EE 583
PATTERN RECOGNITION

Statistical Pattern Recognition

Bayes Decision Theory

Supervised Learning

Linear Discriminant Functions

Unsupervised Learning

Supervised Learning

- Supervised Learning == Training
  - Parametric approaches
    - Maximum likelihood estimation
    - Bayesian parameter estimation
  - Non-parametric approaches
    - Direct pdf (multi-D histogram) estimation
    - Parzen window pdf estimation
    - k_n-nearest neighbor pdf estimation
    - Nearest-neighbor rule

Parametric Approaches

- "*Curse of dimensionality*" : We need lots of training data to determine the completely unknown statistics for multi-D problems
  - A rule of thumb : *"use at least 10 times as many training samples per class as the number of features (i.e. D)"*

- Hence, with some a priori information, it is possible to estimate the *parameters* of the known distribution by using less number of samples

Maximum Likelihood Estimation (1/4)

Assume c sets of samples, drawn according to $p(x \mid \omega_j)$ which has a known parametric form.

e.g. pdf is known to be Gaussian; mean & variance values are unknown

Let $\vec{\Theta}_j$ be unknown deterministic parameter set of pdf for class-j

$$p(x \mid \omega_j) = p(x \mid \omega_j, \vec{\Theta}_j)$$: shows the dependence

Aim : Use the information provided by the observed samples to estimate the unknown parameter

Note that all sets of samples have independent pdf's,

→ there are c separate problems

Maximum Likelihood Estimation (2/4)

For an arbitrary class, let an observed sample set, X, contain n samples, $X=\{x_1,\ldots,x_n\}$.

Assume the samples are independently drawn from their density, $p(x \mid \vec{\Theta})$

The likelihood of the observed sample set, X :

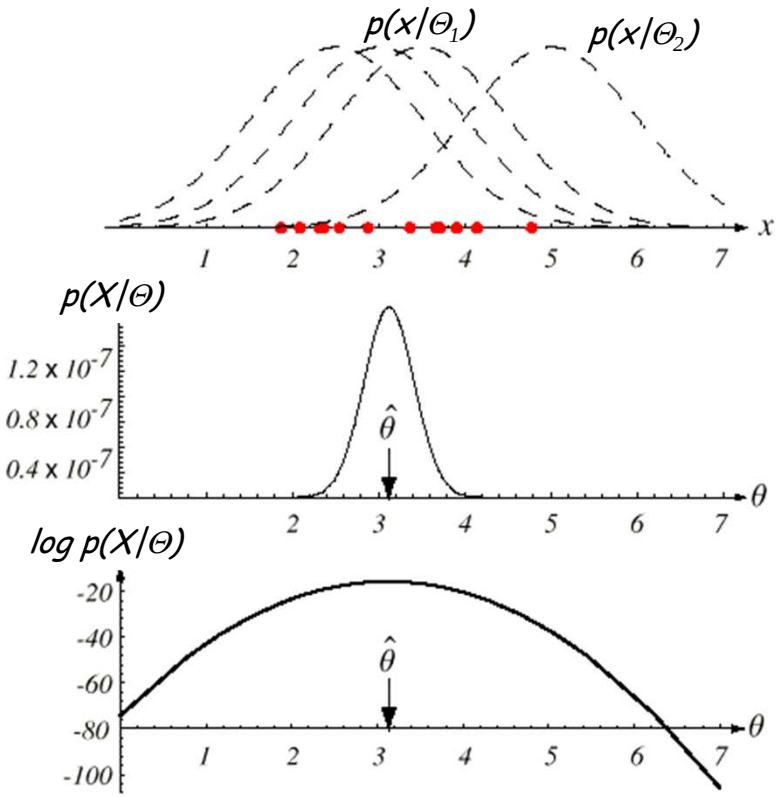
$$p(X \mid \vec{\Theta}) = \prod_{k=1}^{n} p(x_k \mid \vec{\Theta})$$

Find value of the parameter that maximizes $p(X \mid \vec{\Theta})$

→ In order to find the parameter that maximizes its value, differentiate the conditional probability and equate to zero

Maximum Likelihood Estimation (3/4)

Find value of unknown parameter maximizes $p(X \mid \vec{\Theta})$



- For different Θ, the observed samples gives different $p(X|\Theta)$ values for $p(x_k|\Theta)$ densities
- The argument for the maximum of such products is ML estimate
- $\log p(X|\Theta)$ will not differ the argument of this maxima

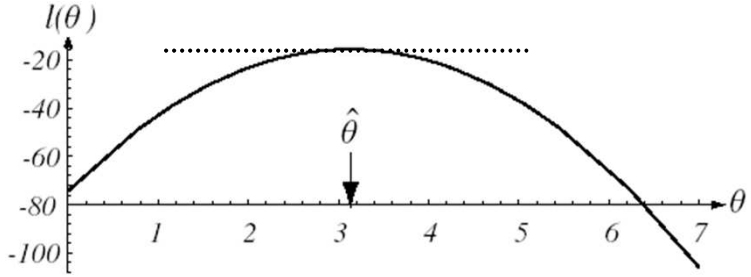
Maximum Likelihood Estimation (4/4)

Better to work with logarithm for analytical purposes.

$$l(\vec{\Theta}) = \log\, p(X \mid \vec{\Theta}) = \sum_{k=1}^{n} \log\, p(x_k \mid \vec{\Theta})$$

Note: Taking logarithm does not effect finding the maxima

Differentiate $l(\Theta)$ and equate it to zero.



$$\nabla_{\Theta} l(\vec{\Theta}) = \sum_{k=1}^{n} \nabla_{\Theta} \log\, p(x_k \mid \vec{\Theta}) = 0$$

ML Estimate of Univariate Normal :

Assume mean θ_1 & variance θ_2^2 are unknown for a Gaussian pdf:

$$\log p(x_k \mid \Theta) = -\frac{1}{2}\log\{(2\pi)\theta_2\} - \frac{1}{2\theta_2}(x_k - \theta_1)^2$$

Differentiate wrt θ_1 and θ_2 : $\nabla_\Theta \log p(x_k \mid \Theta) = \begin{bmatrix} \frac{1}{\theta_2}(x_k - \theta_1) \\ -\frac{1}{2\theta_2} + \frac{(x_k - \theta_1)^2}{2\theta_2^{\,2}} \end{bmatrix}$

Maximum likelihood estimates of the parameters :

$$\sum_{k=1}^{n} \frac{1}{\theta_2}(x_k - \theta_1) = 0 \quad \Rightarrow \quad \hat{\theta}_1 = \frac{1}{n}\sum_{k=1}^{n} x_k$$

$$-\sum_{k=1}^{n} \frac{1}{\theta_2} + \sum_{k=1}^{n} \frac{(x_k - \theta_1)^2}{\theta_2^{\,2}} = 0 \quad \Rightarrow \quad \hat{\theta}_2 = \frac{1}{n}\sum_{k=1}^{n} (x_k - \hat{\theta}_1)^2$$

ML estimates of mean and variance

ML Estimate of Multivariate Normal :

Assume only mean vector is unknown :

$$\log p(\vec{x}_k \mid \vec{\mu}) = -\frac{1}{2}\log\{(2\pi)^d \mid \Sigma \mid\} - \frac{1}{2}(\vec{x}_k - \vec{\mu})^t \Sigma^{-1}(\vec{x}_k - \vec{\mu})$$

Differentiate

$$\nabla_\mu \log p(\vec{x}_k \mid \vec{\mu}) = \Sigma^{-1}(\vec{x}_k - \vec{\mu})$$

Maximum likelihood estimate of the unknown mean vector :

$$\sum_{k=1}^{n} \Sigma^{-1}(\vec{x}_k - \vec{\mu}) = 0 \quad \Rightarrow \quad \hat{\vec{\mu}} = \frac{1}{n}\sum_{k=1}^{n} \vec{x}_k$$

MLE of mean is the arithmetic average of vector samples

Bayesian Parameter Estimation (1/3)

Can we incorporate a priori knowledge about the unknown parameters into the formulation?

Remember, Bayesian minimum error rate classifier maximizes $p(\omega_i|x)$

Assume the role of the observed sample set, X, is emphasized :

$$P(\omega_i \mid \vec{x}, X) = \frac{p(\vec{x} \mid \omega_i, X) P(\omega_i \mid X)}{\sum_{j=1}^{c} p(\vec{x} \mid \omega_j, X) P(\omega_j \mid X)}$$

Assume a priori probabilities are known : $P(\omega_i \mid X) = P(\omega_i)$

Assume sample sets of classes are independent,

→ c separate problems $p(\vec{x} \mid \omega_i, X) = p(\vec{x} \mid \omega_i, X_i)$

$= p(\vec{x} \mid X)$

Bayesian Parameter Estimation (2/3)

$$P(\omega \mid \vec{x}, X) = \frac{p(\vec{x} \mid X)P(\omega)}{\sum_{j=1}^{c} p(\vec{x} \mid \omega_j, X)P(\omega_j)}$$

Main aim is to compute $p(\vec{x} \mid X)$

$$p(\vec{x} \mid X) = \int p(\vec{x}, \Theta \mid X)d\Theta = \int \underbrace{p(\vec{x} \mid \Theta)}_{\text{form is known}} \underbrace{p(\Theta \mid X)}_{?} d\Theta$$

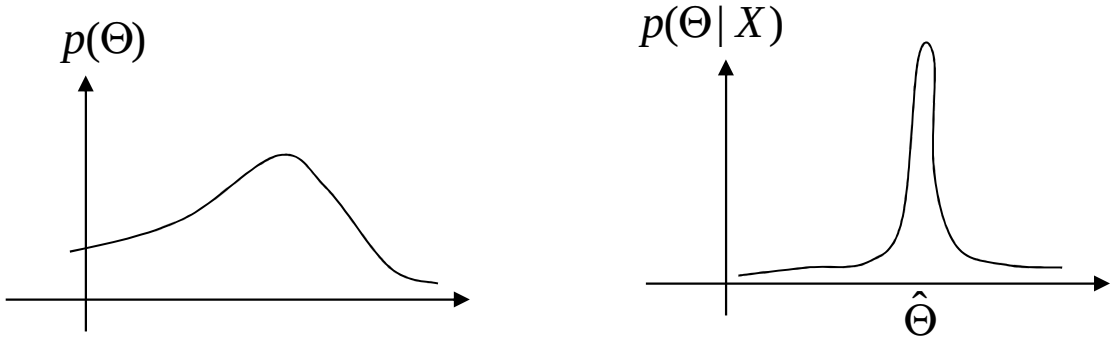
Samples are drawn independently according to $p(\vec{x} \mid \Theta)$ whose parametric form is known

Bayesian approach assumes that the unknown parameter is a random variable with a known density $p(\Theta)$

Bayesian Parameter Estimation (3/3)

If $p(\Theta|X)$ is peakly sharped at some value $\hat{\Theta}$, we obtain

$$p(\vec{x}|X)=\int p(\vec{x}|\Theta)p(\Theta|X)d\Theta \approx p(\vec{x}|\hat{\Theta})$$



If we are not sure about the value (i.e. no sharp peak), the result is the average over possible values of Θ

How to determine $p(\Theta|X)$?

For various densities, different analytical results exist

Bayesian Parameter Estimation Univariate Normal Distribution (1/3)

A univariate normal distribution with unknown μ

$$p(x|\mu) \sim N(\mu, \sigma^2)$$

A priori information about μ is expressed by density

$$p(\mu) \sim N(\mu_0, \sigma_0^{\,2})$$

Observing the sample set, D, $p(\mu|D)$ becomes

$$p(\mu|D) = \frac{p(D|\mu)\,p(\mu)}{\int p(D|\mu)\,p(\mu)\,d\mu} = \alpha \prod_{k=1}^{n} p(x_k|\mu)\,p(\mu)$$

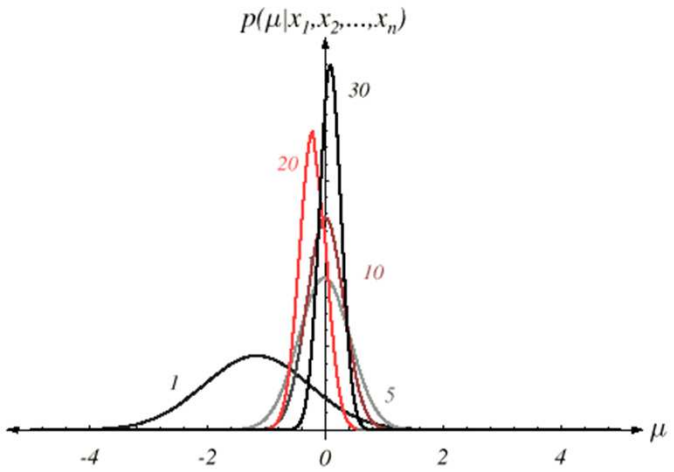
$$p(\mu|D) = \left(\alpha \prod_{k=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x_k-\mu}{\sigma}\right)^2} \right) \left(\frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2} \right)$$

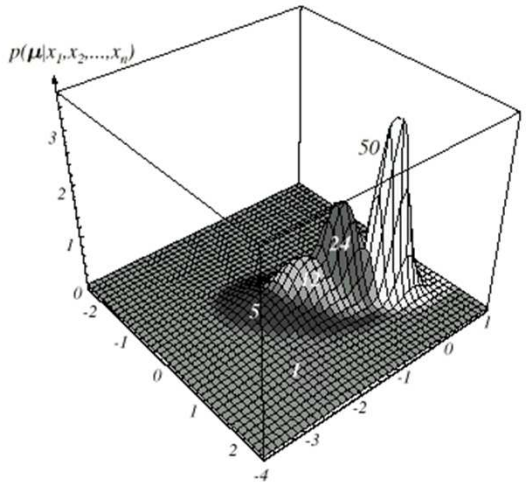
Bayesian Parameter Estimation Univariate Normal Distribution (2/3)

$$p(\mu \mid D) = \alpha' e^{-\frac{1}{2}\left(\sum_{k=1}^{n}(\frac{x_k-\mu}{\sigma})^2+(\frac{\mu-\mu_0}{\sigma_0})^2\right)} = \alpha'' e^{-\frac{1}{2}\left((\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2})\mu^2-2(\frac{1}{\sigma^2}\sum_{k=1}^{n}x_k+\frac{\mu_0}{\sigma_0^2})\mu\right)}$$

$$\Rightarrow p(\mu \mid D) \sim N(\mu_n, \sigma_n^2), \quad \mu_n = \frac{n\sigma_0^2}{n\sigma_0^2+\sigma^2}\underbrace{m_n}_{\frac{1}{n}\sum x_k} + \frac{\sigma^2}{n\sigma_0^2+\sigma^2}\mu_0; \sigma_n^2 = \frac{\sigma^2\sigma_0^2}{n\sigma_0^2+\sigma^2}$$

Increasing number of samples➔ $p(\mu|D)$ sharper peak





As $n \to \infty$, $p(\mu|D) \to \delta(\mu)$ ➔ Bayesian Learning

Bayesian Parameter Estimation Univariate Normal Distribution (3/3)

After determining $p(\mu|D)$, $p(x|D)$ is obtained by

$$p(x|D) = \int p(x|\mu)\,p(\mu|D)\,d\mu$$

$$\Rightarrow p(x|D) = \int \left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \right) \left(\frac{1}{\sqrt{2\pi}\sigma_n} e^{-\frac{1}{2}\left(\frac{\mu-\mu_n}{\sigma_n}\right)^2} \right) d\mu$$

$$\Rightarrow p(x|D) = \frac{1}{2\pi\sigma\sigma_n} e^{-\frac{1}{2}\frac{(x-\mu_n)^2}{\sigma^2+\sigma_n^2}} f(\sigma,\sigma_n)$$

$$\Rightarrow p(x|D) \sim N(\mu_n, \sigma^2+\sigma_n^2)$$

Compared to the initial knowledge, $p(x|\mu)$, about μ, $p(x|D)$ has additional uncertainty due to lack of exact knowledge of μ.

General Bayesian Learning

In summary :

- The form of the density, $p(x|\Theta)$, is assumed to be known, but the value of parameter, Θ, is unknown
- Our initial knowledge about the parameter, Θ, is assumed to be contained in a known a priori density, $p(\Theta)$.
- The rest of our knowledge about the parameter, Θ, is contained in n samples, drawn according to the unknown probability $p(x|\Theta)$

Comparison : ML vs. Bayesian

- ML avoids many assumptions and analytically easier to solve, although some estimates can be biased
- Bayesian parameter estimation permits including a priori information about the unknown, but the analytical derivations are cumbersome.
- For ordinary cases, both approaches give similar results with sufficient sample data

Non-Parametric Approaches

- Parametric approaches require
  - Knowing the form of the density
  - Finding the parameter of the density

- In many cases,
  - The form is not known
  - The form does not let you to find a unique solution (multi-modal densities)

Non-Parametric Approaches

- The solution is to use non-parametric approaches which do not assume a form

- There are 2 main directions :
  - Estimating densities non-parametrically
    - Direct estimation of density
    - Parzen window
    - k-NN estimation
  - Nearest Neighbor Rules

Non-Parametric Approaches Density Estimation (1/3)

Probability P of a vector x falling into region R :

$$P = \int_{\Re} p(\vec{x}')\, d\vec{x}'$$

N samples of x independently drawn according to $p(x)$

Probability of k independent samples fall into R (Binomial):

$$P_k = \binom{n}{k} P^k (1-P)^{n-k} \quad and \quad E[k] = nP, \quad \mathrm{var}(k) = nP(1-P)$$

Since Binomial distribution peaks very sharply around the expected value, the number of observed samples (k_{obs}) in R should be approximately equal $k_{obs} \approx E[k] = nP$

Note that probability P can be estimated via $P \approx k_{obs} / n$, but we need density, $p(x)$

Non-Parametric Approaches Density Estimation (2/3)

Assume $p(x)$ is almost constant in R : $\int_{\Re} p(\vec{x}')\,d\vec{x}' \approx p(\vec{x})V$
where V is the volume of R

Hence, one will obtain the obvious result by combining previous relations : $p(\vec{x}) \approx \frac{k_{obs}/n}{V}$

There are two approximations (≈) in previous relations

- If k (or n) goes to infinity or V goes to zero

then those approximations will converge to exact values

For finite n, fixing V and k independent of n yields problems :

- If $V \rightarrow 0$ then $p(x) \approx 0$ (useless)

Non-Parametric Approaches Density Estimation (3/3)

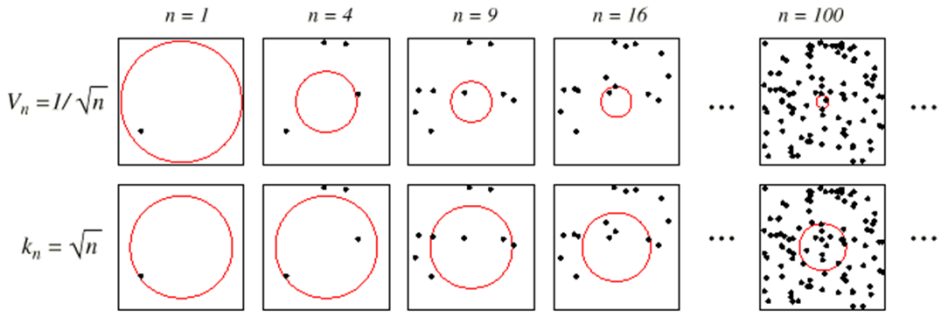
Form a sequence of regions, R_n ,centered at x for n samples

3 conditions under which $\lim_{n\to\infty}\int_{\Re} p_n(\vec{x}')\,d\vec{x}' = p(\vec{x})$ $\qquad p_n(\vec{x}) \equiv \frac{k_n / n}{V_n}$

$$(1)\quad \lim_{n\to\infty} V_n = 0 \qquad (2)\quad \lim_{n\to\infty} k_n = \infty \qquad (3)\quad \lim_{n\to\infty} \frac{k_n}{n} = 0$$

Examples that achieve these conditions :

- Parzen : Initial Vo volume is shrinking $V_n = \frac{V_0}{\sqrt{n}}$
- k-NN : R_n is grown until it contains k_n samples $k_n = \sqrt{n}$



Non-Parametric Approaches Parzen Windows (1/2)

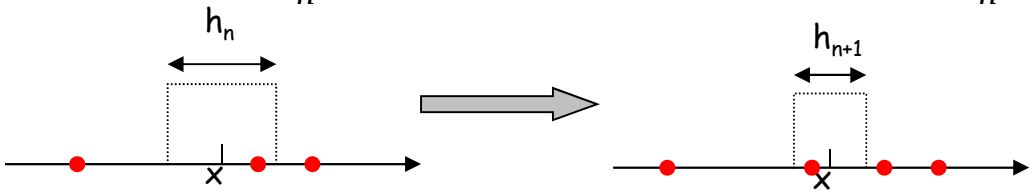
- Assume region R_n is a d-dimensional hypercube with the length of an edge as h_n

- The number of samples falling in R_n can be obtained analytically by using the *window* function :

$$\Phi(\vec{u}) = \begin{cases} 1 & |u_j| \le 1/2 \quad j = 1,\ldots,d \\ 0 & otherwise \end{cases}$$

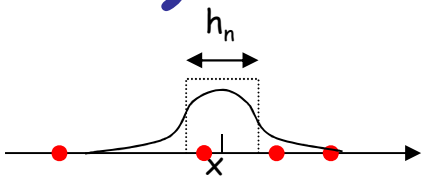
- For a hypercube (centered at x), number of samples and estimate for the density are obtained as :

$$k_n = \sum_{i=1}^{n} \Phi\left(\frac{\vec{x} - \vec{x}_i}{h_n}\right) \quad and \quad p_n(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{V_n} \Phi\left(\frac{\vec{x} - \vec{x}_i}{h_n}\right)$$

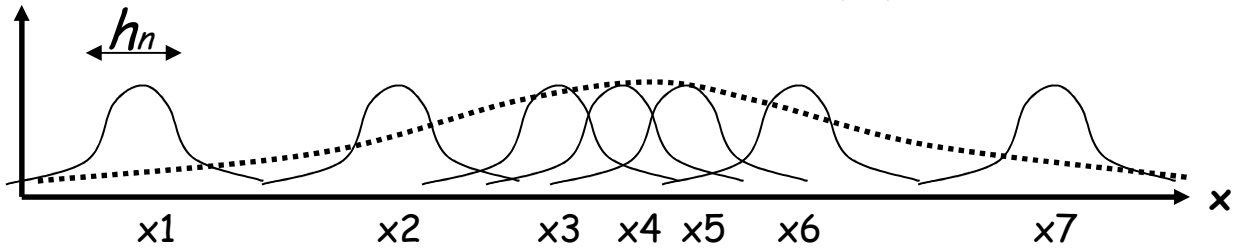


Non-Parametric Approaches Parzen Windows (2/2)

The window function can be generalized for better *interpolation* of the density : each sample contribute to the estimate based on its distance to *x*.



- If h_n is very large, then $p_n(x)$ is a superposition of slowly changing functions & an "out-of-focus" estimate
- If h_n is very small, then window function is a Dirac delta function and estimate is sum of sharp pulses

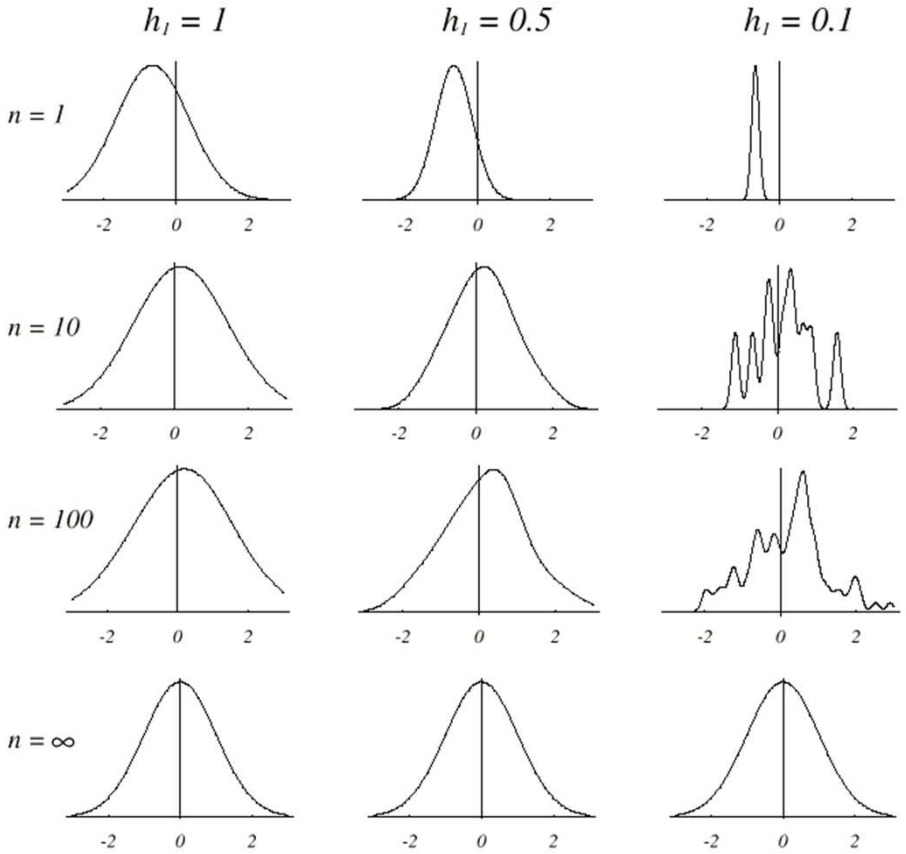


With unlimited number of samples, $p_n(x)$ converges to the unknown density for any value of h_n

With limited number of samples, the best option is to seek for an acceptable compromise

Non-Parametric Approaches
Example : Parzen Windows (1/2)

Window function : $\Phi(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$ $\quad h_n = \frac{h_1}{\sqrt{n}}$

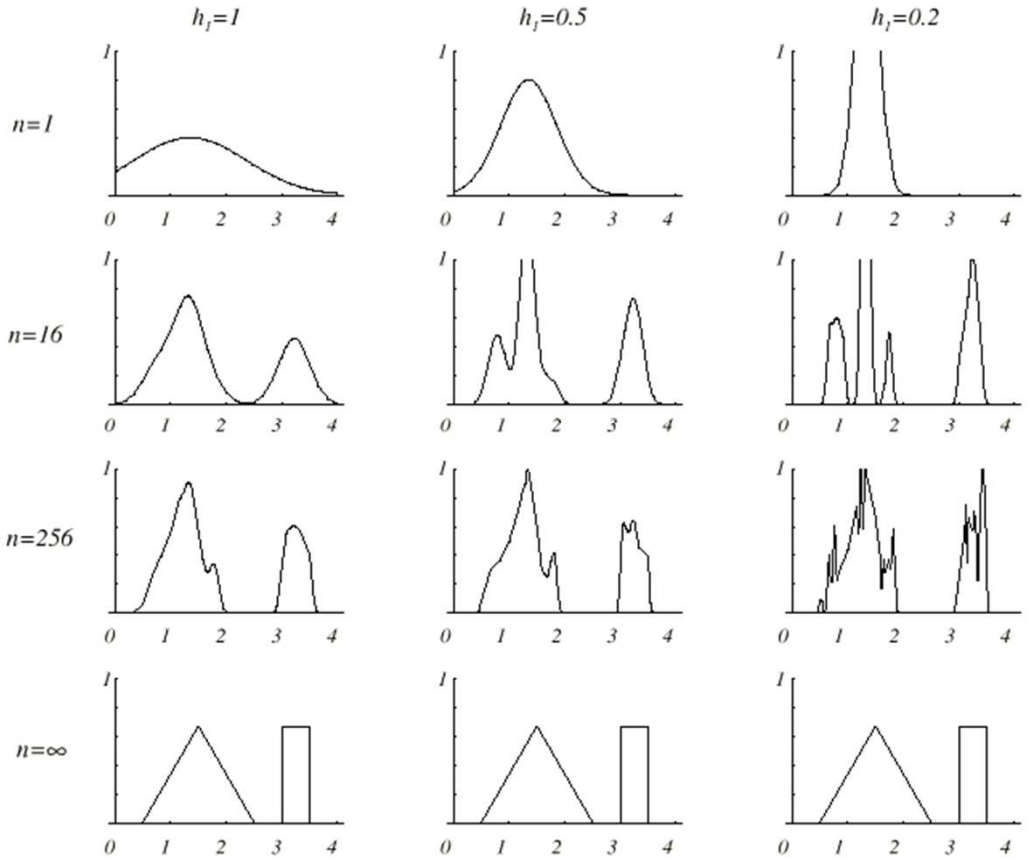


Normal density

Non-Parametric Approaches

Example : Parzen Windows (2/2)

Window function : $\Phi(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$ $\quad h_n = \frac{h_1}{\sqrt{n}}$



Bi-modal density

Non-Parametric Approaches
k_n-Nearest Neighbor

Parzen window approach depends on the initial selection of the cell volume, V

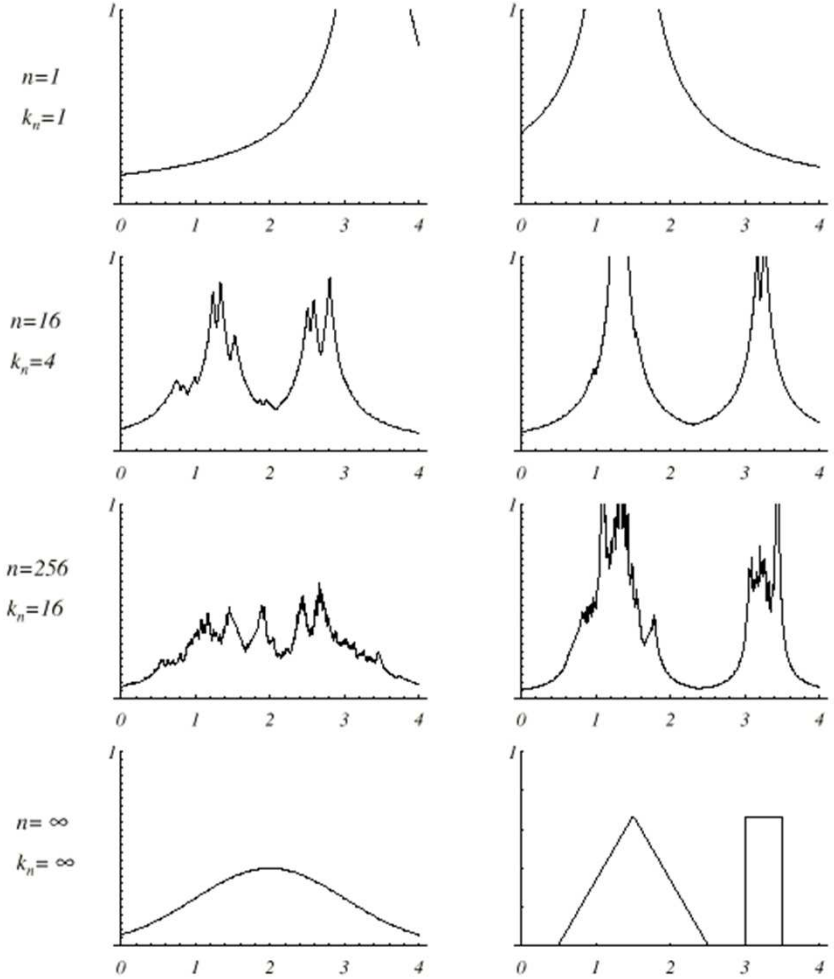
One remedy is to choose the cell volume as a function of the data, rather than an arbitrary function of number of samples

In order to estimate $p(x)$ from n samples, center a cell around x and grow until it captures k_n nearest samples (k_n is a function of n). Resulting $p(x)$: $p_n(x) = \frac{k_n / n}{V_n}$

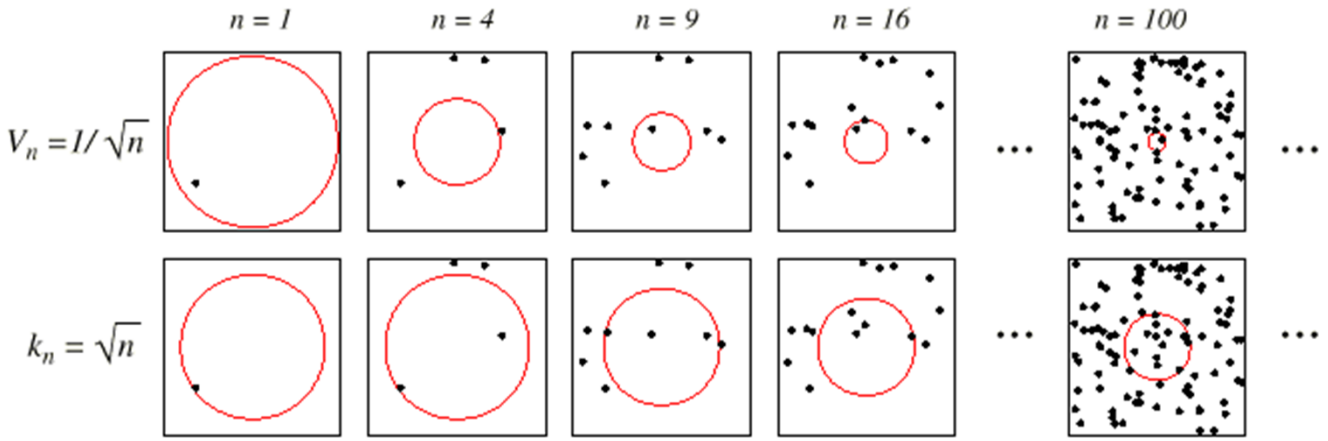
Necessary conditions for convergence :

$$\lim_{n\to\infty} k_n = \infty \quad and \quad \lim_{n\to\infty} \frac{k_n}{n} = 0 \quad (e.g.\ k_n = \sqrt{n})$$

Non-Parametric Approaches
Example : k_n-Nearest Neighbor



Non-Parametric Approaches
Parzen vs k_n-Nearest Neighbor



Both methods do converge, but it is very difficult to make meaningful statements about their finite-sample behaviour

Non-Parametric Approaches Classification Rule

All 3 methods (direct, Parzen, k_n-NN) can be used to obtain a posteriori probabilities for n-sample data

At each cell, total k samples; k_i samples for each class

$$p_n(x,\omega_i)=\frac{k_i/n}{V_n} \qquad P_n(\omega_i \mid x)=\frac{p_n(x,\omega_i)}{\sum_{j=1}^{c} p_n(x,\omega_j)}=\frac{k_i}{k}$$

Cell size selection can be achieved by using either Parzen window or k_n-NN approach

Using arbitrarily large number of samples, unknown probabilities can be obtained with optimum performance

Non-Parametric Approaches Nearest Neighbor Rule (1/3)

All 3 methods (direct, Parzen, k_n-NN) can be used to obtain a posteriori probabilities by using n-sample data so that this density is utilized for Bayes Decision Rule

A radical approach is to use the *nearest neighbor* out of the sample data to classify the unknown test data (*Nearest Neighbor Rule [NN-R]*)

While Bayes Rule (minimum-error rate) is optimal while choosing between different classes, NN-R is suboptimal

Non-Parametric Approaches Nearest Neighbor Rule (2/3)

Assume that there are unlimited number of labeled "*prototypes*" for each class

If the test point x is nearest to one of these prototypes, x' ➔ $p(w_i|x) \approx p(w_i|x')$ for all i

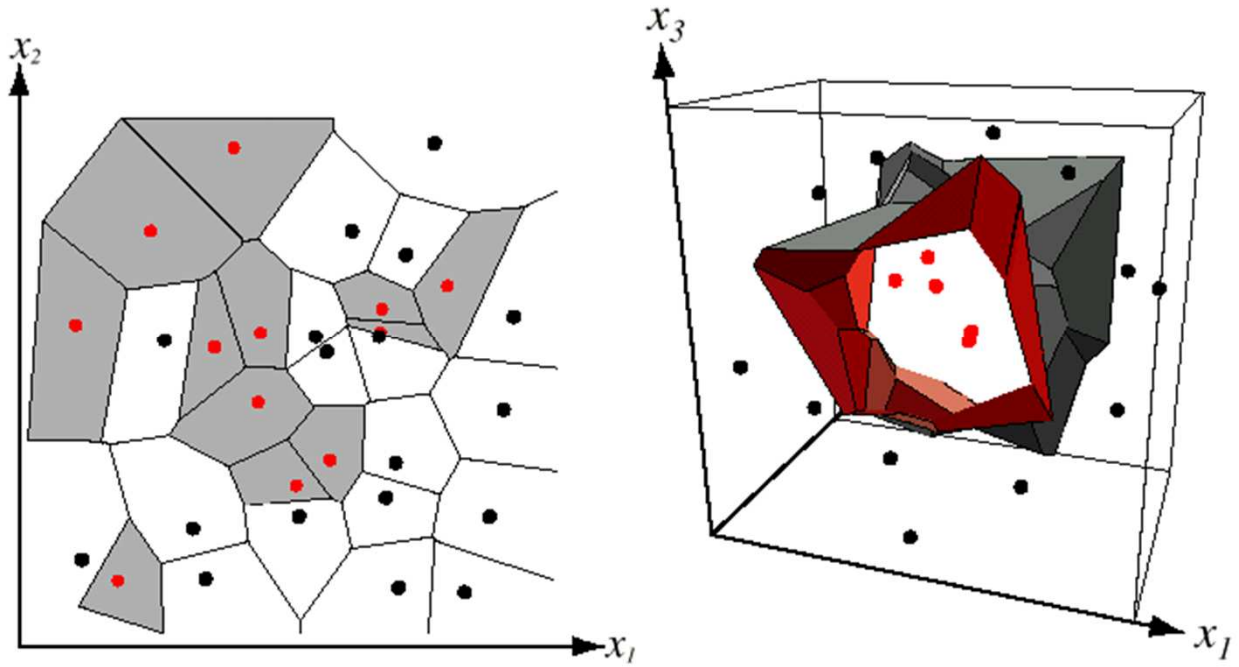
Obviously, x' labeled with m gives $p(w_m|x') > p(w_j|x')$ for all $j \neq m$

➔ one should expect $p(w_m|x) > p(w_j|x)$ for all $j \neq m$

For unlimited samples, the error rate for NN-R is <u>less than twice</u> the error rate of Bayes decision rule

Non-Parametric Approaches Nearest Neighbor Rule (3/3)

NN-rule allows to partition the feature space into cells consisting of all points closer to a given training point than any other training point (Voronoi tessellation)



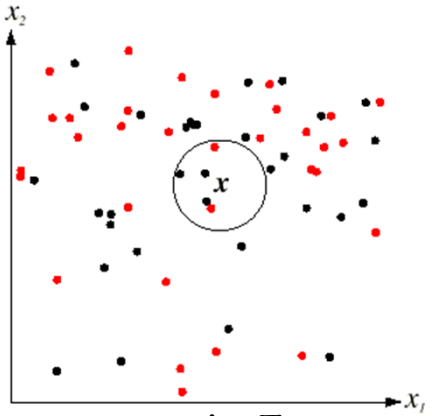
Non-Parametric Approaches
k-Nearest Neighbor Rule

A straight forward extension to Nearest Neighbor rule is using k-neighbors instead of only one.

The classification is achieved by voting k neighbors (k is usually selected as odd to avoid ties)

Selecting k requires a compromise :

- If k is too high → some of these k neighbors may have different probabilities, for finite n
- If k is too low → estimation may not be reliable



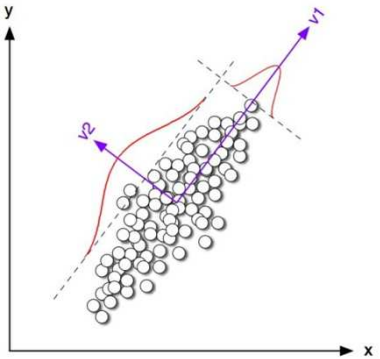
k=5

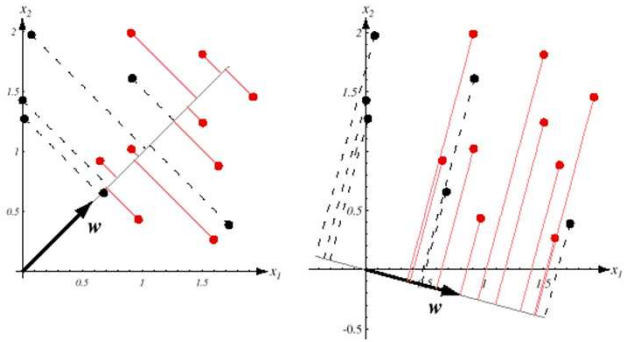
The optimal behavior is obtained as both k and n approaches to infinity.

Dimension Reduction

In supervised learning, excessive dimensionality of features should be decreased. The main approaches are

- Principal Component Analysis
    - Unsupervised
- Fisher's Linear Discriminant
    - Supervised (data with class info is required)





Principal Component Analysis

Assume there n vectors in d-dimensions: $\{ \vec{x}_1, \dots , \vec{x}_n \}$

These vectors are represented by their projections onto a line passing, e, through their sample mean, m

$$\vec{x} = \vec{m} + a \vec{e}$$

For a fixed line, the optimal a coefficients that minimize the distance between points and the line :

$$\min_{a_i \dots a_n} J(a_1, \dots , a_n, \vec{e}) = \min_{a_i \dots a_n} \sum_{k=1}^{n} \left\| (\vec{m} + a_k \vec{e}) - \vec{x}_k \right\|^2$$

$$\Rightarrow \quad J(.) = \sum_{k=1}^{n} a_k^2 \left\| \vec{e} \right\|^2 - 2 \sum_{k=1}^{n} a_k \vec{e}^{\,t} (\vec{x}_k - \vec{m}) + \sum_{k=1}^{n} \left\| \vec{x}_k - \vec{m} \right\|^2$$

$$\frac{\partial J(.)}{\partial a_k} = 0 \quad \Rightarrow \quad a_k = \vec{e}^{\,t} (\vec{x}_k - \vec{m})$$

Principal Component Analysis

Assume a coefficients are obtained; the same cost function, $J(.)$, is minimized wrt to the line direction, e

$$\min_{\vec{e}} \; J(a_1, \ldots, a_n, \vec{e}) = \min_{\vec{e}} \; \sum_{k=1}^{n} \left\| (\vec{m} + a_k \vec{e}) - \vec{x}_k \right\|^2$$

$$\text{where} \quad a_k = \vec{e}^{\,t} (\vec{x}_k - \vec{m})$$

Define *scatter matrix*, S, (similar to covariance) as

$$S \equiv \sum_{k=1}^{n} (\vec{x}_k - \vec{m})(\vec{x}_k - \vec{m})^t$$

$$\Rightarrow \; J(.) = \sum_{k=1}^{n} a_k^2 \underbrace{\left\| \vec{e} \right\|^2}_{=1} - 2 \sum_{k=1}^{n} a_k \underbrace{\vec{e}^{\,t} (\vec{x}_k - \vec{m})}_{=a_k^t} + \sum_{k=1}^{n} \left\| \vec{x}_k - \vec{m} \right\|^2$$

$$= -\sum_{k=1}^{n} \left(\vec{e}^{\,t} (\vec{x}_k - \vec{m}) \right)^2 + \sum_{k=1}^{n} \left\| \vec{x}_k - \vec{m} \right\|^2$$

$$= -\vec{e}^{\,t} S \vec{e} + \sum_{k=1}^{n} \left\| \vec{x}_k - \vec{m} \right\|^2 \quad \Rightarrow \quad \min_{\vec{e}} \; J(.) = \max_{\vec{e}} \; \vec{e}^{\,t} S \vec{e}$$

Principal Component Analysis

$$\min_{\vec{e}} \; J(.) = \max_{\vec{e}} \; \vec{e}^{\,t} S \vec{e}$$

Maximum of e^tSe must be obtained by the constraint $|e|=1$

$$\text{Lagrange mul.}: u \equiv \vec{e}^{\,t} S \vec{e} + \lambda(1 - \vec{e}^{\,t}\vec{e}) \quad \Rightarrow \quad \frac{\partial u}{\partial e} = 0 \Rightarrow 2S\vec{e} - 2\lambda\vec{e} = 0$$

Solution is equal to e which is the eigenvector of S, corresponding its largest eigenvalue

Result can be generalized to d'-dimensional projection by minimizing the following relation

$$J_{d'} = \sum_{k=1}^{n} \left\| \left(\vec{m} + \sum_{i=1}^{d'} a_{ki} \vec{e}_i \right) - x_k \right\|^2$$

where $\vec{x} = \vec{m} + \sum_{i=1}^{d'} a_i \vec{e}_i$, such that e_i's are eigenvectors

Principal Component Analysis

Remember n vectors in d-dimensions: $X = [\vec{x}_1, \ldots, \vec{x}_n]$

Note difficulty during calculation of S, if $d>>n$ (S is dxd)

$$S \equiv \sum_{k=1}^{n} (\vec{x}_k - \vec{m})(\vec{x}_k - \vec{m})^t = XX^t$$

➔ instead of solving $Se=\lambda e$ or $XX^te=\lambda e$, try solving

$$X^t X \vec{f} = \lambda \vec{f} \quad \overset{\text{multiply by } X \text{ from left}}{\Rightarrow} \quad X X^t X \vec{f} = \lambda X \vec{f}$$

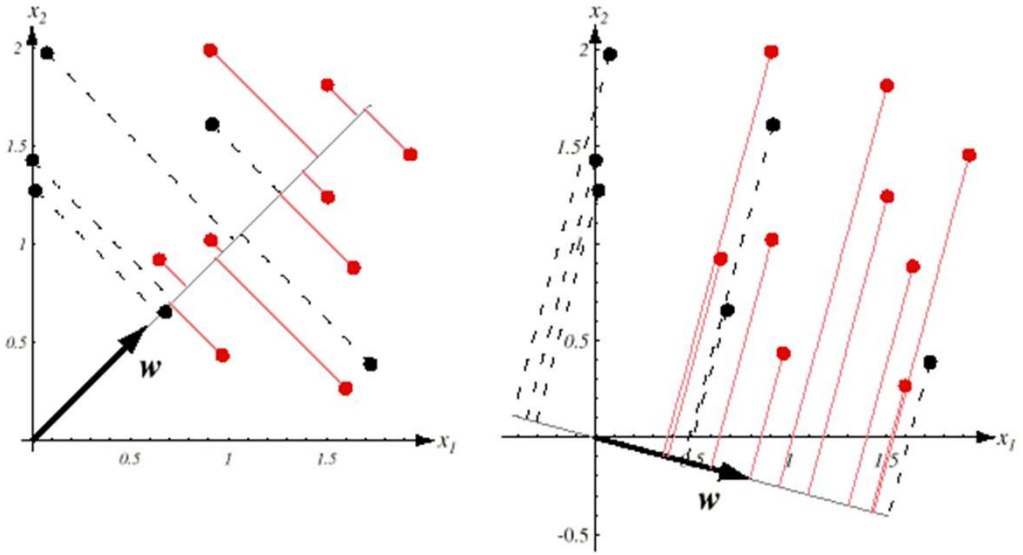
Note that XX^t *is* dxd, *whereas* X^tX *is* nxn

$$X X^t (X \vec{f}) = \lambda (X \vec{f}) \Leftrightarrow X X^t \vec{e} = \lambda \vec{e}$$

$$\Rightarrow X \vec{f} = \vec{e}$$

Fisher's Linear Discriminant (1/8)

- The Fisher's approach aims to project d-dimensional data onto a line (1-D), which is defined by w
- The projected data is expected to be well separated between two classes after such a dimension reduction



Fisher's Linear Discriminant (2/8)

- Feature vector projections : $y_i = \vec{w}^t \vec{x}_i \quad i = 1, \ldots, n$
- Measures for separation based on w :
  - Difference between projection means
  - Variance of within-class projection data
- Choose projection (w) in order to maximize J

$$J(\bullet) = \frac{(m_1 - m_2)^2}{\bar{s}_1^2 + \bar{s}_2^2}$$

where m_i : projection means for class i

$$s_i^2 = \sum_{y \in Y_i} (y - m_i)^2 : \text{scatter}$$

Fisher's Linear Discriminant (3/8)

- Relation between sample & projection means :

$$\vec{m}_i = \frac{1}{n_i}\sum_{x\in\aleph_i}\vec{x} \quad \Rightarrow \quad m_i = \frac{1}{n_i}\sum_{y\in Y_i} y = \frac{1}{n_i}\sum_{x\in\aleph_i}\vec{w}^t\vec{x} = \vec{w}^t\vec{m}_i$$

- Define scatter matrices S_i

$$S_i = \sum_{x\in\aleph_i}(\vec{x}-\vec{m}_i)(\vec{x}-\vec{m}_i)^T \quad and \quad S_W = S_1 + S_2$$

- Note that s_i and S_i are related as

$$s_i^2 = \sum_{y\in Y_i}(y-m_i)^2 = \sum_{x\in\aleph_i}\left(\vec{w}^T\vec{x}-\vec{w}^T\vec{m}_i\right)^2$$

$$= \sum_{x\in\aleph_i}\vec{w}^T(\vec{x}-\vec{m}_i)(\vec{x}-\vec{m}_i)^T\vec{w} = \vec{w}^T S_i\vec{w}$$

Fisher's Linear Discriminant (4/8)

- Similarly, the relation between m_1 and m_2 becomes

$$(m_1 - m_2)^2 = (\vec{w}^T \vec{m}_1 - \vec{w}^T \vec{m}_2)^2 = \vec{w}^T (\vec{m}_1 - \vec{m}_2)(\vec{m}_1 - \vec{m}_2)^T \vec{w}$$

$$= \vec{w}^T S_B \vec{w} \quad \text{(Note that } S_B \text{ has rank 1)}$$

- The initial criterion function : $J(\bullet) = \dfrac{(m_1 - m_2)^2}{\bar{s}_1^2 + \bar{s}_2^2}$

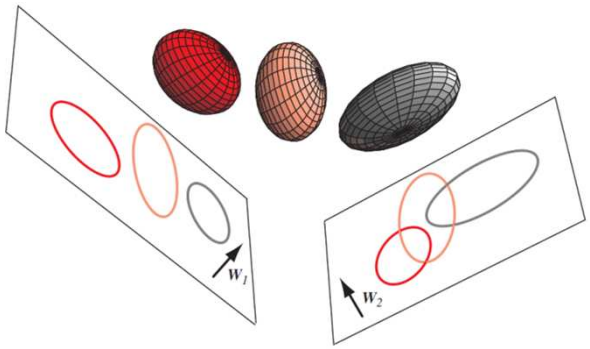
➔ This function can be written as $J(\vec{w}) = \dfrac{\vec{w}^T S_B \vec{w}}{\vec{w}^T S_W \vec{w}}$

- w vector maximizes J must satisfy $S_B \vec{w} = \lambda S_W \vec{w}$
  (see distributed notes for its proof)
- If S_W is non-singular, then

$$S_W^{-1} \underbrace{S_B \vec{w}}_{\substack{direction \\ \vec{m}_1 - \vec{m}_2}} = \lambda \vec{w} \qquad \Rightarrow \qquad \vec{w} = S_W^{-1}(\vec{m}_1 - \vec{m}_2)$$

Fisher's Linear Discriminant (5/8)

- For a 2-class problem, d-dimensional data is projected on a line
- As an extension to c-class problem, it is possible to project data onto $(c\text{-}1)$-dimensions, instead of a line.



- For $(c\text{-}1)$-dimensions :

$$y_i = \vec{w}_i^T \vec{x}\ , i = 1,\ldots c-1 \quad \Rightarrow \quad \vec{y} = W^T \vec{x}$$

- Define new scatter matrices in d-dimensional space

$$S_T = \sum_{\vec{x} \in Whole} (\vec{x} - \vec{m})(\vec{x} - \vec{m})^T \quad , \quad S_W = \sum_{i=1}^{c} S_i$$

$$= \sum_{i=1}^{c} \sum_{\vec{x} \in D_i} (\vec{x} - \vec{m}_i + \vec{m}_i - \vec{m})(\vec{x} - \vec{m}_i + \vec{m}_i - \vec{m})^T$$

$$= S_W + S_B \quad \text{where} \quad S_B = \sum_{i=1}^{c} n_i (\vec{m}_i - \vec{m})(\vec{m}_i - \vec{m})^T$$

(Note that S_B has rank c-1)

Fisher's Linear Discriminant (6/8)

- Remember the samples are projected by $\vec{y} = W^T \vec{x}$
- Resulting projected mean vectors in $(c\text{-}1)$-dimensions :

$$\tilde{\vec{m}}_i = \frac{1}{n_i} \sum_{\vec{y} \in Y_i} \vec{y} \quad , \quad \tilde{\vec{m}} = \frac{1}{n} \sum_{i=1}^{c} n_i \tilde{\vec{m}}_i$$

$$\Rightarrow \quad \tilde{\vec{m}}_i = \frac{1}{n_i} \sum_{x \in \aleph_i} W^T \vec{x} = W^T \vec{m}_i \quad , \quad \tilde{\vec{m}} = \frac{1}{n} \sum_{i=1}^{c} n_i W^T \vec{m}_i = W^T \vec{m}$$

- Scatter matrices in $(c\text{-}1)$-dimensions can defined as

$$\tilde{S}_W = \sum_{i=1}^{c} \sum_{\vec{y} \in Y_i} \left(\vec{y} - \tilde{\vec{m}}_i \right) \left(\vec{y} - \tilde{\vec{m}}_i \right)^T, \quad \tilde{S}_B = \sum_{i=1}^{c} n_i \left(\tilde{\vec{m}}_i - \tilde{\vec{m}} \right) \left(\tilde{\vec{m}}_i - \tilde{\vec{m}} \right)^T$$

Fisher's Linear Discriminant (7/8)

- Scatter matrices in the projected space are

$$\tilde{S}_W = \sum_{i=1}^{c} \sum_{\vec{y} \in Y_i} \left(\vec{y} - \tilde{\vec{m}}_i\right)\left(\vec{y} - \tilde{\vec{m}}_i\right)^T, \quad \tilde{S}_B = \sum_{i=1}^{c} n_i \left(\tilde{\vec{m}}_i - \tilde{\vec{m}}\right)\left(\tilde{\vec{m}}_i - \tilde{\vec{m}}\right)^T$$

- Relation between scatter matrices are equal to

$$\tilde{S}_W = \sum_{i=1}^{c} \sum_{\vec{y} \in Y_i} \left(\vec{y} - \tilde{\vec{m}}_i\right)\left(\vec{y} - \tilde{\vec{m}}_i\right)^T$$

$$= \sum_{i=1}^{c} \sum_{\vec{x} \in \aleph_i} \left(W^T \vec{x} - W^T \vec{m}_i\right)\left(W^T \vec{x} - W^T \vec{m}_i\right)^T = W^T S_W W,$$

$$\tilde{S}_B = \sum_{i=1}^{c} n_i \left(\tilde{\vec{m}}_i - \tilde{\vec{m}}\right)\left(\tilde{\vec{m}}_i - \tilde{\vec{m}}\right)^T$$

$$= \sum_{i=1}^{c} n_i \left(W^T \vec{m}_i - W^T \vec{m}\right)\left(W^T \vec{m}_i - W^T \vec{m}\right)^T = W^T S_B W$$

Fisher's Linear Discriminant (8/8)

- Relation between scatter matrices are obtained as

$$\tilde{S}_W = W^T S_W W, \quad \tilde{S}_B = W^T S_B W$$

- For better discrimination in the projected space:

$$\min |\tilde{S}_W| \;\&\; \max |\tilde{S}_B| \qquad |.| : \text{determinant}$$

$$\Rightarrow \quad J(\bullet) = \frac{|\tilde{S}_B|}{|\tilde{S}_W|} \quad \Rightarrow J(W) = \frac{|W^T S_B W|}{|W^T S_W W|}$$

Note that determinant is product of scatter along principal directions

- Solution for $J(W)$: Columns of the optimal W are generalized $(c\text{-}1)$ eigenvectors that correspond to the largest eigenvalues of $S_B \vec{w}_i = \lambda_i S_W \vec{w}_i$