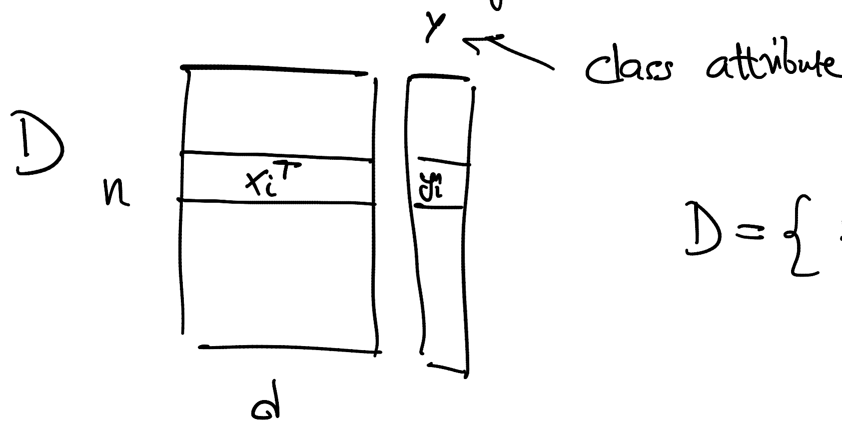
Lecture 6

Thursday, September 14, 2023 9:55 AM

Linear Discriminant Analysis



$D = \{ x_i^T, \; y_i \}$

class

$y_i \in \{ C_1, C_2, \ldots, C_K \}$

y_i is binary

$y_i \in \{ P, N \}$

$+1, -1$

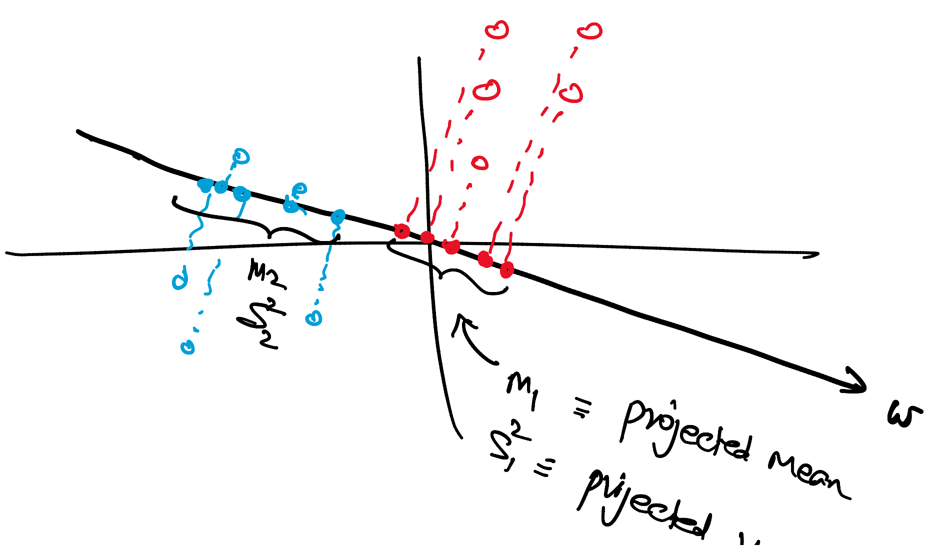
$0 \quad 1$

D_1 $\quad$ D_2

$\{ x_i^T : y_i = C_1 \}$ $\quad$ $\{ x_i^T : y_i = C_2 \}$

Q: find $\vec{\omega}$ to separate the two classe

$\mathbb{R}^d \Longrightarrow \mathbb{R}$



$s_1 \equiv$ projected variance ... Mean

projected variance $(s_1^2 = n_1 \cdot \sigma_1^2)$

$S_i^2 =$ Scatter

total squared deviation

$\sigma_i^2 \equiv$ variance

avg. squared deviation

$$\max_{\vec{\omega}} J \equiv \frac{(m_1 - m_2)^2}{s_1^2 + s_2^2}$$

$x_i^T \in D_1$

$x_i^T \in D_2$

$\vec{\omega}$

$$a_i = \frac{\vec{\omega}^T x_i}{\vec{\omega}^T \vec{\omega}} = \vec{\omega}^T x_i$$

projected point

$\vec{\omega}^T \vec{\omega} = 1$

$$m_1 = \frac{1}{n_1} \sum_{x_i \in D_1} a_i = \frac{1}{n} \sum_{x_i \in D_1} \omega^T x_i = \omega^T \left(\frac{1}{n} \sum_{x_i \in D_1} x_i \right)$$

$$m_1 = \vec{\omega}^T \vec{\mu}_1$$

projected Mean

mean of D_1

$$m_2 = \vec{\omega}^T \vec{\mu}_2$$

$\vec{\mu}_1 \equiv \text{mean}(D_1) \in \mathbb{R}^d$

$\vec{\mu}_2 \equiv \text{mean}(D_2) \in \mathbb{R}^d$

$\vec{\omega} \in \mathbb{R}^d$

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

$$= \left(\vec{\omega}^T (\vec{\mu}_1 - \vec{\mu}_2)\right)^2$$

$$= \vec{\omega}^T \boxed{(\vec{\mu}_1 - \vec{\mu}_2)(\mu_1 - \mu_2)^T} \vec{\omega}$$

$$= \boxed{\vec{\omega}^T \left(\quad B \quad \right) \vec{\omega}}$$

$B \equiv$ between class $\underline{\underline{\text{scatter}}}$

outer product of the difference vector

$(\vec{\mu}_1 - \vec{\mu}_2)$ with itself

$$\underbrace{s_1^2}_{\text{projected space}} = \text{total deviation} = \sum_{x_i \in D_1} (a_i - m_1)^2$$

$$= \sum_{x_i \in D_1} \left(\vec{\omega}^T \vec{x}_i - \vec{\omega}^T \vec{\mu}_1\right)^2$$

$$= \sum_{D_1} \left(\vec{\omega}^T (x_i - \vec{\mu}_1)\right)^2$$

$$= \vec{\omega}^T \left(\underbrace{\sum (\vec{x}_i - \vec{\mu}_1)(\vec{x}_i - \mu_1)^T}_{S_1 = n_1 \cdot \Sigma_1} \right) \vec{\omega}$$

$$s_2^2 = \vec{\omega}^T (S_2) \vec{\omega}$$

$$= \vec{\omega}^T (n_2 \Sigma_2) \vec{\omega}$$

$$s_1^2 + s_2^2 = \vec{\omega}^T \left(\underline{S_1} + \underline{S_2} \right) \vec{\omega}$$

$$= \vec{\omega}^T S \vec{\omega}$$

pooled within-class scatter

$$\max_{\vec{\omega}} \quad \boxed{J = \frac{\vec{\omega}^T B \vec{\omega}}{\vec{\omega}^T S \vec{\omega}}} = \frac{f}{g} \qquad \frac{f'g - g'f}{g^2}$$

$$\nabla_{\vec{\omega}} = \frac{\partial J}{\partial \vec{\omega}} = \frac{2B\vec{\omega}(\vec{\omega}^T S\vec{\omega}) - (2S\vec{\omega})(\vec{\omega}^T B\omega)}{(\vec{\omega}^T S\vec{\omega})^2} = 0$$

gradient

(direction where J increases w.r.t $\vec{\omega}$)

$$\Rightarrow \quad B\vec{\omega}(\vec{\omega}^T S\omega) = S\vec{\omega}(\vec{\omega}^T B\vec{\omega})$$

divide by $\vec{\omega}^T S\vec{\omega}$ on both sides

$$\Rightarrow \quad B\vec{\omega} = \boxed{\left(\frac{\vec{\omega}^T B\omega}{\vec{\omega}^T S\omega}\right)} S\vec{\omega}$$

$J = \lambda$

$$\Rightarrow \quad B\vec{\omega} = J\, S\vec{\omega}$$

$$\Rightarrow \quad \boxed{B\vec{\omega} = \lambda S\vec{\omega}}$$

generalized eigen equation

$B\vec{\omega} = \lambda\vec{\omega}$

$B\vec{\omega} = \lambda \boxed{S} \vec{\omega}$

If S^{-1} exists

$$\Rightarrow \quad S^{-1}B\vec{\omega} = \lambda S^{-1}S\omega$$

$B\vec{\omega} = \lambda \boxed{S} \vec{\omega}$

$$\Rightarrow \quad S^{-1} B \vec{\omega} = \lambda \underline{\underline{S^{-1} S}}\, \omega$$

$$\Rightarrow \quad \boxed{(S^{-1} B)\, \vec{\omega} = \lambda\, \omega}$$

<u>Solution:</u> the optimal direction is the <u>dominant</u> eigenvector of $S^{-1}B$ matrix

$\Downarrow$

why? because $\max_{\vec{\omega}} J$

$\lambda_1 = J$ is the optimal <u>objective value</u>

Simpler solution (w/o eigenvalues)

$B\vec{\omega} = \lambda S \vec{\omega}$ solve for $\vec{\omega}$

<u>Sol 1</u>: $S^{-1}B$ and find dominant <u>eigenvector</u> $\vec{\omega}$

<u>Sol 2</u>

$$B = \underbrace{\underbrace{(\vec{\mu}_1 - \vec{\mu}_2)(\vec{\mu}_1 - \vec{\mu}_2)^T}_{\vec{\delta} \cdot \vec{\delta}^T}}_{\text{Outer product}}$$

$$B\vec{\omega} = \vec{\delta}\,\underbrace{\left(\vec{\delta}^{T}\vec{\omega}\right)}_{\alpha}$$

α ← Some scalar

$$\underbrace{B\vec{\omega}} = \lambda S \vec{\omega}$$

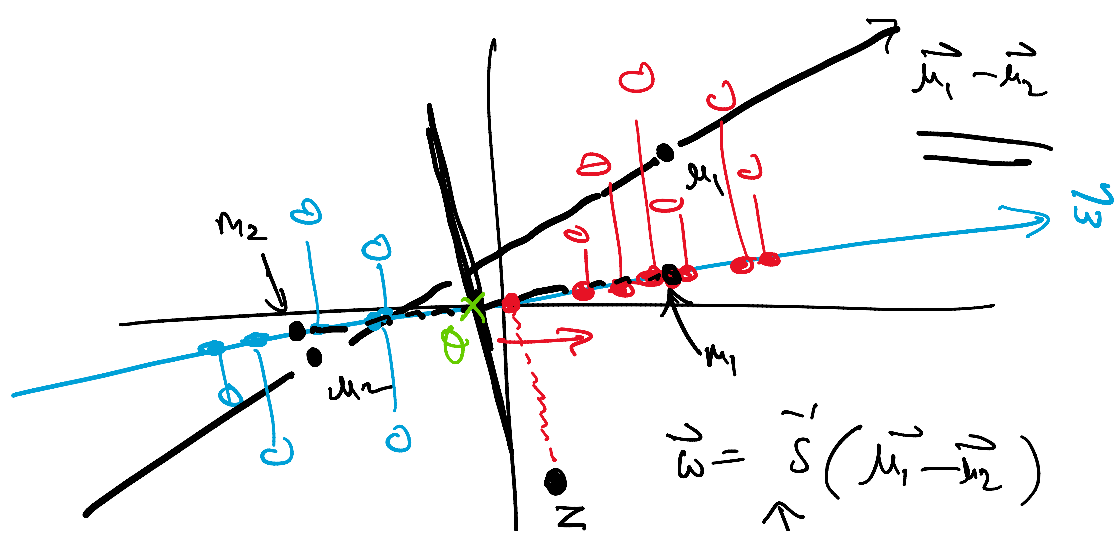
$$\alpha \vec{\delta} = \lambda S \vec{\omega}$$

$$\left(\vec{\mu}_1 - \vec{\mu}_2\right) = \frac{\lambda}{\alpha} S \vec{\omega}$$

$$\boxed{\cancel{\frac{\alpha}{\lambda}}\, S^{-1}\left(\vec{\mu}_1 - \vec{\mu}_2\right)} = \vec{\omega}$$

$$\vec{\omega} = S^{-1}\left(\vec{\mu}_1 - \vec{\mu}_2\right)$$

the normalize $\vec{\omega}$ to be a unit vector!



$\vec{w} = S(\mu_1 - \mu_2)$

rotated/scaled difference vector

Solution: $\vec{w}$: Linear Discriminant Direction

Use to classify : what is the label for $\vec{z}$ (test point)

classifier: $\theta = \frac{1}{2}(m_1 - m_2) = \frac{1}{2}\vec{w}^T(\vec{\mu}_1 - \vec{\mu}_2)$

$$M(\vec{z}) = \begin{cases} \text{red} & \text{if } w^T z \geq \theta \\ \text{blue} & \text{if } \vec{w}^T z < \theta \end{cases}$$

Can be generalized to k classes = $\{C_1\, C_2 \cdots C_k\}$

Solution: k eigenvector of $S^{-1}B'$

gradient descent algorithm

$$\min_{\vec{z}} J = \sum_{i=1}^{n} \|x_i - z\|^2$$

find an estimate for the mean

$\nabla_{\vec{z}} = \frac{\partial J}{\partial \vec{z}} = \vec{0}$

~~Solution~~ $= \underbrace{\vec{z} = \frac{1}{n}\sum_{i=1}^{n} x_i}$

closed-form

n is extremely large

$\vec{\nabla}_z$

$\nwarrow$ gradient vector $\underbrace{\text{for all } n \text{ points}}_{\text{batch gradient descent}}$ all of D

$\nabla_z \longrightarrow B = n \leftarrow$ I.

batch size

$$J = \sum_{i=1}^{B} \frac{1}{2} \|x_i - z\|^2$$

$B = 1$ II.

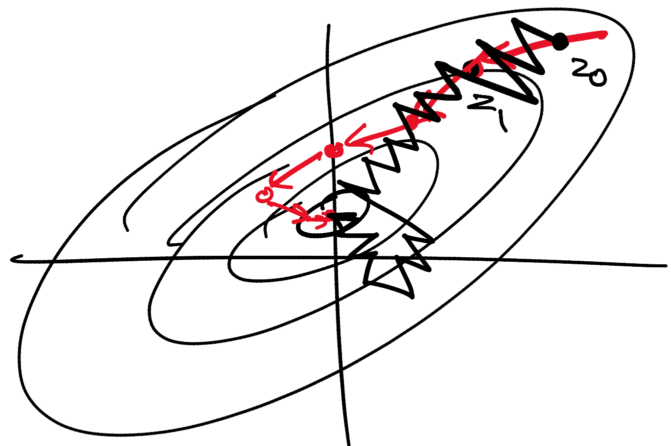
Stochastic gradient descent

$$J = \frac{1}{2} \|x_i - z\|^2$$

Some random point from D

$\vec{\nabla}_z \leftarrow$ gradient by looking at a single random point

$$\vec{z}_t = \vec{z}_{t-1} \underbrace{- \eta \cdot \vec{\nabla}_z}_{\text{opposite to the gradient by a small step } \eta}$$



when η is very large

Q1 how many iterations? t

Q2, how much does each iteration cost?

Batch Gradient descent:

$B = n$ (full dataset)

$\vec{\nabla}_z = O(n)$

total cost: $O(n) \cdot T$ $\qquad$ T : total # of steps

$T = 10$

$\boxed{10 \cdot O(n)}$

$n = 10^9$

total $= 10^{10}$

SGD: $O(1)$ per update

T is going to be large but $T = 10^4$

T is going to be large but $T = 10^4$

total cost = 10^4

mini-batch SGD

↳ $B = 128$

$$\boxed{\frac{\partial J}{\partial \vec{z}} = \vec{\nabla}_z = z - \vec{x}_i} \leftarrow B = 1$$

$$J = \frac{1}{2} \| x_i - z \|^2$$

$$J = \frac{1}{2} \sum_{i=1}^{B} \| x_i - \vec{z} \|^2 \leftarrow B$$

$$\vec{\nabla}_z = \frac{\partial J}{\partial \vec{z}} = \sum_{i=1}^{B} (z - x_i) = B \cdot z - \sum_{i=1}^{B} x_i$$

$$\vec{z}_{t+1} = \vec{z}_t - \eta \vec{\nabla}_z$$

$$\| \vec{z}_{t+1} - \vec{z}_t \| \leq \varepsilon$$ Stop!

$\varepsilon = 10$

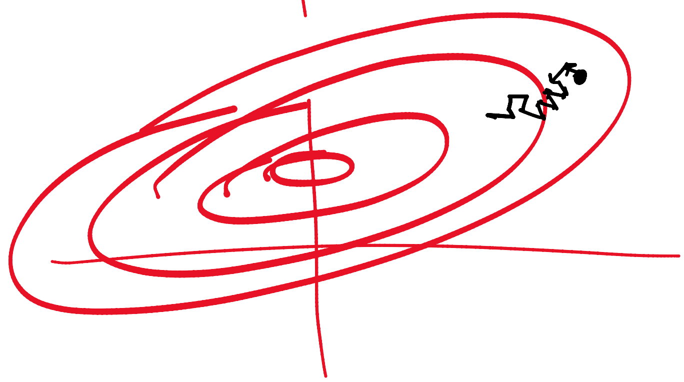
large η

10^{-2}

10^{-2}

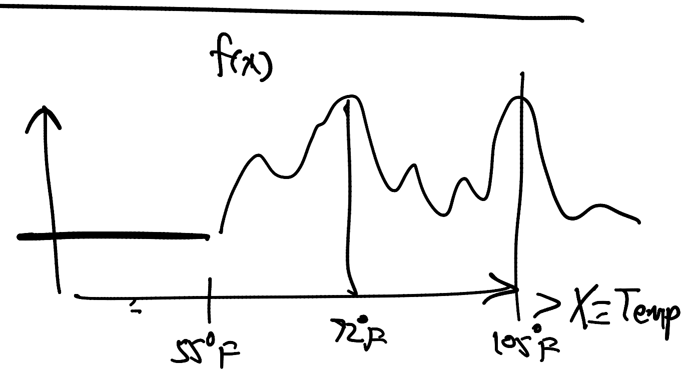
very small η

✓



Probability Distributions

Normal distribution



$\mu \leftarrow$ true mean

Unknown prob density function

Sample 1

$$\left\{ \begin{aligned} x_1 &= 71 \\ x_2 &= 65.5 \\ x_3 &= 60.2 \end{aligned} \right. \Rightarrow \hat{\mu}_1$$

$\text{Sample}_2 \Rightarrow \hat{\mu}_2$

$\vdots$

$\text{Sample}_k \Rightarrow \hat{\mu}_k$

$$\mu \in \underbrace{(lb, ub)}_{\text{99\% interval}}$$

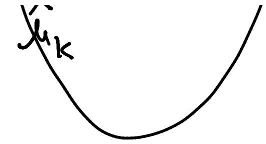
true mean

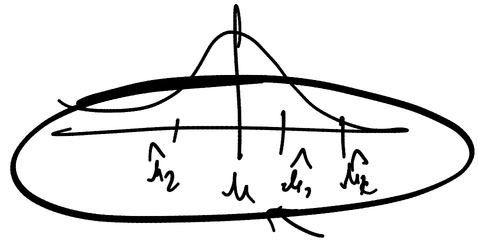
Confidence

true mean (unknown)

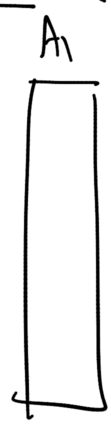
Confidence interval

Sample $k \Rightarrow \hat{\mu}_k$





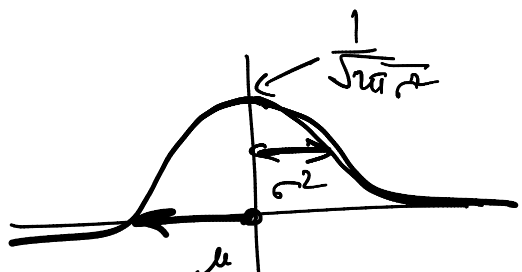
Univariate (1d)



$$f(x \mid \mu, \sigma^2) = \boxed{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}}$$

Σ

$\underbrace{\text{mean} \quad \text{variance}}_{\text{parameters}}$



d-dim (Multivariate)

$$\boxed{f(\vec{x} \mid \vec{\mu}, \Sigma) = \frac{1}{(\sqrt{2\pi})^d (\det(\Sigma))^{1/2}} e^{-\frac{(\vec{x}-\vec{\mu})^T \Sigma^{-1} (\vec{x}-\vec{\mu})}{2}}}$$

$\mathbb{R}^d \quad \mathbb{R}^d \quad \mathbb{R}^{d\times d}$

$\Sigma^{-1} \equiv$ Connection to PCA