Lecture 17:

Step 1: Devise a parametric model:

$$Pr\{0|x\} = \frac{e^{B_0 + B_1 x}}{1 + e^{B_0 + B_1 x}} \leftarrow$$

$$Pr\{1|x\} = 1 - Pr\{0|x\}$$

Step 2:

Define an objective for the "goodness of fit" → minimize the objective → find $\hat{B}_0, \hat{B}_1$

$$Pr\{y_i | x_i, B_0, B_1\} \begin{cases} y_i = 0 \rightarrow \frac{e^{B_0 + B_1 x_i}}{1 + e^{B_0 + B_1 x_i}} \\ y_i = 1 \rightarrow \frac{1}{1 + e^{B_0 + B_1 x_i}} \end{cases}$$

$$Pr\{x_i, y_i | B_0, B_1\} = Pr\{x_i\} \cdot Pr\{y_i | x_i, B_0, B_1\}$$

$$P(\underset{x_i}{A}, \underset{y_i}{B}) = P(\underset{x_i}{A}) \cdot Pr(\underset{y_i}{B} | \underset{x_i}{A})$$

as we learned in module 1, mle can be used to estimate parameters of the distribution that generates the data

mle:

$$lik(\beta_0, \beta_1) = \prod_{i=1}^{n} Pr\{x_i, y_i \mid \beta_0, \beta_1\}$$

$$= \prod_{i=1}^{n} Pr\{x_i\} \, Pr\{y_i \mid x_i; \beta_0, \beta_1\}$$

$$= \underbrace{\prod_{i=1}^{n} Pr\{x_i\} \cdot \prod_{i=1}^{n} Pr\{y_i \mid x_i; \beta_0, \beta_1\}}_{\text{objective to be optimized}}$$

$$\underset{\beta_0, \beta_1}{\operatorname{argmax}} \; lik(\beta_0, \beta_1)$$

the same value for any choice of β_0, β_1

$$= \underset{\beta_0, \beta_1}{\operatorname{argmax}} \; \underbrace{\prod_{i=1}^{n} Pr\{x_i\}}_{C} \cdot \prod_{i=1}^{n} Pr\{y_i \mid x_i; \beta_0, \beta_1\}$$

$$= \cancel{C}' \underset{\beta_0 \beta_1}{\operatorname{argmax}} \; \prod_{i=1}^{n} Pr\{y_i \mid x_i; \beta_0, \beta_1\}$$

$$= \underset{\beta_0, \beta_1}{\operatorname{argmax}} \; \sum_{i=1}^{n} \log Pr\{y_i \mid x_i; \beta_0, \beta_1\}$$

$$= \underset{\beta_0, \beta_1}{\operatorname{argmax}} \; \sum_{i: y_i = 0} \log\left(\frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}\right) + \sum_{i: y_i = 1} \log\left(\frac{1}{1 + e^{\beta_0 + \beta_1 x_i}}\right)$$

$$= \underset{\beta_0, \beta_1}{\operatorname{argmax}} \underbrace{\left\{ \sum_{i: y_i = 0} (\beta_0 + \beta_1 x_i) - \sum_{i=1}^{n} \log\left(1 + e^{\beta_0 + \beta_1 x_i}\right) \right\}}_{g(\beta_0, \beta_1)}$$

↳ Convex in terms of β_0, β_1

$$\begin{cases} \dfrac{\partial g}{\partial \beta_0} = 0 \\ \dfrac{\partial g}{\partial \beta_1} = 0 \end{cases} \Longrightarrow$$

- two equations in terms of the two unknowns β_0, β_1.
- no closed-form solution.
  ⇒ we need to solve it numerically / iteratively
- the solution is unique.

$\Longrightarrow \hat{\beta}_0, \hat{\beta}_1$ ← the best fit to the data.

given a new data point x →

$$\hat{Pr}\{0|x\} = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x}}$$

$$\hat{Pr}\{1|x\} = \frac{1}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x}}$$

$$\rightarrow \hat{y}_{\text{logistic\_reg}}(x) = \underset{y \in \{0,1\}}{\operatorname{argmax}} \hat{Pr}\{y|x\}$$

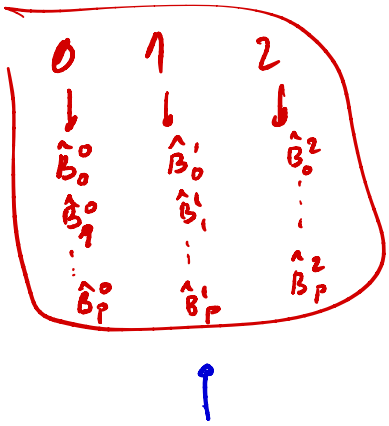
- logistic - regression for $x \in \mathbb{R}^p \rightarrow x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} \in \mathbb{R}^p$

- Binary labels: $y \in \{0,1\}$

$$Pr\{y=0 \mid x\} = \frac{e^{B_0 + B_1 x_1 + \cdots + B_p x_p}}{1 + e^{B_0 + B_1 x_1 + \cdots + B_p x_p}}$$

$$Pr\{y=1 \mid x\} = 1 - Pr\{y=0 \mid x\}$$

- general label set:



$y \in \{0,1,2\}$

$$Pr\{i \mid x\} = \frac{e^{\hat{B}_0^i + \hat{B}_1^i x_1 + \cdots + \hat{B}_p^i x_p}}{\sum_{j=0,1,2} e^{\hat{B}_0^j + \hat{B}_1^j x_1 + \cdots + \hat{B}_p^j x_p}}$$

logistic regression is typically used for settings with binary labels ($y \in \{0,1\}$), and it does not perform well when we have multiple classes.

- Linear / Quadratic Discriminant Analysis:

- Linear Discriminant Analysis (LDA):

$$Pr\{ y \mid x \}$$

Logistic regression directly models this conditional probabilities.

LDA/QDA try to estimate the conditional probabilities through a different route → Baye rule.

Baye's rule

$$Pr\{ y=k \mid X=x \} = \frac{Pr\{ X=x \mid y=k \} \cdot Pr\{ y=k \}}{Pr\{ X=x \}}$$

LDA/QDA aim at estimating these probabilities.

$$Pr\{ X=x \} = \sum_{k=1}^{K} Pr\{ y=k \} Pr\{ X=x \mid y=k \}$$

Assume K classes:

$f_k(x) = \text{gaussian}(\mu_k, \sigma^2)$

$$\Pr\{y=k \mid x=x\} = \frac{\Pr\{x=x \mid y=k\}\,\Pr\{y=k\}}{\Pr\{x=x\}}$$

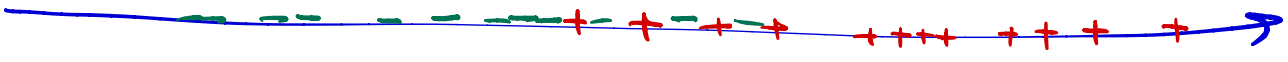
$\Pr\{y=k\}$

$k = 1, \dots, K$

$$\text{Data} = \{(x_i, y_i)\}_{i=1}^{n}$$

example $(x_i \in \mathbb{R},\ y_i \in \{-,+\})$

$\Pr\{x=x \mid y=+\}$



$$\Pr\{y=+\} \overset{\text{estimate}}{\approx} \frac{\#\{i : y_i = +\}}{n}$$

$$Pr\{y=k\} \simeq \frac{\#\{i : y_i = k\}}{n}$$

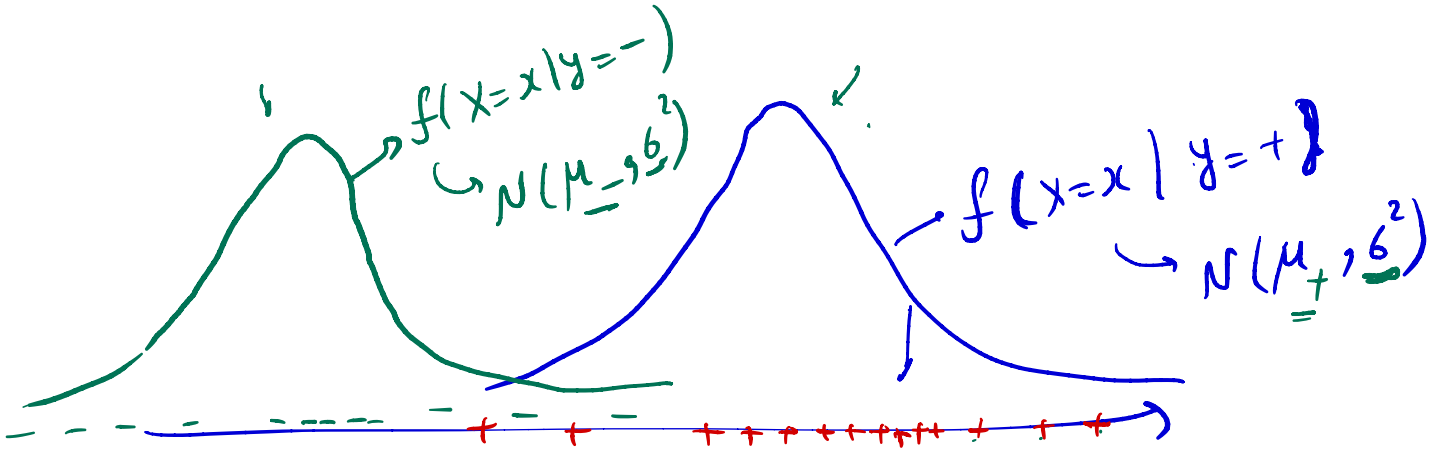
↓

Unbiased estimate of $pr\{y=k\}$

$$f\{X=x \mid Y=k\} \overset{LDA}{=} f_k(x)$$

$$= \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu_k)^2}{2\sigma^2}}$$

LDA "assumes" / "models" that data, given a class k, is generated according to the gaussian distribution.



In LDA, we assume that the variance of the data x is the same over all the classes:

For a gaussian, we need to estimate two parameters: μ_k, σ.

μ_k ← mean of the density

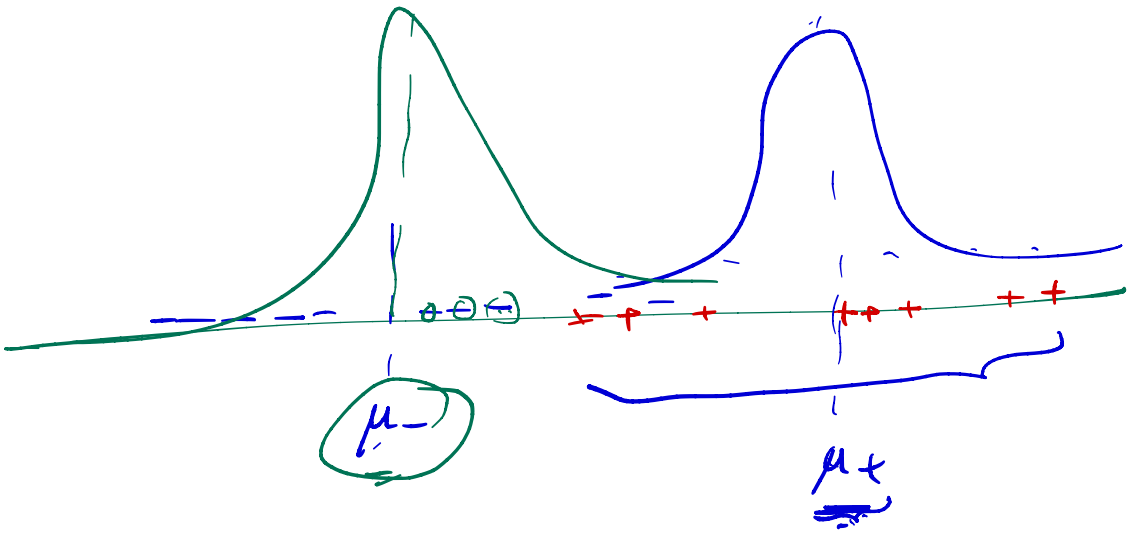
$$\hat{\mu}_k \simeq \frac{\sum_{i: y_i = k} x_i}{n_k}$$ → sample mean

$$n_k = \#\{ i : y_i = k \}$$

Let's now estimate σ:

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

$\sum_{i: y_i = k}$: variance across class k



---

* LDA uses the same Σ for each class, $\Rightarrow$ LDA requires to learn less parameters that the case where we assign different Σ's to the classes.

$\Rightarrow$ QDA → uses different Σ's for the classes.

LDA :