Outline

Recap LDA/QDA
Covariance Estimation
Naive Bayes

$$\arg\max_k \; P(Y=k)\cdot P(X \mid Y=k)$$

Gaussian density μ_k, Σ_k

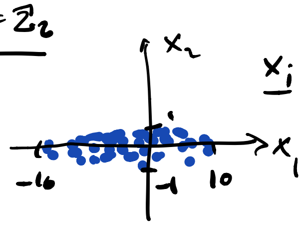
(LDA/QDA)
$$= \arg\max_k \; P(Y=k)\cdot \phi(x \mid Y=k)$$

$$= \arg\max_k \; \log \hat{\pi}_k - \underbrace{\frac{1}{2}(x-\hat{\mu}_k)^T \cdot \hat{\Sigma}_k^{-1}(x-\hat{\mu}_k)}_{\frac{1}{2}\cdot \text{Mahalanobis distance squared}}$$

$$= \frac{1}{2}\,\left\| \hat{\Sigma}_k^{-1/2}(x-\hat{\mu}_k)\right\|_2^2$$

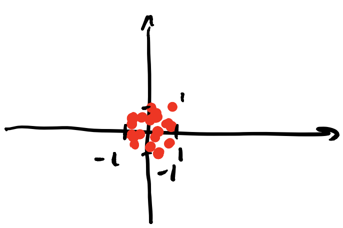
$\Sigma_1 = \Sigma_2$

$x_i \quad i=1,\dots,n$



$$\hat{\Sigma} = \begin{bmatrix} 100 & 0 \\ 0 & 1 \end{bmatrix} \qquad \hat{\Sigma}^{-1/2} = \begin{bmatrix} \frac{1}{10} & \\ & 1 \end{bmatrix}$$

$$\tilde{x}_i = \hat{\Sigma}^{-1/2}\cdot x_i \qquad i=1,\dots,n$$



LDA $\quad \Sigma_i = \Sigma_j \;\; \forall\, i,j \qquad \hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(x-\hat{\mu}_{y_i})(x-\hat{\mu}_{y_i})^T$

QDA $\quad \Sigma_1 \neq \Sigma_2 \neq \dots \qquad \hat{\Sigma}_k = \frac{1}{n_k}\cdot\sum_{i:\,y_i=k}(x-\hat{\mu}_k)(x-\hat{\mu}_k)^T$

Example in 1D

2 classes

$$\hat{\Sigma}_k = \hat{\sigma}_k^2 = \frac{1}{n}\sum_{\substack{i=1 \\ i: y_i = k}}^{n} (x_i - \mu_{y_i})^2$$

$$\log \hat{\pi}_1 - \frac{1}{2}\frac{(x-\hat{\mu}_1)^2}{\hat{\sigma}_1^2} \underset{2}{\overset{1}{\gtrless}} \log \hat{\pi}_2 - \frac{1}{2}\frac{(x-\hat{\mu}_2)^2}{\hat{\sigma}_2^2}$$

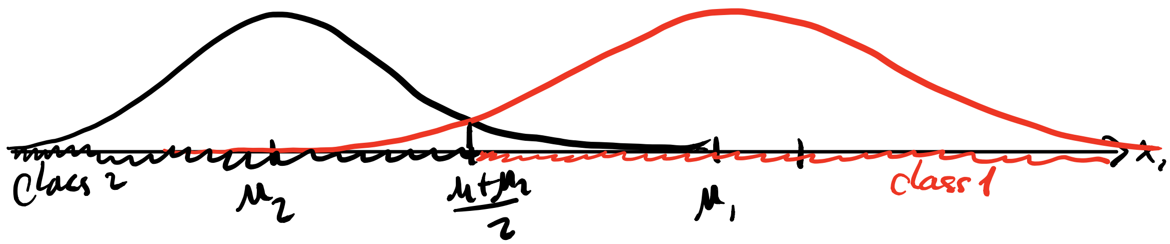
LDA

$\hat{\sigma}_1 = \hat{\sigma}_2 = \hat{\sigma}$

$\hat{\pi}_1 = \hat{\pi}_2$

$$\underbrace{\log \hat{\pi}_1 - \log \hat{\pi}_2}_{0} + \frac{x}{\sigma^2}(\mu_1 - \mu_2) - \frac{1}{2\sigma^2}(\mu_1^2 - \mu_2^2) \underset{2}{\overset{1}{\gtrless}} 0$$

$$-\frac{1}{2}\log \sigma_1^2 + \frac{1}{2}\log \sigma_2^2$$



$$\cancel{(\mu_1 - \mu_2)} \cdot \frac{x}{\cancel{\sigma^2}} > \frac{\cancel{(\mu_1 - \mu_2)}(\mu_1 + \mu_2)}{2\cancel{\sigma^2}}$$

$$x \underset{2}{\overset{1}{\gtrless}} \frac{\mu_1 + \mu_2}{2}$$

QDA

$\hat{\sigma}_1 = \hat{\sigma}_2$

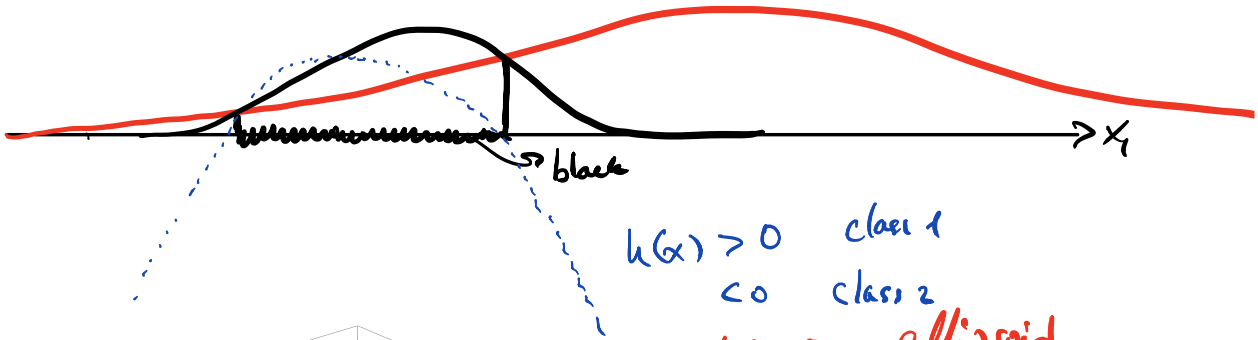
$\hat{\pi}_1 = \hat{\pi}_2$

$$h(x) = \frac{x^2}{2}\left(-\frac{1}{\hat{\sigma}_1^2} + \frac{1}{\hat{\sigma}_2^2}\right) + x \cdot \left(\frac{\mu_1}{\hat{\sigma}_1^2} - \frac{\mu_2}{\hat{\sigma}_2^2}\right)$$

$$-\frac{1}{2}\left(\frac{\mu_1^2}{\hat{\sigma}_1^2} - \frac{\mu_2^2}{\hat{\sigma}_2^2}\right) \underset{2}{\overset{1}{\gtrless}} 0$$

$$-\frac{1}{2}\log \sigma_1^2 + \frac{1}{2}\log \sigma_2^2$$

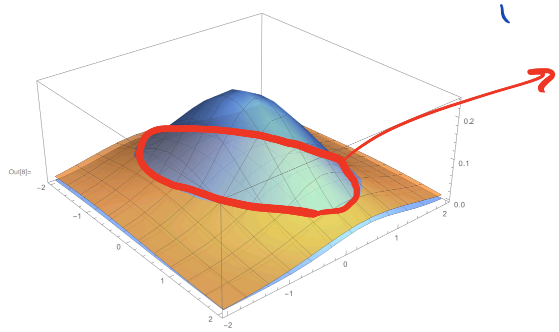
$\hat{\sigma}_1 < \hat{\sigma}_2$



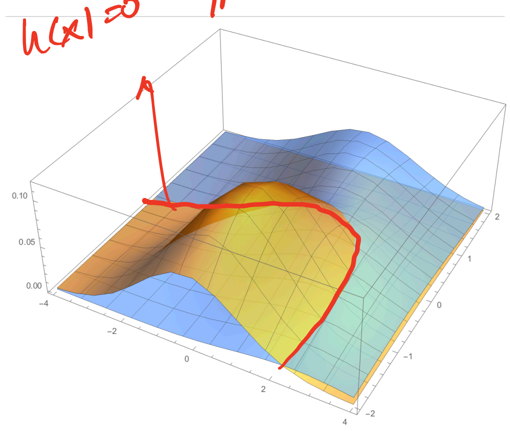
$h(x) > 0$ class 1

< 0 class 2

$h(x) = 0$ ellipsoid



$h(x) = 0$ hyperbola



Regularized Covariance Estimation

$$\hat{\Sigma}_{ML} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_{y_i})(x_i - \mu_{y_i})^T$$

$\hat{\Sigma}^{-1}$ does not exist if $d > n$

$$\hat{\Sigma} = (1-\gamma)\hat{\Sigma}_{ML} + \gamma I$$

γ : regularization parameter $[0,1]$

$\hat{\Sigma}$ is invertible if $\gamma > 0$

Naive Bayes (cont'd)

$X^{(1)}, \ldots, X^{(d)}$ features

NB assumes $X^{(1)}, \ldots, X^{(d)}$ independent given Y

$$\arg\max_k P(Y=k) \overbrace{P(X^{(1)}, \ldots X^{(d)} \mid Y=k)}^{g_k(x)}$$

$$= \arg\max_k P(Y=k) \prod_{j=1}^{d} \underbrace{P(X^{(j)} \mid Y=k)}_{g_k^{(j)}(x)}$$

$$= \arg\max_k \log P(Y=k) + \sum_{j=1}^{d} \log P(X^{(j)} \mid Y=k)$$

1) Binary features ✓

2) Multinomial features: Each feature represents a count

$$X_1 = (100, 20, 0, 3)$$

ex: Text classification $\quad c(d, w_j) =$ # word w_j occurs in document d.

document = [Michigan, football, win, ..., Michigan]

$8/20 \quad 5/20 \quad 3/20 \quad 8/20$

$w_1 =$ football

20 sided die
(20 words in vocabulary)

w_2 = win

w_3 = Michigan.

$P(w_j \mid Y = \text{sports})$ can be estimated by

$$\frac{\sum_{k \text{ in sports}} c(d_k, w_j)}{\sum_j \sum_{k \text{ in sports}} c(d_k, w_j)}$$

Laplace smoothing: add 1 to $c(d_k, w_j)$