Logistic regression

binary classification, $\underline{X}$ – predictors ($\in \mathbb{R}^n$), $Y \in \{0,1\}$ – class

model for $p(X) = \Pr(Y=1 \mid X)$

If X is a discrete random variable, then it's simple: suppose $X \in \{a_1, \dots, a_m\}$

$$p(a_m) = \Pr(Y=1 \mid X = a_m) = \frac{\Pr(Y=1 \wedge X = a_m)}{\Pr(X = a_m)}$$

otherwise,

heuristically: $p(\cancel{x}) \Pr(\underline{X} = x) = \Pr(Y=1 \wedge \underline{X} = x)$

$$\int_A p(x)\,d\cancel{x} \cdot \Pr(X\in A) \overset{?}{=} \int_A p(x)\, P(X=x)\, dx \overset{?}{=} \int_A \Pr(Y=1 \wedge X=x)\, dx \overset{?}{=} \underline{\Pr(Y=1, X\in A)}$$

$$\int_A p(x)\,dx = \frac{\Pr(Y=1, X\in A)}{\Pr(X\in A)}$$

$$\varphi(x) = \frac{1}{1+e^{-x}} \in (0,1)$$

$$p(X) = \varphi(\beta_0 + \beta_1 X) = \frac{1}{1+e^{-\beta_0-\beta_1 X}}$$

$$\Rightarrow \beta_0 + \beta_1 X = \log \frac{p(X)}{1-p(X)}$$

β_0, β_1 are chosen to maximise the likelyhood function:

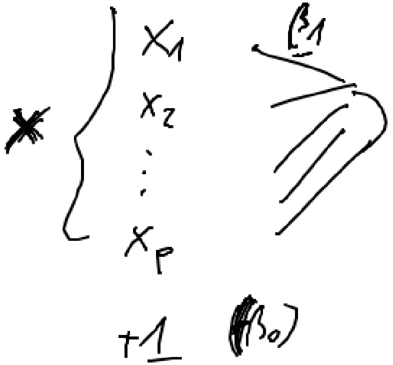
$$\ell(\beta_0, \beta_1) = \prod_{i:\, y_i=1} p(x_i) \cdot \prod_{i:\, y_i=0} \left(1 - p(x_i)\right)$$

$$p(x) = \Pr(Y = 1 \mid X = x)$$

$$\| $$

$$\frac{1}{1+e^{-\beta_0-\beta_1 x}}$$

This corresponds to a single neuron with logistic activation function (and $-\ell$ as the loss function)



$\rightarrow \frac{1}{2}$ (artificial neuron with const. output) $\rightsquigarrow 0$

$\nearrow$ (zero weights)

$$\varphi(+\beta_0 + \beta_1 x) = \frac{1}{1+e^{-\beta_0-\beta_1 x}}$$

$\searrow 1$

Multinomial logistic regression

$Y \in \{0, 1, \dots, k-1\}$ – classes

model:

$$\log \frac{Pr(Y=j \mid X)}{Pr(Y=k-1 \mid X)} = \beta_0^{(j)} + \beta_1^{(j)} X \quad , j = 0, 1, \dots, k-2$$

$$\Rightarrow \quad Pr(Y=j \mid X) = e^{\beta_0^{(j)} + \beta_1^{(j)} \cdot X} \cdot Pr(Y=k-1 \mid X) \qquad \Big| \sum_{j=0}^{k-2}$$

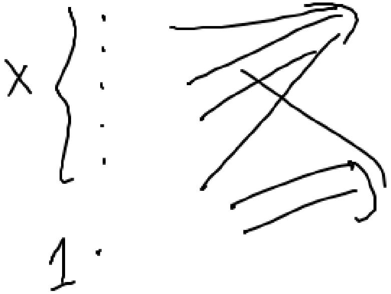
$$1 - Pr(Y=k-1 \mid X) = \left(\sum_{j=0}^{k-2} e^{\beta_0^{(j)} + \beta_1^{(j)} X} \right) Pr(Y=k-1 \mid X)$$

$$\begin{cases} Pr(Y=j \mid X) = \dfrac{e^{\beta_0^{(j)} + \beta_1^{(j)} \cdot X}}{1 + \sum_{j=0}^{k-1} e^{\beta_0^{(j)} + \beta_1^{(j)} X}} & , j = 0, \dots, k-2 \\ Pr(Y=k-1 \mid X) = \dfrac{1}{1 + \sum_{j=0}^{k-1} e^{\beta_0^{(j)} + \beta_1^{(j)} X}} \end{cases}$$

$\rightsquigarrow$ neural net with $k-1$ neurons with logistic function and 1 additional neuron with 0 weights and also logistic act.-f.

$$x : \quad \Rightarrow \quad \frac{1}{1+e^{-\beta_0^{(j)} - \beta_1^{(j)} x}} \quad , j = 0, \dots, k-2$$

$$1 \cdot \quad \Rightarrow \quad \frac{1}{2} \text{ for } k-1$$



$$\left[\frac{1}{1+e^{-\beta_0^{(j)}-\beta_1^{(j)}\cdot x}}\right] \quad \text{for class } j=0,\dots,k-2 \quad \cancel{\text{class}}$$

$$\left[\varphi(0)=\frac{1}{2}\right] \quad \text{for } k-1 \text{ class}$$

$$=\frac{1}{1+1}$$

→ classification: the the class with smallest value ○ wins

→ $e^{\beta_0^{(j)}+\beta_1^{(j)}x}$, $j=0,\dots,k-2$

or 1 for $k-1$

has to ~~the~~ the largest

Bayes classifier

assign each observation to the most likely class, given its predictor values

$$\underset{k}{\operatorname{argmax}} \Pr(Y=k \mid X=x_0)$$

Thm. The average test error rate, $E\left(\frac{1}{N}\mathbb{1}(\underset{\substack{\uparrow\\ \text{actual}\\ \text{value (class)}\\ \text{in the obs. } (x_0,y_0)}}{y_0} \neq \underset{\substack{\uparrow\\ \text{predicted}\\ \text{class}}}{\hat{y}_0})\right)$, is the smallest for the Bayes classifier.

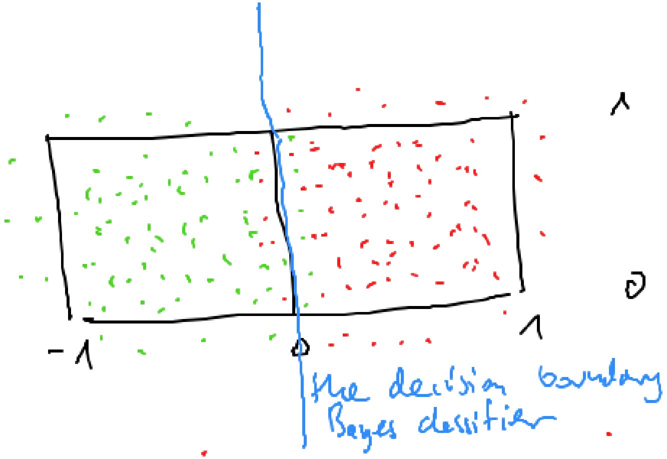
The problem is, in practice we don't know $\Pr(Y=k \mid X=x_0)$

Example: Take $(\underline{x}, y)$ in the following way: $y \in \{0,1\}$

If $y=0$, then $\underline{x} = u_0 + n$, with $u_0 \sim \text{Uniform}([-1,0]\times[0,1])$, $n \sim N(0, \sigma^2 \underset{\begin{bmatrix}1&0\\0&1\end{bmatrix}}{I_d})$

$u_0 \perp n$

If $y=1$, then $\underline{x} = u_1 + n$, with $u_1 \sim \text{Uniform}([0,1]\times[0,1])$, $u_1 \perp n$



the decision boundary for the Bayes classifier

σ^2 - small

Here, one could compute the densities of the random variables u_0+n on u_1+n

Bayes theorem:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

Linear Discriminant Analysis (LDA)

Model the distribution of X separately in each of the response classes (i.e., given Y) and then use Bayes' theorem:

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

where

$$\pi_k = \Pr(Y=k)$$

$$f_k(x) = \underbrace{\Pr(X=x \mid Y=k)}_{\text{density}}$$

i.e.

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

π_k are simple to estimate: just take $\dfrac{\#\{i \in \{1,\dots,N\} : Y_i = k\}}{N}$

For f_k: assume that $X \mid Y=k$ ~~comes from some~~ are drawn from some distribution with some parameters and then estimate the parameters

In LDA we assume that the distribution is Gaussian with the covariance matrix being the same for all classes

LDA for $p=1$ predictors: $X\in\mathbb{R}$

- assume that $f_k(x)=\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)$ for some $\mu_k,\sigma\in\mathbb{R}$, $\sigma>0$

(i.e., we assume that ~~all~~ the variance σ^2 does not depend on k)

Then

$$\Pr(Y=k|X=x)=\frac{\pi_k\cdot\frac{1}{\sqrt{2\pi}\,\sigma}\cdot\exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum_{l=1}^{K}\pi_l\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)} \qquad (*)$$

and we classify the sample to be of class k for which $(*)$ is the largest.

Since the denominator is the same for all k, so we classifey to be of class k for which the foll. expression is the largest.

$$\log\pi_k+\underbrace{\log\frac{1}{\sqrt{2\pi}\,\sigma}}_{\text{does not depend on }k}-\frac{1}{2\sigma^2}\underbrace{(x-\mu_k)^2}_{(\underset{\uparrow\ \text{does not depend on }k}{x^2}-2x\mu_k+\mu_k^2)} \qquad \text{or} \qquad \underbrace{\log\pi_k+\frac{\mu_k}{\sigma^2}x-\frac{\mu_k^2}{2\sigma^2}}$$

Ex. suppose that $K=2$, $\underline{\pi_1=\pi_2=\frac{1}{2}}$

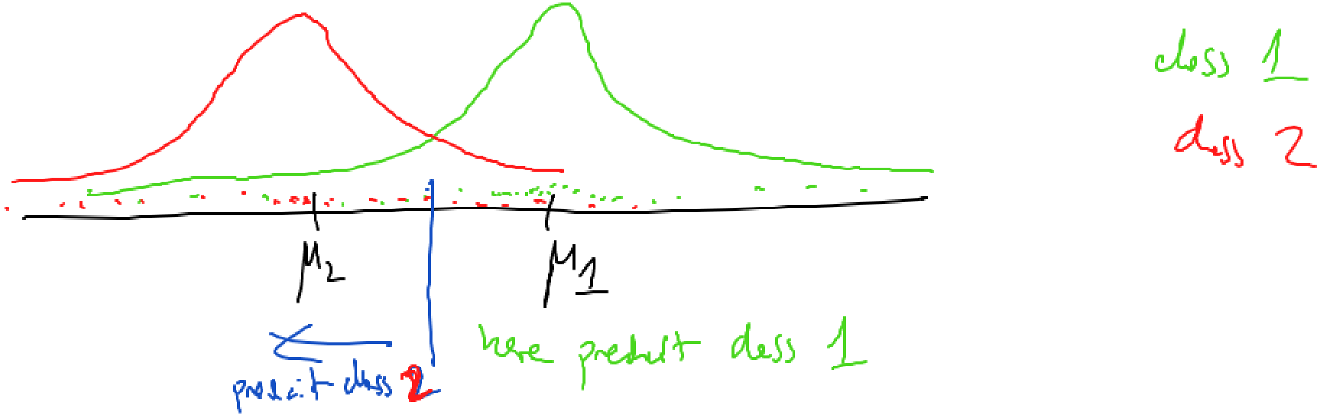
$$\cancel{\log\pi_1} + \frac{\mu_1}{\sigma^2}x - \frac{\mu_1^2}{2\sigma^2} \overset{?}{<} \cancel{\log\pi_2} + \frac{\mu_2}{\sigma^2}x - \frac{\mu_2^2}{2\sigma^2}$$

$$\frac{\mu_1-\mu_2}{\sigma^2}x < \frac{\mu_1^2-\mu_2^2}{2\sigma^2} \qquad \Big| : (\mu_1-\mu_2) > 0$$

Suppose $\mu_1 > \mu_2$:

$$\frac{x}{\sigma^2} < \frac{\mu_1+\mu_2}{2\sigma^2}$$

if $\underline{x < \frac{\mu_1+\mu_2}{2}}$ we predict to be in class 2



Estimation of $\hat{\mu}_k$, $\hat{\sigma}^2$, $\hat{\pi}_k$:

the number of observation in clas k

$$\hat{\pi}_k = \frac{n_k}{n} \quad , k=1,\dots,K$$

the number of all observations

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

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