- Labs are now online till the end of 2021
  we will use the same Zoom link as for the lecture

LDA/QDA

model distribution of $X \in \mathbb{R}^p$ separately in each class $(k = 1, 2, \dots, K)$ and use Bayes' theorem

$$P_r(Y=k \mid X=x) = \frac{\pi_k f_k(x)}{\sum_{\ell=1}^{K} \pi_\ell f_\ell(x)}$$

$$\pi_k = P_r(Y=k) \overset{\text{estimate}}{=} \frac{\#\{Y=k\}}{\#\{Y\}}$$

$f_k(x)$ – estimated density for

$$P(X \in A \mid Y=k) = \int_A f_k(x)\,dx$$

LDA/QDA: we assume that f_k are Gaussian densities

assume $p=1$ $\quad f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\right)$, $k = 1, \dots, K$

In LDA we additionally assume that $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2 = \sigma^2$
(but in QDA)

LDA, $p=1$

model: $$Pr(Y=k \mid X=x) = \frac{\pi_k \cdot \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum_{\ell=1}^{K} \pi_\ell \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_\ell)^2\right)} \qquad (*)$$

classification: $\underset{k}{\operatorname{argmax}}\ Pr(Y=k \mid X=x)$

Since denominators in (*) are the same for all k, we may look at the numerator only, taking log:

$$\log \pi_k + \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2}(x-\mu_k)^2$$

$$\log \pi_k + \frac{1}{\sigma^2} x\mu_k - \frac{1}{2\sigma^2}\mu_k^2$$

linear in x

$-\frac{1}{2\sigma_k^2} x^2 + \dots$

find k for which this is the largest

$\leftarrow$ ——

assuming $K=2$, $\pi_1 = \pi_2$:

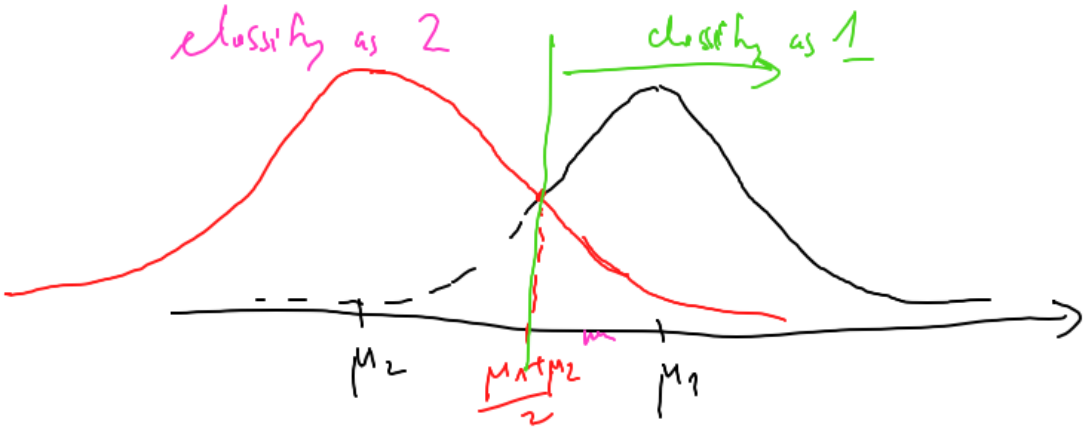
$\mu_1 > \mu_2$

$$\log\pi_1 + \frac{1}{\sigma^2}\mu_1 x - \frac{1}{2\sigma^2}\mu_1^2 \;?\; \log\pi_2 + \frac{1}{\sigma^2}\mu_2 x - \frac{1}{2\sigma^2}\mu_2^2$$

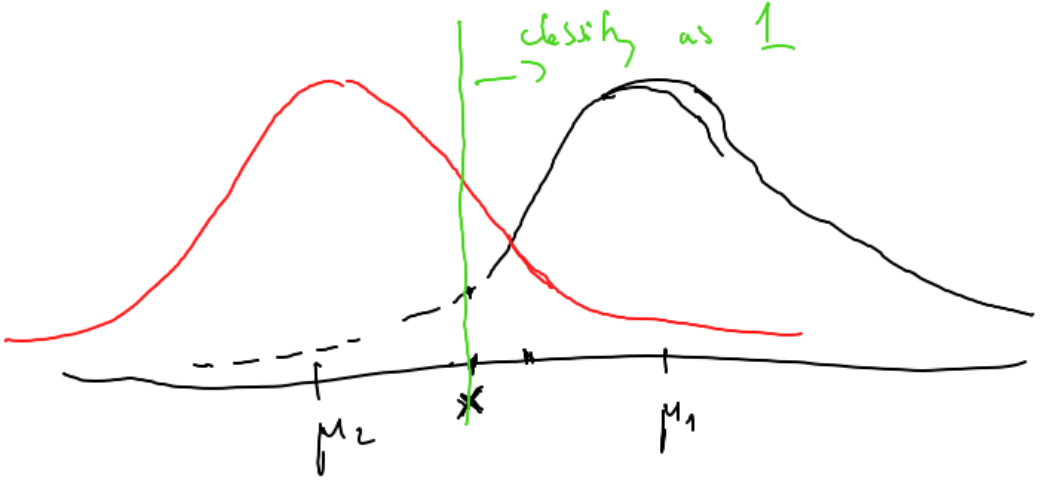
$$\frac{1}{\sigma^2} x(\mu_1-\mu_2) \;?\; \frac{1}{2\sigma^2}(\mu_1^2-\mu_2^2) + \log\pi_2 - \log\pi_1 \quad | : (\mu_1-\mu_2)$$

$$x \;?\; \frac{1}{2}(\mu_1+\mu_2)$$

$p=2$
$K=2$



$\pi_1 = \pi_2$



$\pi_1 > \pi_2$

How to obtain $\hat{\pi}_k, \hat{\mu}_k, \hat{\sigma}^2$:

$$\hat{\pi}_k = \frac{n_k}{n}$$

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i: y_i = k} x_i$$

$n_k = \#\{i : y_i = k\}$ – number of samples in class k

$n = n_1 + n_2 + \dots + n_K$

$$\hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i: y_i = k} (x_i - \mu_k)^2$$

$p > 1$

$$f_k(x) = \frac{1}{(2\pi)^{p/2} |\det \Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_k)^T \Sigma^{-1} (x-\mu_k)\right)$$

$$\Sigma = \mathrm{Cov}\, X = [\mathrm{Cov}(X_i, X_j)]_{i,j}$$

$$\mathrm{Cov}(X_i, X_j) = E\left((X_i - EX_i)(X_j - EX_j)\right)$$

$$\rightsquigarrow \quad \delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k$$

$\operatorname{argmax}_k \delta_k(x) \rightsquigarrow$ prediction

$\hat{\pi}_k, \hat{\mu}_k$ – as before

matrix $p \times p$

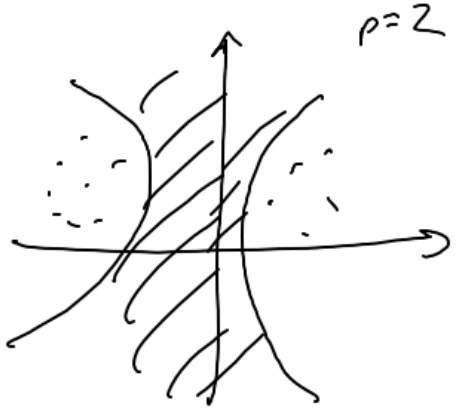
$$\hat{\Sigma} = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i: y_i = k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T$$

$[\,]_{I,J}$

QDA: as LDA, but Σ depends on the class: Σ_k, $k=1,\dots,K$

$$\delta_k(x) = -\frac{1}{2}(x-\mu_k)^T \cdot \Sigma_k^{-1}(x-\mu_k) - \frac{1}{2}\log|\det \Sigma_k| + \log \pi_k$$

$$= \underbrace{\sum a_{ij}^{(k)} x_i \cdot x_j}_{\text{quadratic}} + \underbrace{b_j^{(k)} x_j}_{\text{linear}} + \underbrace{c^{(k)}}_{\text{const.}}$$



$$\hat{\Sigma}_k = \frac{1}{n_k - 1} \sum_{i:\, y_i = k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T$$

in QDA the number of parameters to estimate may be larger than in LDA:

LDA: $\Sigma \sim p \times p = p^2$

QDA: $\Sigma_k \sim p \times p$

$K \cdot p^2$

Naive Bayes

The assumption is that on each $\{Y=k\}$, $X_1, X_2, \ldots, X_p$ (the predictors) are independent

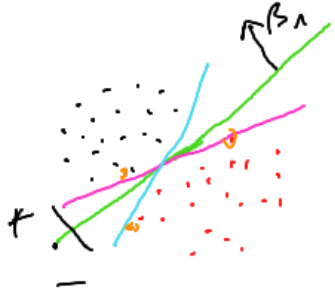
to classify an observation $x \in \mathbb{R}^p$:

$$k: \underset{k}{\operatorname{argmax}} \underbrace{P(Y=k)}_{\pi_k} \cdot \underbrace{P(X_1=x_1 \mid Y=k)}_{\text{estimated}} \cdot \ldots \cdot P(X_p=x_p \mid Y=k)$$

e.g. by assuming some form of the density and by estimating its parameters

Maximal margin classifier

{ binary classification



- Suppose that $(x_i, y_i) \in \mathbb{R}^p \times \{-1, 1\}$
- for (*) simplicity of interpretation let us assume that there exist $\beta_0 \in \mathbb{R}$ and $\beta_1 \in \mathbb{R}^p$ such that

$$\beta_0 + \beta_1 x_i > 0 \quad \text{if } y_i = 1$$
$$\beta_0 + \beta_1 \cdot x_i < 0 \quad \text{if } y_i = -1$$

($\beta_1 \cdot x_i$: scalar product)

$\{x : \beta_0 + \beta_1 x = 0\}$ – hyperplane in $\mathbb{R}^p$

($\beta_0 + \beta_{11} \cdot x_1 + \beta_{12} x_2 = 0$ in $\mathbb{R}^2$)

(i.e., we assume that there exists a separating hyperplane)

this makes sense without (*)

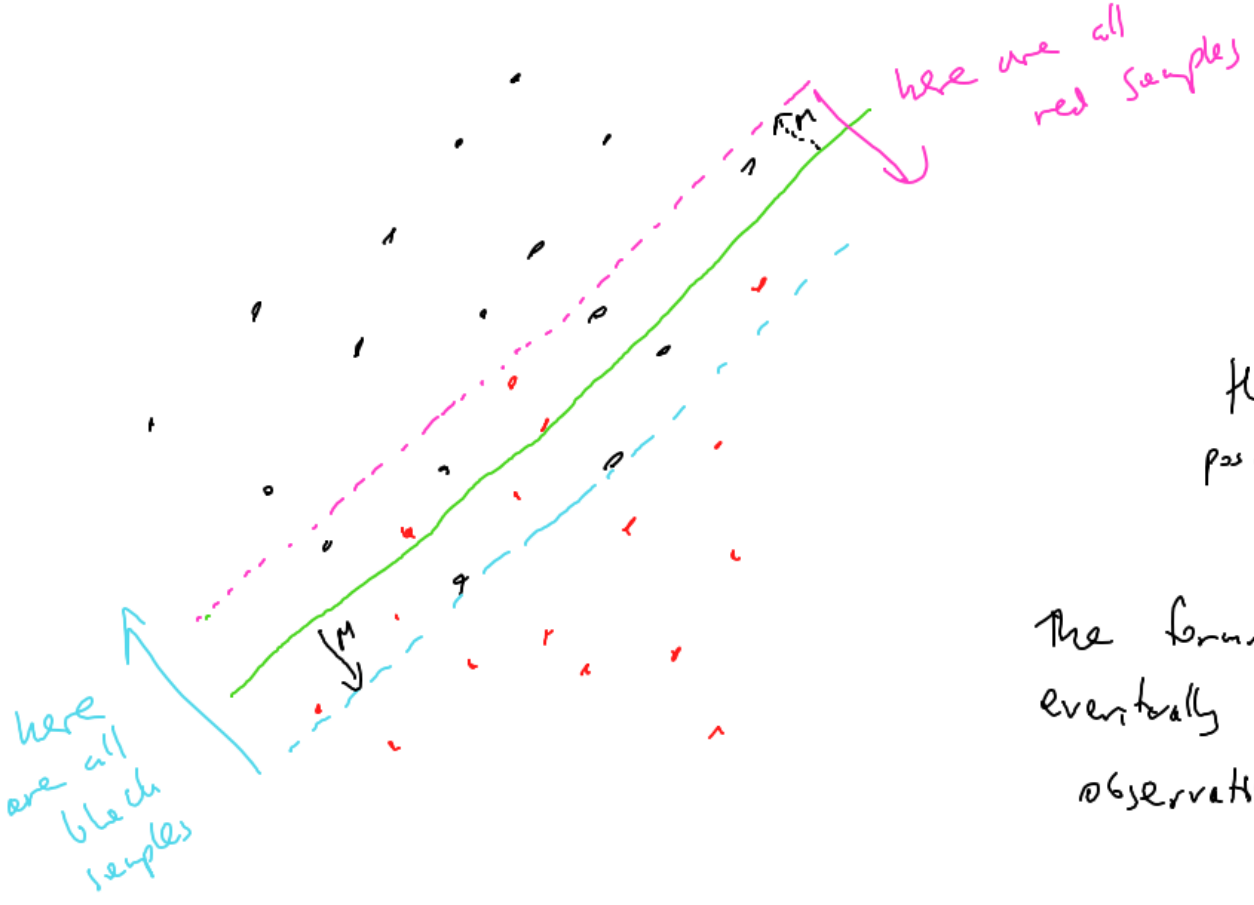
For the max. margin classifier we choose β_0, β_1 so that

maximize M for which $\exists \beta_0, \beta_1$:

$$\min_i (\beta_0 + \beta_1 x_i) \cdot y_i \geq M \qquad \|\beta_1\|_2 = 1$$

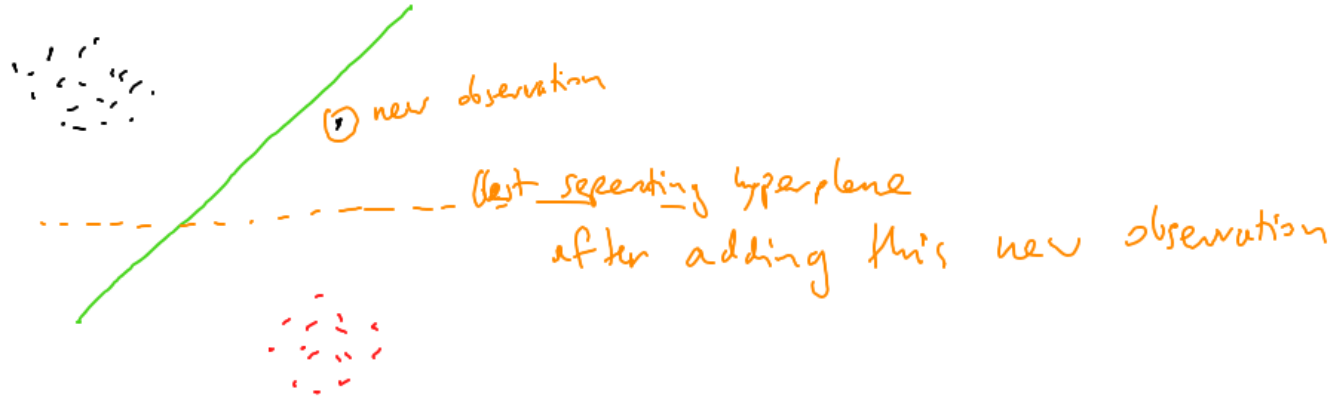
($(\beta_0 + \beta_1 x_i) \cdot y_i$: distance between x_i and the hyperplane $\{x : \beta_0 + \beta_1 x = 0\}$)

$\rightsquigarrow$ classification: $\operatorname{sgn}(\beta_0 + \beta_1 x)$



Here the maximum M possible will be negative.

The formula for the classifier eventually depends on just few observations


new observation
Best separating hyperplane
after adding this new observation