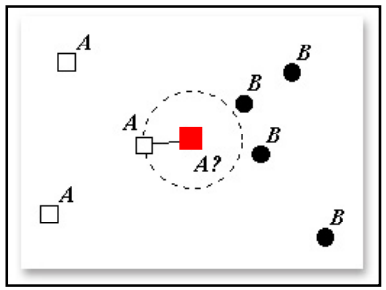

- it works a lot better than what one might predict



# Nearest neighbor classifier

- to define it mathematically we need to define
  - a training set $\mathcal{D} = \{(x_1,y_1), \ldots, (x_n,y_n)\}$
  - $x_i$ is a vector of observations, $y_i$ is the label
  - a vector $x$ to classify
- the "decision rule" is

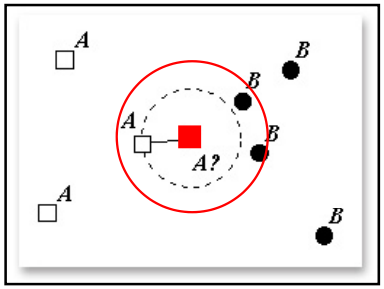
$$\begin{aligned} &set \quad y = y_{i^*} \\ &where \\ &i^* = \arg\min_{i \in \{1,\ldots,n\}} d(x, x_i) \end{aligned}$$






# k-nearest neighbors

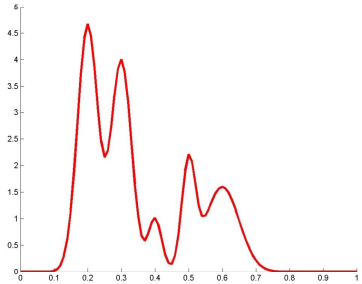
- instead of the NN, assigns to the majority vote of the k nearest neighbors
- in this example
  - NN rule says "A"
  - but 3-NN rule says "B"
- for x away from the border does not make much difference




- usually best performance for k > 1, but there is no universal number
- k large: performance degrades (no longer neighbors)
- k should be odd, to prevent ties

# Mixture density estimates

- back to BDR-based classifiers
  Consider the problem of instrument classification
- summary:
  - Estimate instrument type (brass, string, percu) from audio
  - Measure some audio feature
  - estimate pdf
  - use BDR
- clearly this is not Gaussian
- possible solution: use a kernel-based model

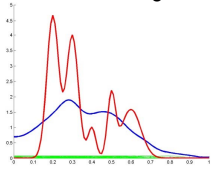




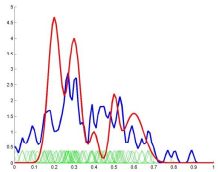
# Kernel-based estimate

- simple learning procedure
  - measure audio feature X
  - place a Gaussian on top of each measurement
- can be overkill
  - spending all degrees of freedom (# of training points) just to get the Gaussian means
  - cannot use the data to determine variances
- handpicking of bandwidth can lead to too much bias or variance

bandwidth too large: bias



bandwidth too small: variance



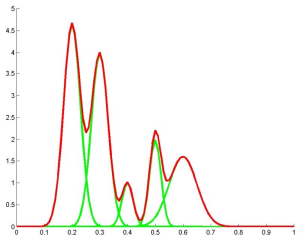


# mixture density estimate

- it looks like we could do better by just picking the right # of Gaussians
- this is indeed a good model:
  - density is multimodal because there is a hidden variable Z
    - Z can determine the type of intermediate musical instruments (for example)

$$Z \in \{Violin, Piano, Saxophone, Flute, Drum\}$$

    - Note that this is different from *Y* which is the instrument type (brass, string, percussion)
    - For a given instrument type, the density is approximate Gaussian here.
    - The density is a *mixture of Gaussians*

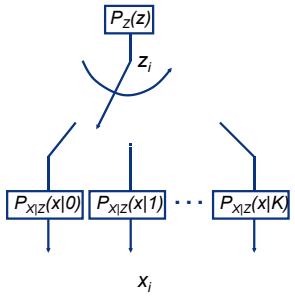




# mixture model

- two types of random variables
  - *Z* – hidden state variable
  - *X* – observed variable
- observations sampled with a two-step procedure
  - a state (class) is sampled from the distribution of the hidden variable

$$P_Z(z) \rightarrow z_i$$

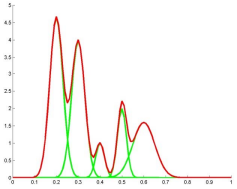



  - an observation is drawn from the class conditional density for the selected state

$$P_{X|Z}(x|z_i) \rightarrow x_i$$

## mixture model

- the sample consists of pairs $(x_i, z_i)$
  $D = \{(x_1, z_1), \ldots, (x_n, z_n)\}$
  but we never get to see the $z_i$



- the pdf of the observed data is

$$P_{\mathbf{X}}(\mathbf{x}) = \sum_{c=1}^{C} P_{\mathbf{X}|Z}(\mathbf{x}|c) P_Z(c) = \sum_{c=1}^{C} P_{\mathbf{X}|Z}(\mathbf{x}|c)\pi_c$$

# of mixture components

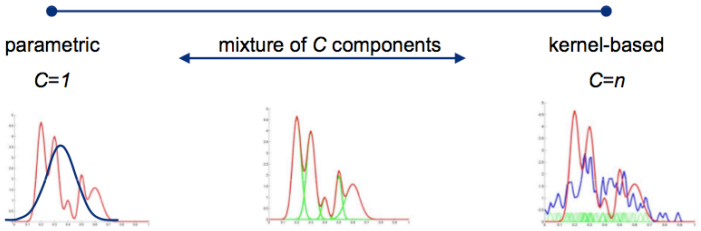
component "weight"

$c^{th}$ "mixture component"



## Mixtures vs Kernel and parametric

- A parametric model is a mixture with *one* component
  - The weight is *one*
  - The mixture density is the parametric density itself!
  - More degrees of freedom in mixture => less bias
- A mixture density is like a kernel density less components
  - less components => less learning parameters, less variance
- Mixture is a compromise between these two extremes:






## mixture disadvantages

- main disadvantage is learning complexity
- non-parametric estimates
  - simple: store the samples (NN); place a kernel on top of each point (kernel-based)
- parametric estimates
  - small amount of work: if ML equations have closed-form
  - substantial amount of work: otherwise (numerical solution)
- mixtures:
  - there is usually no closed-form solution
  - always need to resort to numerical procedures
- standard tool is the expectation-maximization (EM) algorithm



# Clustering and EM algorithm

# Classes vs. Clusters

- Supervised: $\mathcal{X} = \{ \mathbf{x}^t, \mathbf{r}^t \}_t$
- Classes $C_i$ $i$=1,...,$K$

$$p(\mathbf{x}) = \sum_{i=1}^{K} p(\mathbf{x} \mid C_i) P(C_i)$$

where $p(\mathbf{x} \mid C_i) \sim \mathcal{N}(\mu_i, \Sigma_i)$

- $\Phi = \{P(C_i), \mu_i, \Sigma_i\}_{i=1}^{K}$

$$\hat{P}(C_i) = \frac{\sum_t r_i^t}{N} \quad \mathbf{m}_i = \frac{\sum_t r_i^t \mathbf{x}^t}{\sum_t r_i^t}$$

$$\mathbf{S}_i = \frac{\sum_t r_i^t (\mathbf{x}^t - \mathbf{m}_i)(\mathbf{x}^t - \mathbf{m}_i)^T}{\sum_t r_i^t}$$

- Unsupervised : $\mathcal{X} = \{ \mathbf{x}^t \}_t$
- Clusters $G_i$ $i$=1,...,$k$

$$p(\mathbf{x}) = \sum_{i=1}^{k} p(\mathbf{x} \mid Z_i) P(Z_i)$$

where $p(\mathbf{x} \mid Z_i) \sim \mathcal{N}(\mu_i, \Sigma_i)$

- $\Phi = \{P(Z_i), \mu_i, \Sigma_i\}_{i=1}^{k}$

Labels, $\mathbf{r}^t_i$?



# K-Means Clustering

- Dataset $D = \{x_1, x_2, \ldots, x_n\}$
- Goal: Partition in *K* clusters
  - Cluster prototype: $\mu_k$
  - Binary indicator variable (1-of-K coding scheme)

$$r_{nk} \in \{0, 1\}$$
$$r_{nk} = 1, \quad \text{and } r_{nj} = 0 \text{for } j \neq k$$

*hard assignment*

  - Distortion measure

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} ||x_n - \mu_k||^2$$

- K-means gives *k* reference vectors (prototypes) which can be used as decision rule
  - Our sample communication problem:

$$E(\{\mathbf{m}_i\}_{i=1}^{k} | \mathcal{X}) = \sum_t \sum_i b_i^t \|\mathbf{x}^t - \mathbf{m}_i\|$$
$$b_i^t = \begin{cases} 1 & \text{if } \|\mathbf{x}^t - \mathbf{m}_i\| = \min_j \|\mathbf{x}^t - \mathbf{m}_j\| \\ 0 & \text{otherwise} \end{cases}$$



# K-Means clustering: EM

- Find values for $\{r_{nk}\}$ and $\{\mu_k\}$ to minimize:

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} ||x_n - \mu_k||^2$$

- Iterative procedure:
  - Expectation: Minimize *J* with regards to $\{r_{nk}\}$, keep $\{\mu_k\}$ fixed

$$r_{nk} = \begin{cases} 1, & \text{if } k = \arg\min_j ||x_n - \mu_k||^2 \\ 0, & \text{otherwise} \end{cases}$$

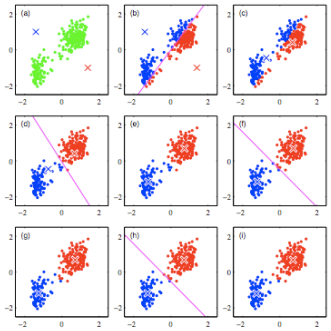
  - Maximization: Minimize J with regards to $\{\mu_k\}$, keep $\{r_{nk}\}$ fixed

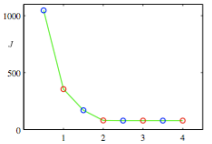
$$2 \sum_{n=1}^{N} r_{nk} (x_n - \mu_k) = 0$$

$$\mu_k = \frac{\sum_n r_{nk} x_n}{\sum_n r_{nk}}$$



# K-Means Clustering: Example





- Each E or M step reduces the value of the objective function J
- Convergence to a global or local maximum

# K-means and clustering

- Problems:
  - How many clusters? (K)
    - Various methods available: *Bayesian Information Criterion, Akaike Information Criterion, Minimum Description Length*
    - Or guessing + cross-validation!
  - Local minimum only
    - Can be a source of head-ache!
  - Initialization of the means
    - Another source of head-ache!
    - Usual method: *mean-splitting...* .
- *Sounds great..*
  - But what about non-classification problems?!



# The basics of EM

- as usual, we start from an iid sample $D = \{x_1, \ldots, x_N\}$
- goal is to find parameters $\Psi^\star$ that maximize likelihood with respect to $D$

$$\Psi^\star = \arg\max_\Psi P_{\mathbf{X}}(\mathcal{D}; \Psi)$$
$$= \arg\max_\Psi \int P_{\mathbf{X}|Z}(\mathcal{D}|z; \Psi) P_Z(z; \Psi) dz$$

- the set

$$D_c = \{(x_1, z_1), \ldots, (x_N, z_N)\}$$

  is called the complete data
- the set

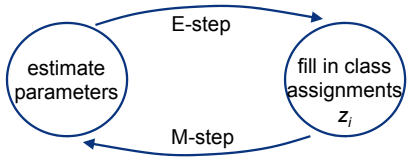
$$D = \{x_1, \ldots, x_N\}$$

  is called the incomplete data



# Learning with incomplete data (EM)

- the basic idea is quite simple
  1. start with an initial parameter estimate $\Psi^{(0)}$
  2. **E-step:** given current parameters $\Psi^{(i)}$ and observations in $D$, "guess" what the values of the $z_i$ are
  3. **M-step:** with the new $z_i$, we have a complete data problem, solve this problem for the parameters, i.e. compute $\Psi^{(i+1)}$
  4. go to 2.
- this can be summarized as






# Classification-maximization

- C-step:
  - given estimates $\Psi^{(i)} = \{\Psi^{(i)}_1, \ldots, \Psi^{(i)}_C\}$
  - determine $z_i$ by the BDR

$$z_l = \arg\max_c P_{\mathbf{X}|Z}\left(\mathbf{x}_l|c; \Psi_c^{(i)}\right) \pi_c^{(i)}, l \in \{1, \ldots, n\}$$

  - split the training set according to the labels $z_i$

    $D^1 = \{x_i|z_i=1\}$, $D^2 = \{x_i|z_i=2\}$, ... , $D^C = \{x_i|z_i=C\}$
- M-step:
  - as before, determine the parameters of each class independently

$$\Psi_c^{(i+1)} = \arg\max_{\Psi,\pi} P_{\mathbf{X}|Z}(\mathcal{D}^c|c, \Psi)\pi$$

## For Gaussian mixtures

- C-step:
  - $z_l = \arg\max_c \left\{ -\frac{1}{2}\left(\mathbf{x}_l - \mu_c^{(i)}\right)^T \left(\Sigma_c^{(i)}\right)^{-1} \left(\mathbf{x}_l - \mu_c^{(i)}\right) -\frac{1}{2} log \left|\Sigma_c^{(i)}\right| + \log \pi_c^{(i)} \right\}, l \in \{1, \ldots, n\}$
  - split the training set according to the labels $z_i$

    $D^1 = \{x_i | z_i = 1\}$, $D^2 = \{x_i | z_i = 2\}$, … , $D^C = \{x_i | z_i = C\}$

- M-step:
  - $\pi_c^{(i+1)} = \frac{|\{\mathbf{x}_i \in \mathcal{D}^c\}|}{n}$ $\mu_c^{(i+1)} = \frac{1}{|\{\mathbf{x}_i \in \mathcal{D}^c\}|} \sum_{i|\mathbf{x}_i \in \mathcal{D}^c} \mathbf{x}_i$

    $\Sigma_c^{(i+1)} = \frac{1}{|\{\mathbf{x}_i \in \mathcal{D}^c\}|} \sum_{i|\mathbf{x}_i \in \mathcal{D}^c} \left(\mathbf{x}_i - \mu_c^{(i+1)}\right)\left(\mathbf{x}_i - \mu_c^{(i+1)}\right)^T$



## K-means

- when covariances are identity and priors uniform
- C-step:
  - $z_l = \arg\min_c ||\mathbf{x}_l - \mu_c^{(i)}||^2, \quad l \in \{1, \ldots, n\}$
  - split the training set according to the labels $z_i$

    $D^1 = \{x_i | z_i = 1\}$, $D^2 = \{x_i | z_i = 2\}$, … , $D^C = \{x_i | z_i = C\}$
- M-step:
  - $\mu_c^{(i+1)} = \frac{1}{|\{\mathbf{x}_i \in \mathcal{D}^c\}|} \sum_{i|\mathbf{x}_i \in \mathcal{D}^c} \mathbf{x}_i$
- this is the K-means algorithm, aka generalized Loyd algorithm, aka LBG algorithm in the vector quantization literature:
  - "assign points to the closest mean; recompute the means"



## Expectation-Maximization

- What about problems that are not about classification?
- EM suggests:
  - Do the most intuitive operation that is ALWAYS possible
  - Don't worry about Z_i directly
  - E-Step: "estimate the likelihood of the complete data by its expected value given the observed data"
  - M-step: "Maximize this expected value"
  - This leads to the so called *Q-function*



## The Q function

- is defined as

  $Q(\Psi; \Psi^{(n)}) = E_{Z|\mathbf{X};\Psi^{(n)}} \left[\log P_{\mathbf{X},Z}(\mathcal{D}, \{z_1, \ldots, z_N\}; \Psi) | \mathcal{D}\right]$
- and is a bit tricky:
  - it is the expected value of likelihood with respect to complete data (joint X and Z)
  - given that we observed incomplete data (X=$\mathcal{D}$)
  - note that the likelihood is a function of $\Psi$ (the parameters that we want to determine)
  - but to compute the expected value we need to use the parameter values from the previous iteration (because we need a distribution for Z|X)
- the EM algorithm is, therefore, as follows

# Expectation-maximization

- E-step:
  - given estimates $\Psi^{(n)} = \{\Psi^{(n)}_1, \ldots, \Psi^{(n)}_C\}$
  - compute expected log-likelihood of complete data

$$Q(\Psi; \Psi^{(n)}) = E_{Z|\mathbf{X};\Psi^{(n)}} \left[ \log P_{\mathbf{X},Z}(\mathcal{D}, \{z_1, \ldots, z_N\}; \Psi) | \mathcal{D} \right]$$

- M-step:
  - find parameter set that maximizes this expected log-likelihood

$$\Psi^{(n+1)} = \arg\max_{\Psi} Q(\Psi; \Psi^{(n)})$$

- let's make this more concrete by looking at the mixture case



# Expectation-maximization

- to derive an EM algorithm you need to do the following
1. write down the likelihood of the COMPLETE data
2. E-step: write down the Q function, i.e. its expectation given the observed data
3. M-step: solve the maximization, deriving a closed-form solution if there is one



# EM for mixtures (step 1)

- the first thing we always do in a EM problem is
  - compute the likelihood of the COMPLETE data
- very neat trick to use when z is discrete (classes)
  - instead of using *z in {1, 2, ..., C}*
  - use a binary vector of size equal to the # of classes

$$\mathbf{z} \in \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}, \ldots, \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} \right\}$$

  - where *z = j* in the *z in {1, 2, ..., C}* notation, now becomes

$$\mathbf{z} = \mathbf{e}_j = \begin{bmatrix} 0 \\ \vdots \\ 1 \quad (j^{th} position) \\ \vdots \\ 0 \end{bmatrix}$$



# EM for mixtures (step 1)

- we can now write the complete data likelihood as

$$\begin{aligned} P_{\mathbf{X},\mathbf{Z}}(\mathbf{x}, \mathbf{z}; \Psi) &= P_{\mathbf{X}|\mathbf{Z}}(\mathbf{x}|\mathbf{z}; \Psi) P_{\mathbf{Z}}(\mathbf{z}; \Psi) \\ &= \prod_{j=1}^{C} \left[ P_{\mathbf{X}|\mathbf{Z}}(\mathbf{x}|\mathbf{e}_j, \Psi) \pi_j \right]^{z_j} \end{aligned}$$

- for example, if *z = k* in the *z in {1, 2, ..., C}* notation,

$$\begin{aligned} P_{\mathbf{X},Z}(\mathbf{x}, k; \Psi) &= P_{\mathbf{X},\mathbf{Z}}(\mathbf{x}, \mathbf{e}_k; \Psi) \\ &= \left[ P_{\mathbf{X}|\mathbf{Z}}(\mathbf{x}|\mathbf{e}_k, \Psi) \pi_k \right]^1 \prod_{j \neq k} \left[ P_{\mathbf{X}|\mathbf{Z}}(\mathbf{x}|\mathbf{e}_j, \Psi) \pi_j \right]^0 \end{aligned}$$

- the advantage is that

$$\log P_{\mathbf{X},\mathbf{Z}}(\mathbf{x}, \mathbf{z}; \Psi) = \sum_{j=1}^{C} z_j \log \left[ P_{\mathbf{X}|\mathbf{Z}}(\mathbf{x}|\mathbf{e}_j, \Psi) \pi_j \right]$$

- becomes LINEAR in the components $z_j$ !!!

# EM for mixtures (step 1)

- for the complete iid dataset $D_c = \{(x_1,z_1), \ldots, (x_N,z_N)\}$

$$P_{\mathbf{X},Z}(\mathcal{D}, \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}; \Psi) = \prod_{i=1}^{N} P_{\mathbf{X},Z}(\mathbf{x}_i, \mathbf{z}_i; \Psi)$$
$$= \prod_{i=1}^{N} \prod_{j=1}^{C} \left[P_{\mathbf{X}|\mathbf{Z}}(\mathbf{x}_i|\mathbf{e}_j, \Psi)\pi_j\right]^{z_{ij}}$$

- and the complete data log-likelihood is

$$\log P_{\mathbf{X},Z}(\mathcal{D}, \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}; \Psi) = \sum_{i,j} z_{ij} \log \left[P_{\mathbf{X}|\mathbf{Z}}(\mathbf{x}_i|\mathbf{e}_j, \Psi)\pi_j\right]$$

- this does not depend on z and simply becomes a constant for the expectation that we have to compute in the E-step



# Expectation-maximization

- to derive an EM algorithm you need to do the following
  1. write down the likelihood of the COMPLETE data
  2. E-step: write down the Q function, i.e. its expectation given the observed data
  3. M-step: solve the maximization, deriving a closed-form solution if there is one
- important E-step advice:
  - do not compute terms that you do not need
  - at the end of the day we only care about the parameters
  - terms of Q that do not depend on the parameters are useless, e.g. in

    *Q = f(z, Ψ) + log(sin z)*

    the expected value of *log(sin z)* appears to be difficult and is completely unnecessary, since it is dropped in the M-step



# EM for mixtures (step 2)

- once we have the complete data likelihood

$$Q(\Psi; \Psi^{(n)}) = E_{Z|\mathbf{X};\Psi^{(n)}}\left[\log P_{\mathbf{X},Z}(\mathcal{D}, \{z_1, \ldots, z_N\}; \Psi)|\mathcal{D}\right]$$
$$= \sum_{i,j} E_{Z|\mathbf{X};\Psi^{(n)}}[z_{ij}|\mathcal{D}] \log \left[P_{\mathbf{X}|\mathbf{Z}}(\mathbf{x}_i|\mathbf{e}_j, \Psi)\pi_j\right]$$

- i.e. to compute the Q function we only need to compute

$$E_{Z|\mathbf{X};\Psi^{(n)}}[z_{ij}|\mathcal{D}], \ \forall i, j$$

- note that this expectation can only be computed because we use $\Psi^{(n)}$
- note that the Q function will be a function of both $\Psi$ and $\Psi^{(n)}$



# EM for mixtures (step 2)

- since $z_{ij}$ is binary and only depends on $x_i$

$$E_{\mathbf{Z}|\mathbf{X};\Psi^{(n)}}[z_{ij}|\mathcal{D}] = P_{\mathbf{Z}|\mathbf{X}}(z_{ij} = 1|\mathbf{x}_i; \Psi^{(n)}) = P_{\mathbf{Z}|\mathbf{X}}(\mathbf{e}_j|\mathbf{x}_i; \Psi^{(n)})$$

- the E-step reduces to computing the posterior probability of each point under each class!
- defining

$$h_{ij} = P_{\mathbf{Z}|\mathbf{X}}(\mathbf{e}_j|\mathbf{x}_i; \Psi^{(n)})$$

- the Q function is

$$Q(\Psi; \Psi^{(n)}) = \sum_{i,j} h_{ij} \log \left[P_{\mathbf{X}|\mathbf{Z}}(\mathbf{x}_i|\mathbf{e}_j, \Psi)\pi_j\right]$$

# Expectation-maximization

- to derive an EM algorithm you need to do the following
  1. write down the likelihood of the COMPLETE data

$$\log P_{\mathbf{X},Z}(\mathcal{D}, \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}; \Psi) = \sum_{i,j} z_{ij} \log \left[ P_{\mathbf{X}|\mathbf{Z}}(\mathbf{x}_i|\mathbf{e}_j, \Psi)\pi_j \right]$$

  2. E-step: write down the Q function, i.e. its expectation given the observed data

$$h_{ij} = P_{\mathbf{Z}|\mathbf{X}}(\mathbf{e}_j|\mathbf{x}_i; \Psi^{(n)})$$

$$Q(\Psi; \Psi^{(n)}) = \sum_{i,j} h_{ij} \log \left[ P_{\mathbf{X}|\mathbf{Z}}(\mathbf{x}_i|\mathbf{e}_j, \Psi)\pi_j \right]$$

  3. M-step: solve the maximization, deriving a closed-form solution if there is one

$$\Psi^{(n+1)} = \arg\max_{\Psi} \sum_{ij} h_{ij} \log \left[ P_{\mathbf{X}|\mathbf{Z}}(\mathbf{x}_i|\mathbf{e}_j, \Psi)\pi_j \right]$$



# EM for Gaussian mixtures

- in summary:
  - CM = EM + hard assignments
  - CM special case, cannot be better
- let's look at the special case of Gaussian mixtures
- **E-step:**

$$\begin{aligned} h_{ij} &= P_{\mathbf{Z}|\mathbf{X}}(\mathbf{e}_j|\mathbf{x}_i; \Psi^{(n)}) \\ &= \frac{\mathcal{G}\left(\mathbf{x}_i, \mu_j^{(n)}, \sigma_j^{(n)}\right)\pi_j^{(n)}}{\sum_{k=1}^{C} \mathcal{G}\left(\mathbf{x}_i, \mu_k^{(n)}, \sigma_k^{(n)}\right)\pi_k^{(n)}} \end{aligned}$$



# M-step for Gaussian mixtures

- **M-step:**

$$\begin{aligned} \Psi^{(n+1)} &= \arg\max_{\Psi} \sum_{ij} h_{ij} \log \left[ \mathcal{G}\left(\mathbf{x}_i, \mu_j, \sigma_j\right)\pi_j \right] \\ &= \arg\min_{\Psi} \sum_{ij} \frac{h_{ij}(\mathbf{x}_i - \mu_j)^2}{2\sigma_j^2} + \frac{h_{ij}}{2} \log \sigma_j^2 - h_{ij} \log \pi_j \end{aligned}$$

- important note:
  - in the M-step, the optimization must be subject to whatever constraint may hold
  - in particular, we always have the constraint $\sum_j \pi_j = 1$
  - as usual we introduce a Lagrangian

$$L = \sum_{ij} \left[ \frac{h_{ij}(\mathbf{x}_i - \mu_j)^2}{2\sigma_j^2} + \frac{h_{ij}}{2} \log \sigma_j^2 - h_{ij} \log \pi_j \right] + \lambda \left( \sum_j \pi_j - 1 \right)$$

Not familiar with Lagrange multipliers? See http://en.wikipedia.org/wiki/Lagrange_multipliers



# M-step for Gaussian mixtures

- Lagrangian

$$L = \sum_{ij} \left[ \frac{h_{ij}(\mathbf{x}_i - \mu_j)^2}{2\sigma_j^2} + \frac{h_{ij}}{2} \log \sigma_j^2 - h_{ij} \log \pi_j \right] + \lambda \left( \sum_j \pi_j - 1 \right)$$

- setting derivatives to zero

$$\begin{aligned} \frac{\partial L}{\partial \mu_j} &= -\sum_i \frac{h_{ij}(\mathbf{x}_i - \mu_j)}{\sigma_j^2} = 0 \\ \frac{\partial L}{\partial \sigma_j^2} &= -\sum_i \left[ \frac{h_{ij}(\mathbf{x}_i - \mu_j)^2}{\sigma_j^4} - \frac{h_{ij}}{\sigma_j^2} \right] = 0 \\ \frac{\partial L}{\partial \pi_j} &= -\sum_i \frac{h_i}{\pi_j} + \lambda = 0 \\ \frac{\partial L}{\partial \lambda} &= \sum_j \pi_j - 1 = 0 \end{aligned}$$

# M-step for Gaussian mixtures

- leads to the update equations

$$\mu_j^{(n+1)} = \frac{\sum_i h_{ij}\mathbf{x}_i}{\sum_i h_{ij}} \qquad \pi_j^{(n+1)} = \frac{1}{n}\sum_i h_{ij}$$

$$\sigma_j^{2(n+1)} = \frac{\sum_i h_{ij}(\mathbf{x}_i - \mu_j)^2}{\sum_i h_{ij}}$$

- comparing to those of CM

$$\pi_c^{(n+1)} = \frac{|\{\mathbf{x}_i \in \mathcal{D}^c\}|}{N} \qquad \mu_c^{(n+1)} = \frac{1}{|\{\mathbf{x}_i \in \mathcal{D}^c\}|}\sum_{i|\mathbf{x}_i\in\mathcal{D}^c}\mathbf{x}_i$$

$$\Sigma_c^{(n+1)} = \frac{1}{|\{\mathbf{x}_i \in \mathcal{D}^c\}|}\sum_{i|\mathbf{x}_i\in\mathcal{D}^c}\left(\mathbf{x}_i - \mu_c^{(n+1)}\right)\left(\mathbf{x}_i - \mu_c^{(n+1)}\right)^T$$

- they are the same up to hard vs soft assignments.



# Expectation-maximization

- note that the procedure is the same for all mixtures
  1. write down the likelihood of the COMPLETE data

$$\log P_{\mathbf{X},Z}(\mathcal{D}, \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}; \Psi) = \sum_{i,j} z_{ij} \log\left[P_{\mathbf{X}|\mathbf{Z}}(\mathbf{x}_i|\mathbf{e}_j, \Psi)\pi_j\right]$$

  2. E-step: write down the Q function, i.e. its expectation given the observed data

$$h_{ij} = P_{\mathbf{Z}|\mathbf{X}}(\mathbf{e}_j|\mathbf{x}_i; \Psi^{(n)})$$

$$Q(\Psi; \Psi^{(n)}) = \sum_{i,j} h_{ij} \log\left[P_{\mathbf{X}|\mathbf{Z}}(\mathbf{x}_i|\mathbf{e}_j, \Psi)\pi_j\right]$$

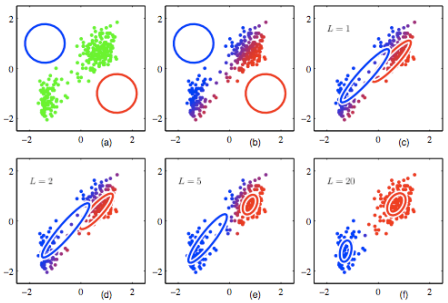
  3. M-step: solve the maximization, deriving a closed-form solution if there is one

$$\Psi^{(n+1)} = \arg\max_{\Psi} \sum_{ij} h_{ij} \log\left[P_{\mathbf{X}|\mathbf{Z}}(\mathbf{x}_i|\mathbf{e}_j, \Psi)\pi_j\right]$$



# EM on Gaussian Mixtures

- Example:






# Group Homework 3

- Derive the EM algorithm for a mixture of exponential distributions:

$$P_X(x) = \sum_{i=1}^{C} \pi_i \lambda_i e^{-\lambda_i x}$$

  - Write down the E-step
  - Write down the M-step, solve the maximization and derive iterative solutions for $\lambda_k$ and $\pi_k$
    - hint: Use the Lagrangian.... .

# Sequential Learning



# Sequential Learning

- Up to now, our approach was rather *static*. Now imagine that you have a problem set where data arrives *sequentially*. *(e.g. Time Series)*
  - *A sequence of observations:* $y_1, y_2, y_3, \ldots, y_t$
  - Considered as independent... :

$$P(y_1, y_2, \ldots, y_t) = \prod_{n=1}^{t} p(y_n | y_1, \ldots, y_{n-1})$$

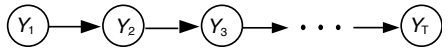
- Many many ways to model sequential data! We will look at some... .



# Markov Models

**First-order Markov model:**

$$P(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_t) = P(\mathbf{y}_1) P(\mathbf{y}_2 | \mathbf{y}_1) \cdots P(\mathbf{y}_t | \mathbf{y}_{t-1})$$




The term *Markov* refers to a conditional independence relationship. In this case, the Markov property is that, given the present observation ($\mathbf{y}_t$), the future ($\mathbf{y}_{t+1}, \ldots$) is independent of the past ($\mathbf{y}_1, \ldots, \mathbf{y}_{t-1}$).

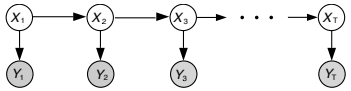
**Second-order Markov model:**

$$P(\mathbf{y}_1, \ldots, \mathbf{y}_t) = P(\mathbf{y}_1) P(\mathbf{y}_2 | \mathbf{y}_1) \cdots P(\mathbf{y}_t | \mathbf{y}_{t-2}, \mathbf{y}_{t-1})$$



# Hidden Variables

Speech recognition:
- $\mathbf{x}$ - underlying phonemes or words
- $\mathbf{y}$ - acoustic waveform

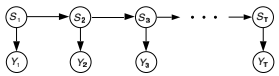



Vision:
- $\mathbf{x}$ - object identities, poses, illumination
- $\mathbf{y}$ - image pixel values

Industrial Monitoring:
- $\mathbf{x}$ - current state of molten steel in caster
- $\mathbf{y}$ - temperature and pressure sensor readings

Two frequently-used tractable models:
- Linear-Gaussian state-space models
- Hidden Markov models

# Hidden Markov Models (HMM)



- Discrete hidden states $s_t \in \{1 \dots, K\}$, and outputs $\mathbf{y}_t$ (discrete or continuous).
  Joint probability factorizes:

$$\mathrm{P}(s_1, \dots, s_\tau, \mathbf{y}_1 \dots, \mathbf{y}_\tau) = \mathrm{P}(s_1)\mathrm{P}(\mathbf{y}_1|s_1) \prod_{t=2}^{\tau} \mathrm{P}(s_t|s_{t-1})\mathrm{P}(\mathbf{y}_t|s_t)$$

- Leading to the following important structural elements:
  - Transition probabilities:

$$a_{ij} = P(s_t = S_j | s_{t-1} = S_i) \,, \quad a_{ij} \geq 0 \; and \; \sum_{j=1}^{N} a_{ij} = 1$$

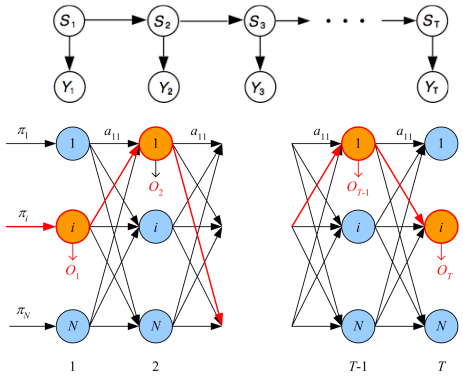
  - Observation probabilities
    - Discrete case: $b_j(m) = P(y_t = Y_m | s_t = S_j)$
  - Initial probabilities (prior):

$$\pi_i = P(s_1 = S_i) \,, \quad \sum_{j=1}^{N} \pi_j = 1$$



# Hidden Markov Models (HMM)

- Notes:
  - Hidden states (S_i) are markov... but not necessarily the output process (Y_i). In reality, they are NOT markov... .
  - For each observation sequence, there are multiple state sequences




- *N*: Number of states
- *M*: Number of observation symbols
- **A** = $[a_{ij}]$: *N* by *N* state transition proba. matrix
- **B** = $b_j(m)$: *N* by *M* observation proba. matrix
- **Π** = $[\pi_i]$: *N* by 1 initial state proba. vector

$\lambda$= (**A**, **B**, **Π**), parameter set of HMM



# HMMs

Three basic HMM problems:

1. Evaluation: Given $\lambda$, and $O$, calculate $P(O \mid \lambda)$
2. State sequence: Given $\lambda$, and $O$, find $Q^*$ such that

$$P(Q^* \mid O, \lambda) = \max_Q P(Q \mid O, \lambda)$$

3. Learning: Given $\mathcal{X}=\{O^k\}_k$, find $\lambda^*$ such that

$$P(\mathcal{X} \mid \lambda^*) = \max_\lambda P(\mathcal{X} \mid \lambda)$$



# Likelihood in HMM

- The likelihood $P(O_1, \dots, O_\tau | \lambda)$ is an extremely hard computation
  - Number of possible paths grow exponentially with time (# of paths=$K^\tau$)
- To compute this likelihood, there exist an efficient *forward* recursion algorithm using dynamic programming:

$$\alpha_t(i) = P(O_1, \dots, O_t, s_t = S_i | \lambda)$$

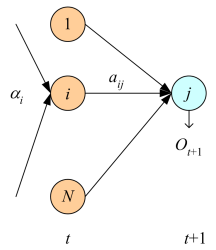
Initialization :

$$\alpha_1(i) = \pi_i b_i(O_1)$$

Recursion :

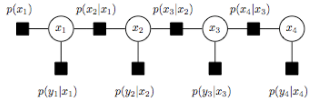
$$\alpha_{t+1}(j) = \left[ \sum_{i=1}^{N} \alpha_t(i) a_i j \right] b_j(O_{t+1})$$

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i)$$



# Likelihood in HMM

or:

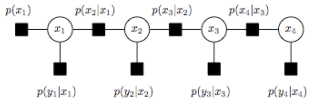


$$
\begin{aligned}
p(y_{1:K}) &= \sum_{x_{1:K}} p(y_{1:K}|x_{1:K})p(x_{1:K}) \\
&= \sum_{x_K} p(y_K|x_K) \sum_{x_{K-1}} p(x_K|x_{K-1}) \cdots \sum_{x_2} p(x_3|x_2)p(y_2|x_2) \sum_{x_1} p(x_2|x_1) \underbrace{p(y_1|x_1)}_{\alpha_{1|1}} \overbrace{p(x_1)}^{\alpha_{1|0}} \\
&= \sum_{x_K} p(y_K|x_K) \sum_{x_{K-1}} p(x_K|x_{K-1}) \cdots \sum_{x_2} p(x_3|x_2)p(y_2|x_2) \sum_{x_1} p(x_2|x_1)\alpha_{1|1}(x_1) \\
&= \sum_{x_K} p(y_K|x_K) \sum_{x_{K-1}} p(x_K|x_{K-1}) \cdots \sum_{x_2} p(x_3|x_2)p(y_2|x_2)\alpha_{2|1}(x_2) \\
&= \sum_{x_K} p(y_K|x_K) \sum_{x_{K-1}} p(x_K|x_{K-1}) \cdots \sum_{x_2} p(x_3|x_2)\alpha_{2|2}(x_2) \\
&= \sum_{x_K} p(y_K|x_K) \sum_{x_{K-1}} p(x_K|x_{K-1}) \cdots \alpha_{3|2}(x_3)
\end{aligned}
$$



# Likelihood in HMM

- Note:



- Predict

$$
\begin{aligned}
\alpha_{k|k-1}(x_k) &= p(y_{1:k-1}, x_k) = \sum_{x_{k-1}} p(x_k|x_{k-1})p(y_{1:k-1}, x_{k-1}) \\
&= \sum_{x_{k-1}} p(x_k|x_{k-1})\alpha_{k-1|k-1}(x_{k-1})
\end{aligned}
$$

- Update

$$
\begin{aligned}
\alpha_{k|k}(x_k) &= p(y_{1:k}, x_k) = p(y_k|x_k)p(y_{1:k-1}, x_k) \\
&= p(y_k|x_k)\alpha_{k|k-1}(x_k)
\end{aligned}
$$



# Most likely state-sequence

- Forward variable gives the likelihood of an observation sequence given model parameters or $P(O|\lambda)$
- What about the most-likely sequence or $P(S^*|O,\lambda) = \max_S P(S|O,\lambda)$
  - *Forward* variable is the most-likely belief from time *0* to *t* (t=present)
  - Need also a belief towards future (from time t to T)
  - *Backward* variable:

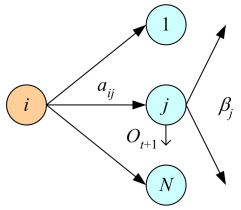
$$\beta_t(i) = P(O_{t+1}, \ldots, O_T, s_t = S_i|\lambda)$$

Initialization :

$$\beta_T(i) = 1$$

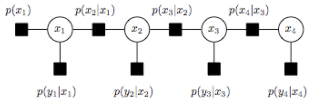
Recursion :

$$\beta_t(j) = \left[\sum_{i=1}^{N} \alpha_{t+1}(i)a_ij\right] b_j(O_{t+1})$$






- Another view:



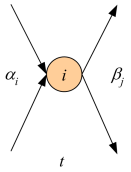
$$
\begin{aligned}
p(y_{1:K}) &= \sum_{x_1} p(x_1)p(y_1|x_1) \ldots \sum_{x_{K-1}} p(x_{K-1}|x_{K-2})p(y_{K-1}|x_{K-1}) \sum_{x_K} p(x_K|x_{K-1})p(y_K|x_K) \underbrace{1}_{\beta_{K|K+1}} \\
&= \sum_{x_1} p(x_1)p(y_1|x_1) \cdots \sum_{x_{K-1}} p(x_{K-1}|x_{K-2})p(y_{K-1}|x_{K-1}) \sum_{x_K} p(x_K|x_{K-1})\beta_{K|K} \\
&= \sum_{x_1} p(x_1)p(y_1|x_1) \cdots \sum_{x_{K-1}} p(x_{K-1}|x_{K-2})p(y_{K-1}|x_{K-1})\beta_{K-1|K} \\
&= \sum_{x_1} p(x_1)p(y_1|x_1) \cdots \sum_{x_{K-1}} p(x_{K-1}|x_{K-2})\beta_{K-1|K-1} \\
&= \sum_{x_1} p(x_1)p(y_1|x_1) \ldots \beta_{K-2|K-1}
\end{aligned}
$$

# Most-likely state-sequence

- Combine the two propagations and use maximum likelihood?

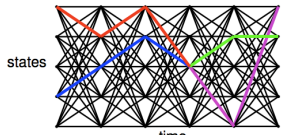
$$s_t(i) = P(s_t = S_i|O,\lambda) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N}\alpha_t(j)\beta_t(j)}$$

$$s_t^* = \arg\max_i s_t(i)$$




Problem:

- Is the set of locally optimal states, equal to the global optimal path?
  - NO! not necessarily... .
    - $\alpha_i(t)$ gives total *inflow* of prob. to node $(t,i)$; $\beta_i(t)$ gives total *outflow* of prob.




- Need to maximize over the WHOLE PATH and not just one state:

$$s_t(i) = \max_{s_1 s_2 \cdots s_{t-1}} P(s_1, s_2 \cdots s_{t-1}, s_t = S_i, O_1, \cdots, O_t|\lambda)$$



# Viterbi Algorithm

$$s_t(i) = \max_{s_1 s_2 \cdots s_{t-1}} P(s_1, s_2 \cdots s_{t-1}, s_t = S_i, O_1, \cdots, O_t|\lambda)$$

- Initialization:

$$s_1(i) = \pi_i b_i(O_1), \quad \psi_1(i) = 0$$

- Recursion:

$$s_t(j) = \max_i s_{t-1}(i) a_{ij} b_j(O_t)$$
$$\psi_t(j) = \arg\max_i s_{t-1}(i) a_{ij}$$

- Termination:

$$p^* = \max_i s_T(i), \; s_T^* = \arg\max_i s_T(i)$$

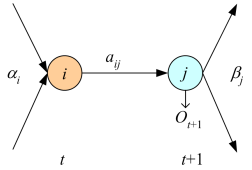
- Backtracking:

$$s_t^* = \psi_{t+1}(s_{t+1}^*), \; t = T-1, T-2, \ldots, 1$$



# Parameter Learning

$$\xi_t(i,j) = P(s_t = S_i, s_{t+1}|O,\lambda) = \frac{\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{\sum_k\sum_\ell \alpha_t(k)a_{k\ell}b_\ell(O_{t+1})\beta_{t+1}(\ell)}$$




- Use EM Algorithm:
  - Auxiliary variable:

$$Z_i^t = \begin{cases} 1, & \text{if } s_t = S_i \\ 0, & \text{otherwise} \end{cases} \qquad Z_{ij}^t = \begin{cases} 1, & \text{if } s_t = S_i \text{ and } s_{t+1} = S_j \\ 0, & \text{otherwise} \end{cases}$$

  - E-Step: $E[z_i^t] = s_t(i) \quad E[z_{ij}^t] = \xi_t(i,j)$
    - recall: $s_t(i) = P(s_t = S_i|O,\lambda) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N}\alpha_t(j)\beta_t(j)}$
  - M-Step:

$$\hat{\pi}_i = \frac{\sum_{k=1}^{K}}{s_t^k(i)}K$$

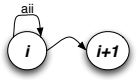
$$\hat{a}_{ij} = \frac{\sum_{k=1}^{K}\sum_{t=1}^{T_k-1}\xi_t^k(i,j)}{\sum_{k=1}^{K}\sum_{t=1}^{T_k-1}s_t^k(i)}$$

$$\hat{b}_j(m) = \frac{\sum_{k=1}^{K}\sum_{t=1}^{T_k-1}s_t^k(j)\delta(O_t^k - v_m)}{\sum_{k=1}^{K}\sum_{t=1}^{T_k-1}s_t^k(i)}$$



# HMM State Duration

- The time an observation would spend in a state is *implicit*:




$$p_i(d) = a_{ii}^{(d-1)}(1 - a_{ii})$$

  - where $p_i(d)$ is the probability of staying 'd' discrete times in state *i*
  - This is an *exponential* model of time
  - Not very desirable for all temporal sequences... such as music!

# HMM Variants

- Discrete observations:

$$P(O_t \mid s_t = S_j, \lambda) = \prod_{m=1}^{M} b_j(m)^{r_m^t} \qquad r_m^t = \begin{cases} 1 & \text{if } O_t = v_m \\ 0 & \text{otherwise} \end{cases}$$

- Continuous observations:

$$P(O_t \mid s_t = S_j, \lambda) \sim \mathcal{N}(\mu_j, \sigma_j^2)$$

  use EM to find parameters... .

- Gaussian Mixtures:

$$P(O_t \mid q_t = S_j, \lambda) = \sum_{l=1}^{L} P(G_{jl}) p(O_t \mid q_t = S_j, G_l, \lambda) \sim \mathcal{N}(\mu_l, \Sigma_l)$$

- Duration-Focused models
  - Transitions as explicit functions of time... .
- And many more... .



# HMMs

- So far, we have assumed that our underlying *models* are *static*!
  - In some cases this works out well as an approximation
  - In many cases it won't!
  - Real-life systems are *dynamic systems*
- Dynamics systems
  - Are generally hard to model... .
  - Much easier if they are linear... .
  - Much harder if they are non-linear... .



# Kalman Filter Models

- Also known as *Linear Dynamical Systems*
- Imagine a sequential framework where latent variable *S* and observations *Y* are continuous
- and underlying dynamics is linear... . (or can be approximated so)
- Example:
  - A one dimensional tracking system:

$$\mathbf{s}_k = \begin{pmatrix} \text{position} \\ \text{velocity} \end{pmatrix}_k = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \mathbf{s}_{k-1} = \mathbf{A}\mathbf{s}_{k-1}$$

$$y_k = \text{position}_k = \begin{pmatrix} 1 & 0 \end{pmatrix} \mathbf{s}_k = \mathbf{C}\mathbf{s}_k$$



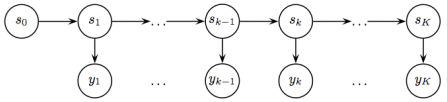
# Kalman Filter Models

- Tracking example:
  - Imagine that we have unknown accelerations (dynamics!)

$$\begin{aligned} \mathbf{s}_k &= \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \mathbf{s}_{k-1} + \epsilon_k \\ &= \mathbf{A}\mathbf{s}_{k-1} + \epsilon_k \end{aligned}$$

$$\begin{aligned} y_k &= \begin{pmatrix} 1 & 0 \end{pmatrix} \mathbf{s}_k + \nu_k \\ &= \mathbf{C}\mathbf{s}_k + \nu_k \end{aligned}$$

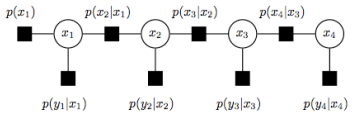
  - Generatively speaking,




$$\begin{aligned} \mathbf{s}_k &\sim \mathcal{N}(\mathbf{s}_k; \mathbf{A}\mathbf{s}_{k-1}, Q) \\ y_k &\sim \mathcal{N}(y_k; \mathbf{C}\mathbf{s}_k, R) \end{aligned}$$

# Kalman Filter Models

Inference is similar to HMM: summations are replaced by integrations




- Forward Pass

$$p(y_{1:K}) = \underbrace{\int_{x_K} p(y_T|x_K) \int_{x_{K-1}} p(x_K|x_{K-1})}_{\alpha_K} \cdots \int_{x_2} p(x_3|x_2) \underbrace{p(y_2|x_2) \overbrace{\int_{x_1} p(x_2|x_1)}^{\alpha_{2|1}} \underbrace{p(y_1|x_1) \overbrace{p(x_1)}^{\alpha_{1|0}}}_{\alpha_1}}_{\alpha_2}$$

- Backward Pass

$$p(y_{1:K}) = \int_{x_1} p(x_1) p(y_1|x_1) \cdots \underbrace{\int_{x_{K-1}} p(x_{K-1}|x_{K-2}) p(y_{K-1}|x_{K-1})}_{\beta_{K-2}} \underbrace{\int_{x_K} p(x_K|x_{K-1}) p(y_K|x_K)}_{\beta_{K-1}} \underbrace{1}_{\beta_K}$$



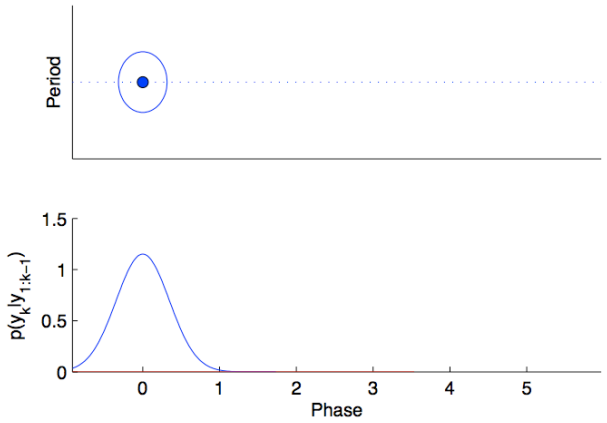
# Kalman Filter Models

- Goal:
  - Find an a posteriori belief based on prior estimate and a weighted difference between the *actual measurement* $y_t$ and a *measurement prediction* $\hat{s}_t$
  - A series of *predictions* <> *corrections*
- *Results into closed form solutions in the Gaussian case...*



# Kalman Filters

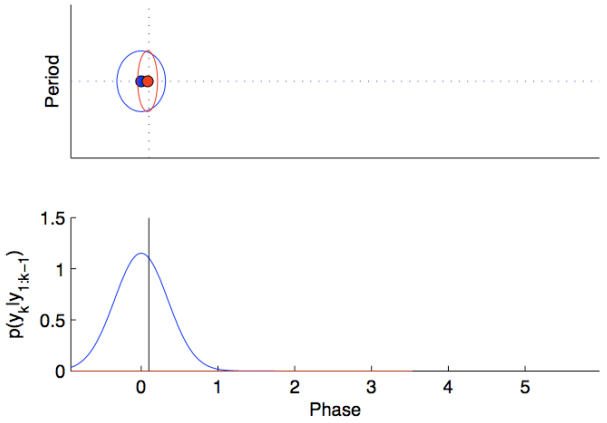
$p(s_1)$





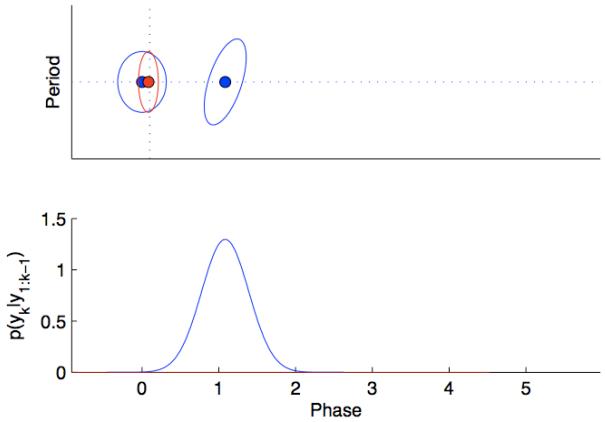

# Kalman Filters

$p(y_1|s_1)p(s_1)$



# Kalman Filters

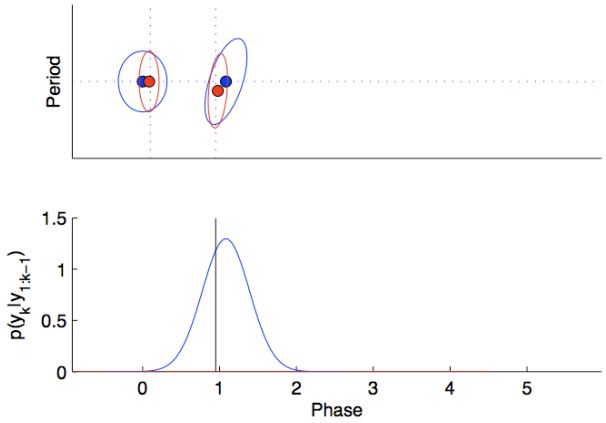
$$p(s_2|y_1) \propto \int ds_1 p(s_2|s_1)p(y_1|s_1)p(s_1)$$






# Kalman Filters

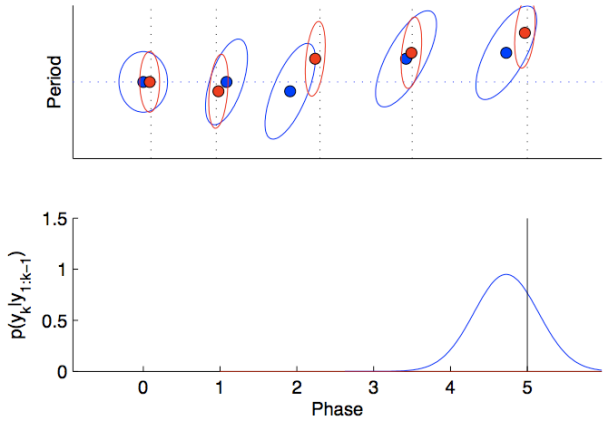
$$p(y_2|s_2)p(s_2|y_1)$$






# Kalman Filters

$$p(s_5|y_{1:5})$$






# Sequential learning applications in audio

- Score Following (AKA real-time alignment of audio to symbolic scores)
- Gesture Following
- Speech Recognition
- Automatic Transcription
- and many many many more... .

# Discriminant Learning

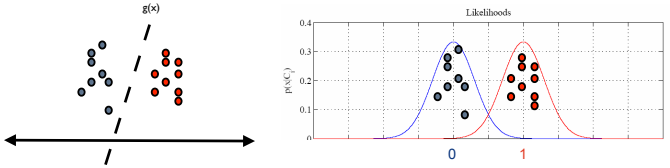
master ATiAM '13

# Likelihood vs. Discriminant

- Likelihood based classification
  - Assume a model for $p(x|C_i)$ and use Bayes' rule to calculate $p(C_i|x)$

$$g_i(x) \sim \log P(C_i|x)$$

- Discriminant based
  - Assume a model for $g_i(x|\Psi_i)$ ; no density estimation
  - Estimating the boundaries is enough
  - No need to accurately estimate the densities inside the boundaries



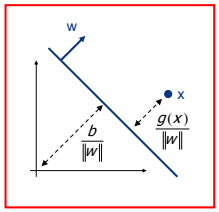



# Linear Discriminant

- Linear discriminant $g_i(x|w,b) = w^T x + b = \sum_{j=1}^{d} w_j x_j + b$
  - in summary, the decision rule

$$h^*(x) = \begin{cases} 0 & \text{if } g(x) > 0 \\ 1 & \text{if } g(x) < 0 \end{cases} \quad \text{with} \quad g(x) = w^T x + b$$

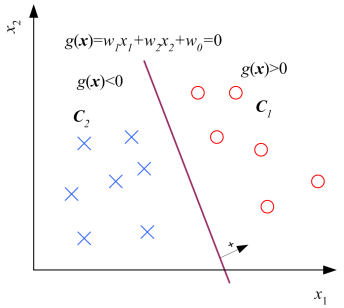
  - has the properties
    - it divides $\mathcal{X}$ into two "half-spaces"
    - boundary is the plane with:
      - normal *w*
      - distance to the origin $b/\|w\|$
    - g(x)/||w|| is the distance from point x to the boundary
      - g(x) = 0 for points on the plane
      - g(x) > 0 on the side *w* points to ("positive side")
      - g(x) < 0 on the "negative side"






# Example

- Two classes:

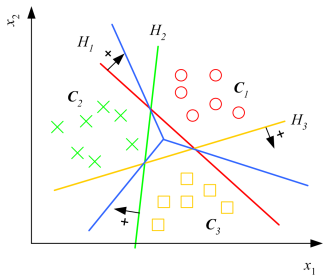



$$\begin{aligned} g(x) &= g_1(x) - g_2(x) \\ &= (w_1^T x + w_{10}) - (w_2^T x + w_{20}) \\ &= (w_1 - w_2)^T x + (w_{10} - w_{20}) \\ &= w^T x + w_0 \end{aligned}$$

$$\text{choose} \begin{cases} C_1 & \text{if } g(x) > 0 \\ C_2 & \text{otherwise} \end{cases}$$

# Example

- Multiple Classes




$$g_i(\mathbf{x} \mid \mathbf{w}_i, w_{i0}) = \mathbf{w}_i^T \mathbf{x} + w_{i0}$$

Choose $C_i$ if

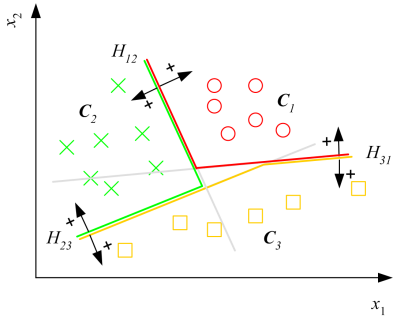
$$g_i(\mathbf{x}) = \max_{j=1}^{K} g_j(\mathbf{x})$$

Classes are linearly separable



# Example

- Pairwise Separation




$$g_{ij}(\mathbf{x} \mid \mathbf{w}_{ij}, w_{ij0}) = \mathbf{w}_{ij}^T \mathbf{x} + w_{ij0}$$

$$g_{ij}(\mathbf{x}) = \begin{cases} > 0 & \text{if } \mathbf{x} \in C_i \\ \leq 0 & \text{if } \mathbf{x} \in C_j \\ \text{don't care} & \text{otherwise} \end{cases}$$

choose $C_i$ if

$$\forall j \neq i, g_{ij}(\mathbf{x}) > 0$$



# Discriminants to Posteriors

- When $p(x|C_i) \sim N(\mu_i, \Sigma)$

$$g_i(x|w_i, b_i) = w_i^T x + b_i$$

$$w_i = \Sigma^{-1}\mu_i \quad , \quad b_i = \frac{-1}{2}\mu_i^T \Sigma^{-1}\mu_i + \log P(C_i)$$

where $y \equiv P(C_1 \mid \mathbf{x})$ and $P(C_2 \mid \mathbf{x}) = 1 - y$

$$\text{choose } C_1 \text{ if } \begin{cases} y > 0.5 \\ y/(1-y) > 1 \\ \log[y/(1-y)] > 0 \end{cases} \text{ and } C_2 \text{ otherwise}$$



# Discriminants to Posteriors

$$\begin{aligned} \text{logit}(P(C_1 \mid \mathbf{x})) &= \log \frac{P(C_1 \mid \mathbf{x})}{1 - P(C_1 \mid \mathbf{x})} = \log \frac{P(C_1 \mid \mathbf{x})}{P(C_2 \mid \mathbf{x})} \\ &= \log \frac{p(\mathbf{x} \mid C_1)}{p(\mathbf{x} \mid C_2)} + \log \frac{P(C_1)}{P(C_2)} \\ &= \log \frac{(2\pi)^{-d/2} |\Sigma|^{-1/2} \exp\left[-(1/2)(\mathbf{x} - \mu_1)^T \Sigma^{-1} (\mathbf{x} - \mu_1)\right]}{(2\pi)^{-d/2} |\Sigma|^{-1/2} \exp\left[-(1/2)(\mathbf{x} - \mu_2)^T \Sigma^{-1} (\mathbf{x} - \mu_2)\right]} + \log \frac{P(C_1)}{P(C_2)} \\ &= \mathbf{w}^T \mathbf{x} + w_0 \end{aligned}$$

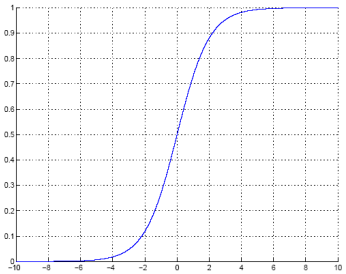
where $\mathbf{w} = \Sigma^{-1}(\mu_1 - \mu_2)$ $w_0 = -\frac{1}{2}(\mu_1 + \mu_2)^T \Sigma^{-1}(\mu_1 - \mu_2)$

The inverse of logit

$$\log \frac{P(C_1 \mid \mathbf{x})}{1 - P(C_1 \mid \mathbf{x})} = \mathbf{w}^T \mathbf{x} + w_0$$

$$P(C_1 \mid \mathbf{x}) = \text{sigmoid}(\mathbf{w}^T \mathbf{x} + w_0) = \frac{1}{1 + \exp\left[-(\mathbf{w}^T \mathbf{x} + w_0)\right]}$$

# Sigmoid (Logistic) Function



1. Calculate $g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0$ and choose $C_1$ if $g(\mathbf{x}) > 0$, or
2. Calculate $y = \text{sigmoid}(\mathbf{w}^T\mathbf{x} + w_0)$ and choose $C_1$ if $y > 0.5$

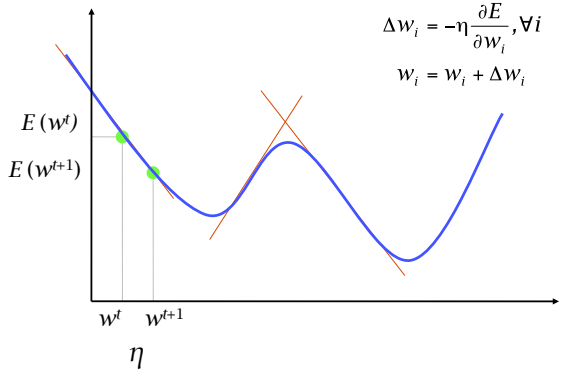


# Gradient-Descent

- E($\mathbf{w}|\mathcal{X}$) is error with parameters $\mathbf{w}$ on sample $\mathcal{X}$
  $\mathbf{w}^* = \arg\min_{\mathbf{w}} E(\mathbf{w} \mid \mathcal{X})$
- Gradient $\nabla_w E = \left[\frac{\partial E}{\partial w_1}, \frac{\partial E}{\partial w_2}, \ldots, \frac{\partial E}{\partial w_d}\right]^T$
- Gradient-descent:
  Starts from random $\mathbf{w}$ and updates $\mathbf{w}$ iteratively in the negative direction of gradient



# Gradient-Descent






# Example

- Two classes:

$$\mathcal{X} = \{\mathbf{x}^t, r^t\}_t \quad r^t \mid \mathbf{x}^t \sim \text{Bernoulli}(y^t)$$

$$y = P(C_1 \mid \mathbf{x}) = \frac{1}{1 + \exp\left[-\left(\mathbf{w}^T\mathbf{x} + w_0\right)\right]}$$

$$l(\mathbf{w}, w_0 \mid \mathcal{X}) = \prod_t \left(y^t\right)^{r^t} \left(1 - y^t\right)^{1 - r^t}$$

$$E = -\log l$$

$$E(\mathbf{w}, w_0 \mid \mathcal{X}) = -\sum_t r^t \log y^t + \left(1 - r^t\right) \log\left(1 - y^t\right)$$

# Example

- Gradient-Descent

$$E(\mathbf{w}, w_0 \mid \mathcal{X}) = -\sum_t r^t \log y^t + (1 - r^t) \log (1 - y^t)$$

$$\text{If } y = \text{sigmoid}(a) \quad \frac{dy}{da} = y(1 - y)$$

$$\Delta w_j = -\eta \frac{\partial E}{\partial w_j} = \eta \sum_t \left( \frac{r^t}{y^t} - \frac{1 - r^t}{1 - y^t} \right) y^t (1 - y^t) x_j^t$$

$$= \eta \sum_t (r^t - y^t) x_j^t, j = 1, \ldots, d$$

$$\Delta w_0 = -\eta \frac{\partial E}{\partial w_0} = \eta \sum_t (r^t - y^t)$$



# Example

- Gradient-Descent Training

```
For j = 0, ..., d
    w_j ← rand(-0.01, 0.01)
Repeat
    For j = 0, ..., d
        Δw_j ← 0
    For t = 1, ..., N
        o ← 0
        For j = 0, ..., d
            o ← o + w_j x_j^t
        y ← sigmoid(o)
        Δw_j ← Δw_j + (r^t − y) x_j^t
    For j = 0, ..., d
        w_j ← w_j + η Δw_j
Until convergence
```



# Generalized Linear Models

- Quadratic:

$$\log \frac{P(x|C_i)}{p(x|C_k)} = x^T W_i x + W_i^T x + w_{i0}$$

- Sum of basis functions $\Phi$

$$\log \frac{P(x|C_i)}{p(x|C_k)} = w_i^T \Phi(x) + w_{i0}$$

  - Kernels in Support Vector Machines (SVM)
  - Hidden units in Neural Networks

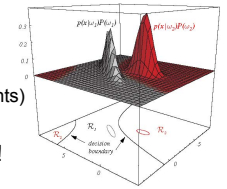


# Generalized Linear Models

- 1) use a higher-order decision function
  - e.g. a quadratic boundary

$$x^T W x + w^T x + w_0 = 0$$

  - is the optimal solution for any Gaussian problem (2 Gaussian classes no constraints)



- looks like we are going to need a very high-order polynomial in general!
  - lots of parameters
  - too much complexity
  - where to stop?
- can we do something else to keep the simplicity of the linear boundary?

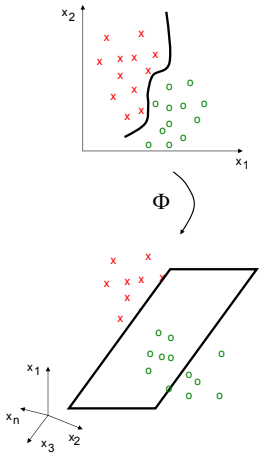
# Linear Discriminants

- 2) transform the space:
- introduce a mapping

$$\Phi: \mathcal{X} \rightarrow \mathcal{Z}$$

such that $dim(\mathcal{Z}) > dim(\mathcal{X})$

- learning a linear boundary in $\mathcal{Z}$ is equivalent to learning a non-linear boundary in $\mathcal{X}$
- e.g.
  - two scalar Gaussians
  - zero mean, different variances






# Linear Discriminants

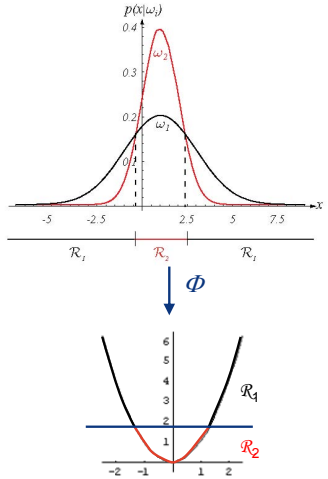
- since $P_{X|Y}(x \mid i) = G(x, 0, \sigma_i)$
- using the BDR

$$h^*(x) = \arg\max_i P_{X|Y}[x \mid i] P_Y[i]$$

leads to this

- cannot be implemented with a linear discriminant
- but becomes feasible by mapping to 2D

$$\Phi: \quad \Re \rightarrow \Re^2$$
$$x \rightarrow (x, x^2)$$






# Linear Discriminants

- But how do we determine $\Phi$?
  - Using KERNELS
  - Kernel functions transform the feature space to a higher-dimensional space!
  - The VERY basic idea is this:
    - Using BDR we know how to solve for an optimal discriminant case ONLY if our two classes are *linearly discriminant*!
    - The *Kernel* transformations, move the world to a higher-dimensional space, and with mathematical care and hope, it that higher-dimensional space, things are linearly discriminant!
    - Once the BDR determines the discriminant factor then we come back to the real-world.

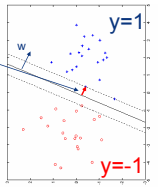


# Linear Discriminants

- But how can Discriminant learning assure generalization if we do not have any models??
  - Optimal Separating Hyperplanes

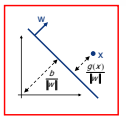
- is the distance from the boundary to the closest point

$$\gamma = \min_i \frac{|w^T x_i + b|}{\|w\|}$$




- there will be no error if it is strictly greater than zero

$$y_i(w^T x_i + b) > 0, \ \forall i \Leftrightarrow \boxed{\gamma > 0}$$




- note that this is ill-defined in the sense that $\gamma$ does not change if both $w$ and $b$ are scaled by $\lambda$
- we need a normalization

make $|g(x)| = 1$ for the closest point, i.e.

$$\boxed{\min_i |w^T x_i + b| = 1}$$

under which

$$\boxed{\gamma = \frac{1}{\|w\|}}$$

# Support Vector Machines

- Under this normalization, the empirical error is zero IF AND ONLY IF

$$|w^T x_i + b| \geq 1 \quad \forall i \Leftrightarrow$$
$$\text{sgn}(w^T x_i + b)(w^T x_i + b) \geq 1 \quad \forall i \Leftrightarrow$$
$$\boxed{y_i (w^T x_i + b) \geq 1 \quad \forall i}$$

- the SVM is the classifier that maximizes the margin under this set of constraints

$$\boxed{\min_{w,b} \|w\|^2 \text{ subject to } y_i (w^T x_i + b) \geq 1 \quad \forall i}$$



# Support Vector Machines

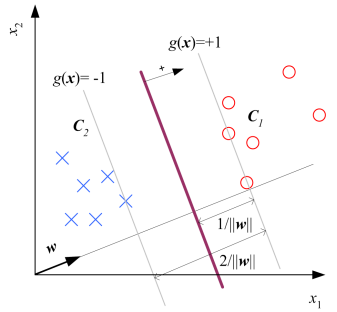
$$\min \frac{1}{2}\|w\|^2 \text{ subject to } r^t (w^T x^t + w_0) \geq +1, \forall t$$

$$L_p = \frac{1}{2}\|w\|^2 - \sum_{t=1}^{N} \alpha^t \left[ r^t (w^T x^t + w_0) - 1 \right]$$

$$= \frac{1}{2}\|w\|^2 - \sum_{t=1}^{N} \alpha^t r^t (w^T x^t + w_0) + \sum_{t=1}^{N} \alpha^t$$

$$\frac{\partial L_p}{\partial w} = 0 \Rightarrow w = \sum_{t=1}^{N} \alpha^t r^t x^t$$

$$\frac{\partial L_p}{\partial w_0} = 0 \Rightarrow \sum_{t=1}^{N} \alpha^t r^t = 0$$






# Support Vector Machines

$$L_p = \frac{1}{2}(\mathbf{w}^T \mathbf{w}) - \mathbf{w}^T \sum_t \alpha^t r^t \mathbf{x}^t - w_0 \sum_t \alpha^t r^t + \sum_t \alpha^t$$

$$= -\frac{1}{2}(\mathbf{w}^T \mathbf{w}) + \sum_t \alpha^t$$

$$= -\frac{1}{2} \sum_t \sum_s \alpha^t \alpha^s r^t r^s (\mathbf{x}^t)^T \mathbf{x}^s + \sum_t \alpha^t$$

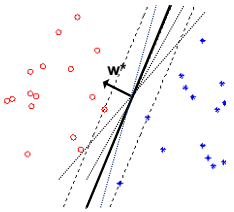
$$\text{subject to } \sum_t \alpha^t r^t = 0 \text{ and } \alpha^t \geq 0, \forall t$$

- Most $\alpha^t$ are zero and only a small number are useful... .
  - These are support vectors



# Intuition

- this is penalizing complexity
  - e.g. the smaller the ||w|| the larger the number of components set to zero
- this is searching for the more stable hyperplane
  - among the ones that have zero training error
  - is the one that has most room for discrepancies between training and testing
  - the margin as a “security gap”
- there are many details which we have not filled



# Finally...

master ATiAM '13

# In this course...

- We looked at some basic problem solving approaches in machine learning literature...
  - Introduction to Bayesian Decision Theory and Learning
  - Gaussian Classifiers, EM Algorithm,
  - Basics of Sequential Learning and Decision theory
  - Introduction to Discriminant Learning Theory
- What we did not see... :
  - Dimensionality Reduction Algorithms: *Principle Component Analysis (PCA), Independent Component Analysis (ICA), Non-negative Matrix Factorization (NMF) etc.*
  - Fuzzy logic based algorithms
  - Some important unsupervised learning approaches: *Spectral Clustering etc.*
  - Important sequential learning algorithms: *Reinforcement Learning (RL), Active Learning etc.*
  - *... and much more ...*



# Do not forget...

- Our goal today is to *introduce* some well-known and well-established approaches in AI and Machine Learning
  - The methods presented today are not *domain-specific* but for every problem, you start with a design, collect *related data* and then define the learning problem. We will not get into *design* today... .
- Keep in mind that,
  - AI is an *empirical science! (See "Science of the Artificial" by* H.A. Simons, MIT Press, 1969)
  - DO NOT apply algorithms blindly to your data/problem set!
    - The MATLAB Toolbox syndrome: Examine the hypothesis and limitation of each approach before hitting enter!
  - Do not forget your own intelligence!

Contact: cont@ircam.fr