LDA

model $f_k(x) = Pr(X=x \mid Y=k)$ $\qquad$ $Pr(X \in A \mid Y=k) = \int_A f_k(y)\,dy$

$\pi_k = Pr(Y=k)$

$$\longrightarrow \quad Pr(Y=k \mid X=x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

classifier: classify x to the class k for which $Pr(Y=k \mid X=x)$ is the largest

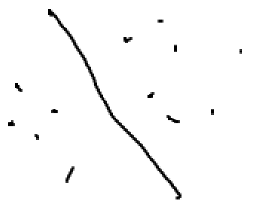
LDA: $(p=1)$

Take $f_k(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(x-\mu_k)^2}{2\sigma^2}\right)$

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

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

$\rightsquigarrow$ decision ~~region~~ boundary is ~~line~~ a hyperplane



LDA for $p \geq 1$

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

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

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

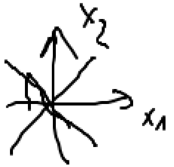
$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k$$

$\rightsquigarrow$ k with the largest $\delta_k(x)$ is the predicted class

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

QDA (Quadratic Discriminant Analysis)

like LDA, but with σ_k^2 or Σ_k depending on the class $k \in \{1, \dots, K\}$



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

Classifier: take k for which $\delta_k(x)$ is the largest.

$x_1^2 - x_2^2 = 0$

$(x_1 - x_2)(x_1 + x_2) = 0$

Ex: $p=1$

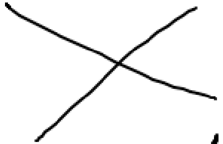
$$-\frac{1}{2}(x-\mu_k)^2 \sigma_k^{-2} - \frac{1}{2}\cdot 2\log\sigma_k + \log\pi_k \gtrless -\frac{1}{2}(x-\mu_j)^2\sigma_j^{-2} - \frac{1}{2}\cdot 2\log\sigma_j + \log\pi_j$$

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

if $\sigma_k \neq \sigma_j$, these do not cancel out

$x_1^2 + x_2^2 + 1 \geq 0$

$-x_1^2 + x_2^2 - 1 \geq 0$



if $p=2$, $K=2$, then the decision region is an ellipse, a parabola, a hyperbola, 2 lines, 1 line or $\emptyset$

compared to LDA, QDA is a more flexible model with more parameters (many more if p is large)

LDA is yet another method that gives a linear decision boundary

Naive Bayes

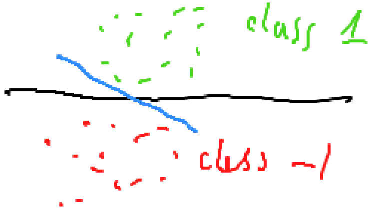
the assumption is that on each $\{Y=k\}$, $X_1, \dots, X_p$ (the predictors) are independent

$$k: \operatorname*{argmax}_k \; P(Y=k) \; \underbrace{P(X_1 = x_{01} \mid Y=k)}_{\text{estimated e.g. by assuming some form of the density and by estimating its parameters}} \cdot \ldots \cdot P(X_p = x_{0p} \mid Y=k)$$

Maximal margin classifier

Suppose that $(X_i, Y_i) \in \mathbb{R}^p \times \{-1, 1\}$ and suppose that the classes can be separated by a hyperplane, i.e., there exists $\beta_0 \in \mathbb{R}$, $\beta_1 \in \mathbb{R}^p \setminus \{(0,0,\dots,0)\}$ such that

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

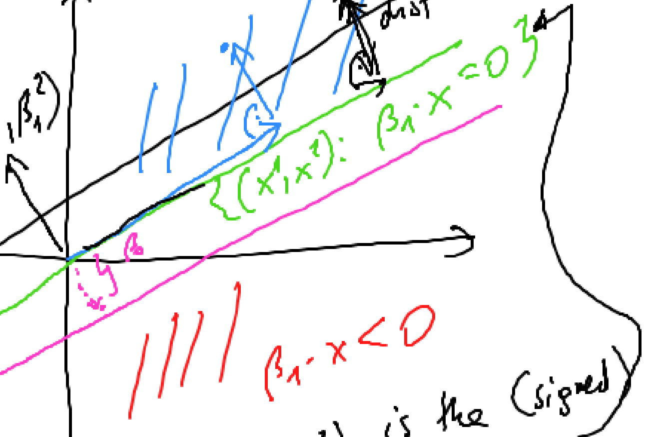


E.g. $p=2$

$$\underbrace{\beta_0 + \beta_1^1 X_i^1 + \beta_1^2 \cdot X_i^2 \gtrless 0}_{\text{a line in } \mathbb{R}^2}$$

$\beta_0 = 0$ $\quad (\beta_1^1, \beta_1^2) \cdot (x^1, x^2) = 0$ $\quad = -\beta_0$

$\beta_0 \neq 0$



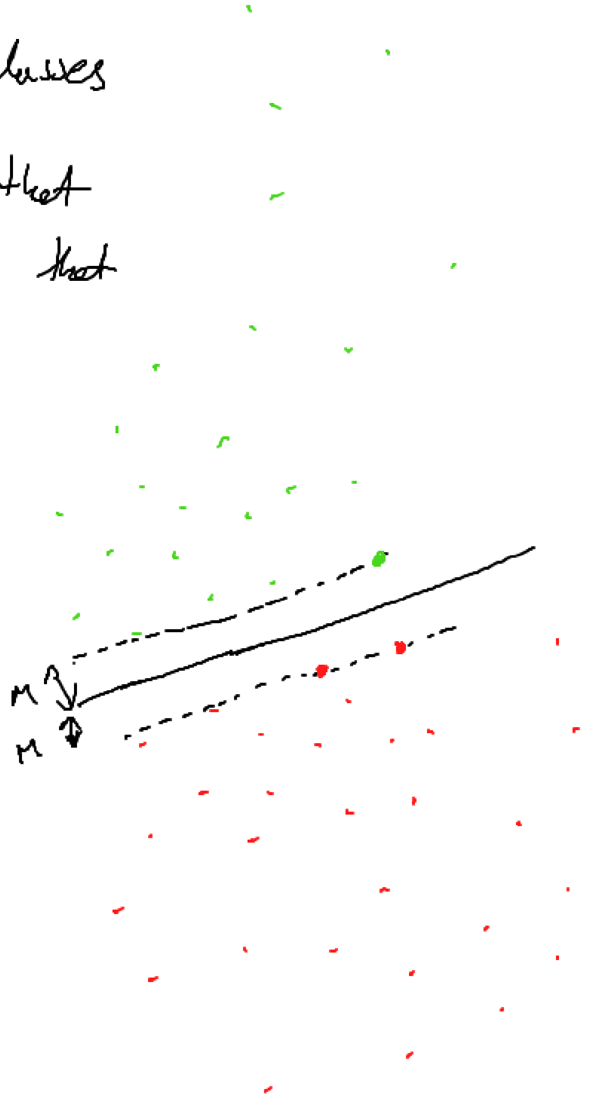
If $\|(\beta_1^1, \beta_1^2)\| = 1$, then $(\beta_1^1, \beta_1^2) \cdot (x^1, x^2)$ is the (signed) distance from the

Then we may choose different hyperplanes that separate the classes

For the maximal margin classifier, we choose β_0, β_1 so that the number M (= margin) is maximised, where M is such that

$$y_i(\beta_0 + \beta_1^T x_i) \geq M \quad \text{for all } i = 1, \dots, n,$$

and $\|\beta_1\|^2 = 1$.



Note: just a few samples influence the final form of the classifier

What if the separating hyperplane does not exist?

Support vector classifier

$C \geq 0$ tuning parameter

We want to maximise M (over $\beta_0, \beta_1, \varepsilon_i$) such that:

- $y_i(\beta_0 + \beta_1 x_i) \geq M(1 - \varepsilon_i)$, $i = 1, \dots, n$
- $\varepsilon_i \geq 0$, $\sum_{i=1}^{n} \varepsilon_i \leq C$,
- $\|\beta_1\|^2 = 1$.

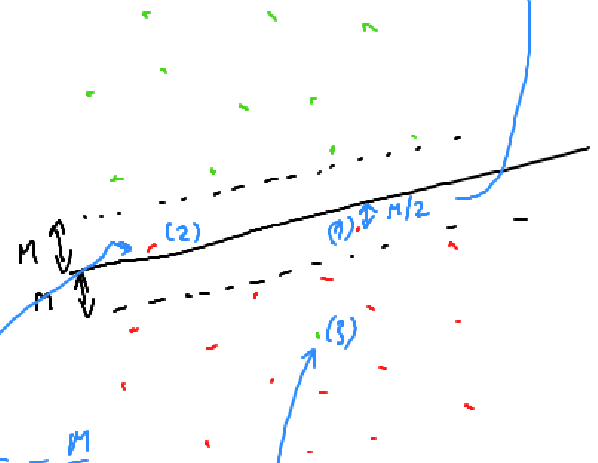
We obtain some $\tilde\beta_0, \tilde\beta_1$ for which the above M is maximised. (provided C is large enough)

the classifier:

$$\operatorname{sgn}(\tilde\beta_0 + \beta_1 \cdot x)$$

$y_1(\beta_0 + \beta_1 x_1) = \frac{M}{2}$

$\varepsilon_1 = \frac{1}{2}$



$y_2(\beta_0 + \beta_1 x_1) = -\frac{M}{4}$

$\varepsilon_2 = \frac{5}{4}$

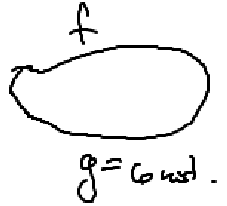
$y_3(\beta_0 + \beta_1 x_3) = -2M$

$\varepsilon_3 = +3$

$\frac{5}{4} + 3 + \frac{1}{2} \leq C$

- C controls the bias-variance tradeoff
  - larger C: smaller variance (more samples influence the form of the classifier)

- How to solve this optimisation problem?
  - Use Karush-Kuhn-Tucker theorem (generalisation of Lagrange multipliers method)



$f + \lambda g$