Introduction to data science & artificial intelligence (INF7100)

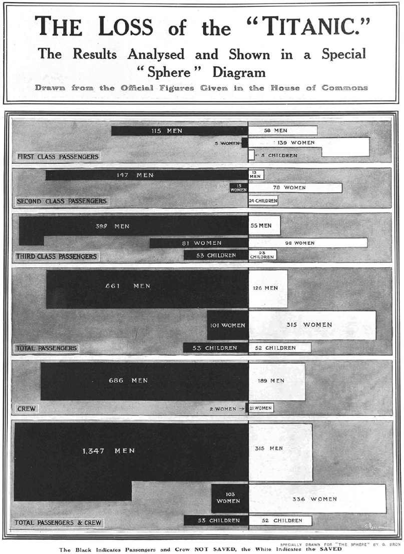
Arthur Charpentier

#351 Classification

été 2020

Titanic, who survived?

$y \in \{0, 1\}$, see
freakonometrics.hypotheses.org



Fisher's Discriminent Analysis

Suppose that we only want to predict the class, $\widehat{y} \in \{0,1\}$ (or more generally $\widehat{y} \in \{a_1, a_2, \cdots, a_J\}$)

$$m^\star(\boldsymbol{x}) = \underset{y\in\{0,1\}}{\text{argmin}}\{\mathbb{P}[Y = y|\boldsymbol{X} = \boldsymbol{x}]\}$$

i.e.

$$m^\star(\boldsymbol{x}) = \underset{y\in\{0,1\}}{\text{argmin}}\left\{\frac{\mathbb{P}[\boldsymbol{X} = \boldsymbol{x}|Y = y]}{\mathbb{P}[\boldsymbol{X} = \boldsymbol{x}]}\right\}$$

(where $\mathbb{P}[\boldsymbol{X} = \boldsymbol{x}]$ is $f(\boldsymbol{x})$ is the continuous case).
If y takes two values – i.e. $\{0,1\}$

$$m^\star(\boldsymbol{x}) = \begin{cases} 1 \text{ if } \mathbb{E}(Y|\boldsymbol{X} = \boldsymbol{x}) > \dfrac{1}{2} \\ 0 \text{ otherwise} \end{cases}$$

Fisher's Discriminent Analysis

The set

$$\mathcal{D}_S = \left\{ \boldsymbol{x} : \ \mathbb{E}(Y|\boldsymbol{X} = \boldsymbol{x}) = \frac{1}{2} \right\}$$

is called decision border.

Assume that $\boldsymbol{X}|Y = 0 \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ and $\boldsymbol{X}|Y = 1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, then, if r_y^2 is Manalahobis distance from $\boldsymbol{x}$ to $\boldsymbol{\mu}_y$

$$r_y^2 = [\boldsymbol{x} - \boldsymbol{\mu}_y]^\top \boldsymbol{\Sigma}_y^{-1} [\boldsymbol{x} - \boldsymbol{\mu}_y] \text{ for } y \in \{0, 1\},$$

$$m^\star(\boldsymbol{x}) = \begin{cases} 1 \text{ if } r_1^2 < r_0^2 + 2\log \dfrac{\mathbb{P}(Y=1)}{\mathbb{P}(Y=0)} + \log \dfrac{|\boldsymbol{\Sigma}_0|}{|\boldsymbol{\Sigma}_1|} \\ 0 \text{ otherwise} \end{cases}$$

Fisher's Discriminent Analysis

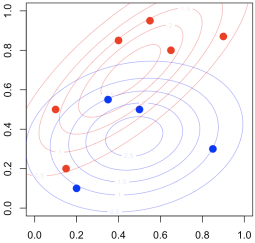
Let δ_y be defined (for $y \in \{0,1\}$) as

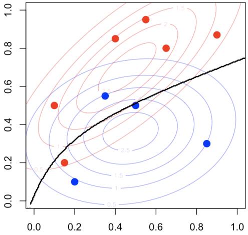
$$\delta_y(\boldsymbol{x}) = -\frac{1}{2}\log|\boldsymbol{\Sigma}_y| - \frac{1}{2}[\boldsymbol{x}-\boldsymbol{\mu}_y]^\top \boldsymbol{\Sigma}_y^{-1}[\boldsymbol{x}-\boldsymbol{\mu}_y] + \log\mathbb{P}(Y=y)$$

so that the decision frontier is

$$\{\boldsymbol{x} \text{ such that } \delta_0(\boldsymbol{x}) = \delta_1(\boldsymbol{x})\}$$

which is quadratic in $\boldsymbol{x}$



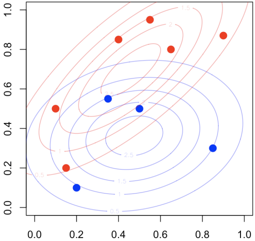


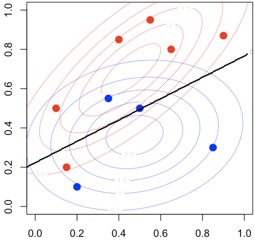
Fisher's Discriminent Analysis

Fisher (1936) added the assumption $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1$. Then

$$\delta_y(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_y - \frac{1}{2}\boldsymbol{\mu}_y^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_y + \log \mathbb{P}(Y=y)$$

and the decision border is linear in $\boldsymbol{x}$





Fisher's Discriminent Analysis

If $\boldsymbol{X}|Y=0 \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})$ and $\boldsymbol{X}|Y=1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ then

$$\log \frac{\mathbb{P}(Y=1|\boldsymbol{X}=\boldsymbol{x})}{\mathbb{P}(Y=0|\boldsymbol{X}=\boldsymbol{x})}$$

is equal to

$$\boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1}[\boldsymbol{\mu}_y] - \frac{1}{2}[\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0]^\top \boldsymbol{\Sigma}^{-1}[\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0] + \log \frac{\mathbb{P}(Y=1)}{\mathbb{P}(Y=0)}$$

which is linear in $\boldsymbol{x}$, i.e.

$$\log \frac{\mathbb{P}(Y=1|\boldsymbol{X}=\boldsymbol{x})}{\mathbb{P}(Y=0|\boldsymbol{X}=\boldsymbol{x})} = \boldsymbol{x}^\top \boldsymbol{\beta}$$

which will be close to the linear regression...

Logistic Regression

$$p_i = \mathbb{E}(Y_i|\boldsymbol{X}_i = \boldsymbol{x}_i) \in [0,1] \quad \neq \boldsymbol{x}_i^\top \boldsymbol{\beta}$$

$\longrightarrow$ use the odds ratio

$$\text{odds}_i = \frac{\mathbb{P}[Y_i = 1]}{\mathbb{P}[Y_i = 0]} = \frac{p_i}{1 - p_i} \in [0, \infty].$$

i.e. if we take the logarithm

$$\log(\text{odds}_i) = \log\left(\frac{p_i}{1 - p_i}\right) \in \mathbb{R}.$$

That transformation is called logit,

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1 - p_i}\right) = \boldsymbol{x}_i^\top \boldsymbol{\beta}$$

or

$$p_i = \text{logit}^{-1}(\boldsymbol{x}_i^\top \boldsymbol{\beta}) = \frac{\exp[\boldsymbol{x}_i^\top \boldsymbol{\beta}]}{1 + \exp[\boldsymbol{x}_i^\top \boldsymbol{\beta}]}.$$

Logistic Regression

Data : $\{(\boldsymbol{x}_i, y_i) = (x_{1,i}, x_{2,i}, y_i), i = 1, \cdots, n\}$

$$\mathbb{P}(Y = 1|\boldsymbol{X} = \boldsymbol{x}) = \frac{\exp[\boldsymbol{x}^\top \beta]}{1 + \exp[\boldsymbol{x}^\top \beta]}$$

Inference using maximum likelihood techniques

$$\widehat{\beta} = \text{argmin}\left\{\sum_{i=1}^{n} \log[\mathbb{P}(Y = y_i|\boldsymbol{X} = \boldsymbol{x}_i)]\right\}$$

and the score model is then

$$s(\boldsymbol{x}) = \frac{\exp[\boldsymbol{x}^\top \widehat{\beta}]}{1 + \exp[\boldsymbol{x}^\top \widehat{\beta}]}$$