Machine Learning II

Session 1 : Supervised Learning: Discriminant Analysis

Mohamad GHASSANY

EFREI PARIS

Course Overview

Format:

- Course sessions: 6 sessions of 5 hours each.
- Sessions are CTP.

Course chapters:

Session 1: Supervised Learning: Discriminant Analysis

Session 2:

Session 3:

Session 4:

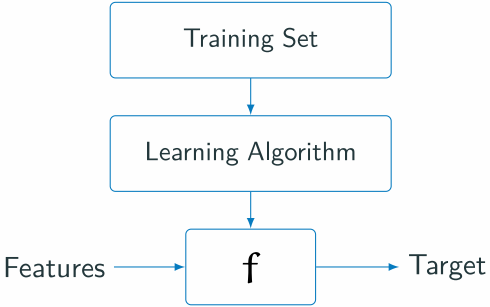
Session 5:

Session 6:

In general, any machine learning problem can be assigned to one of these broad types:

Supervised Learning

The term supervised learning refers to the fact that we gave the algorithm a data set in which the **"right answers"** (known as **labels**) were given.



- Supervised Learning refers to a set of approaches for estimating f.
- f is also called ***hypothesis*** in Machine Learning.

Regression

- The example of the house price prediction is also called a regression problem.
- A regression problem is when we try to predict a quantitative (continuous) value output. Namely the price in the example.

Classification

- The process for predicting qualitative (categorical, discrete) responses is known as classification.
- Methods: Logistic regression, Support Vector Machines, etc..

Classification

- Email: Spam / Not Spam?
- Online Transactions: Fraudulent (Yes/No)?
- Tumor: Malignant / Benign?
- Loan Demand (Credit Risk): Safe / Risky

Classification: categorical output

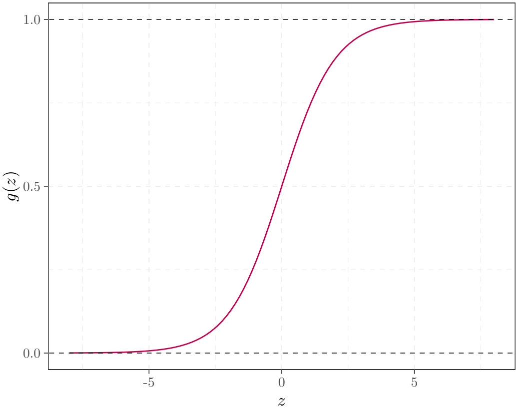
- $y \in \{0, 1\}$
- 0: "Negative class"
- 1: "Positive Class"

.. and also multiclass classification

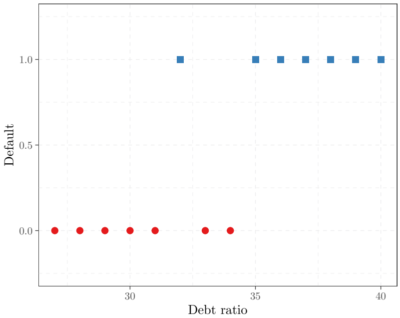
Logistic Regression

The logistic function (sigmoid)

$$g(z) = \frac{e^z}{1 + e^z} = \frac{1}{1 + e^{-z}}$$

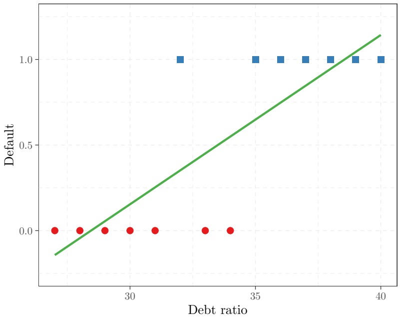


Logistic Regression: why not linear regression

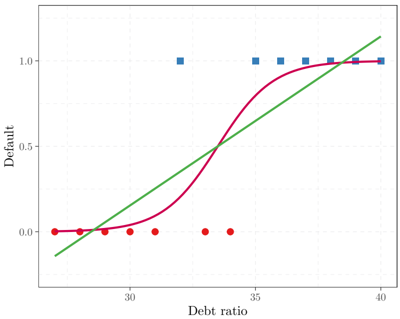


- $y \in \{0, 1\}$:
  - "0": Negative class (here no default)
  - "1": Positive class (here default)
- $f_\omega(x) = \omega' x$ can be > 1 ou < 0 !
- Ideally $0 \leqslant f_\omega(x) \leqslant 1$ s.t.:
  - If $f_\omega(x) \geqslant 0.5$, predict "$y = 1$"
  - If $f_\omega(x) < 0.5$, predict "$y = 0$"

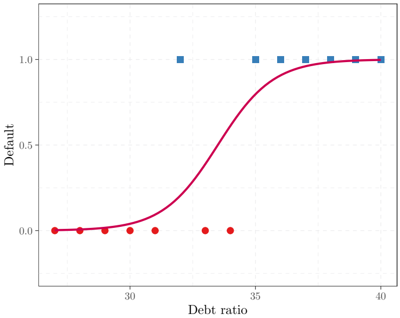
- Let $f_\omega(x) = \omega' x$



- Let $f_\omega(x) = \cancel{\omega' x} = g(\omega' x) = \dfrac{1}{1 + e^{-\omega' x}}$



- $0 \leqslant g(\omega' x) \leqslant 1$
- $f_\omega(x) = g(\omega' x) =$ estimated probability that $y = 1$ on input x
- Probability that $y = 1$, given x, parameterized by ω
- $g(\omega' x) = p(y = 1 \mid x) = p(x)$
- $y \in \{0, 1\}$ so $p(y = 1 \mid x) + p(y = 0 \mid x) = 1$



logistic score

$$p(x) = p(y = 1 \mid x) = \frac{e^{\omega' x}}{1 + e^{\omega' x}} = \frac{1}{1 + e^{-\omega' x}}$$

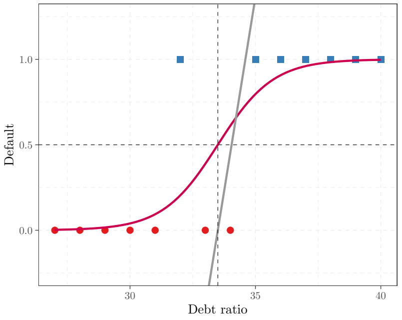
odds (côtes)

$$\frac{p(x)}{1 - p(x)} = e^{\omega' x}$$

log-odds or logit (logarithme des côtes)

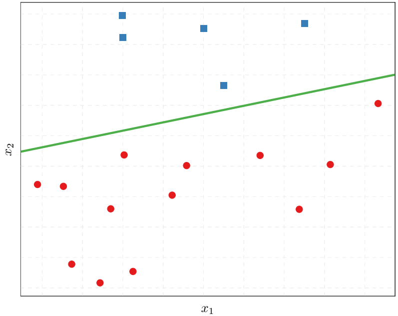
$$\log\left(\frac{p(x)}{1 - p(x)}\right) = \omega' x$$

Logistic Regression: decision boundary



- We predict "$y = 1$" if $p(x) \geqslant 0.5$ which means $\omega' x \geqslant 0$

- $\omega_0 + \omega_1 x \geqslant 0 \Rightarrow x \geqslant -\dfrac{\omega_0}{\omega_1}$

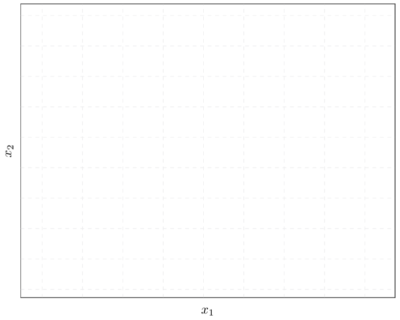


- $p(x) = p(y = 1 \mid x) = f_\omega(x) = g(\omega' x)$
- Predict "$y = 1$" if $p(x) \geqslant 0.5$ which means $\omega' x \geqslant 0$
- $\omega_0 + \omega_1 x_1 + \omega_2 x_2 \geqslant 0$ So

$$x_2 \geqslant -\frac{\omega_1}{\omega_2} x_1 - \frac{\omega_0}{\omega_2}$$

Fun

Identify TP, TN, FP, FN on the figure.



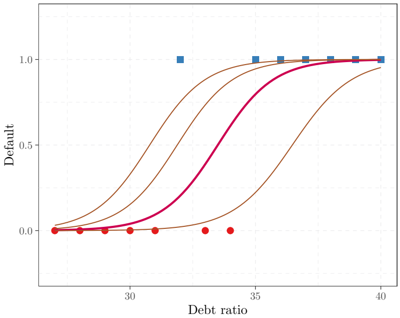
- Let
  $f_\omega(x) = g(\omega_0 + \omega_1 x_1 + \omega_2 x_2 + \omega_3 x_1^2 + \omega_4 x_2^2)$

- For example, predict "$y = 1$" if
  $-1 + x_1^2 + x_2^2 \geqslant 0$

- Or, $f_\omega(x) = g(\omega_0 + \omega_1 x_1 + \omega_2 x_2 + \omega_3 x_1^2 + \omega_4 x_1^2 x_2 + \omega_5 x_1^2 x_2^2 + \ldots)$

Logistic Regression: model estimation

- Parameters to estimate: $\omega = \{\omega_0, \omega_1\}$ if univariate
- $\omega = \{\omega_0, \omega_1, \ldots, \omega_p\}$ if multivariate with p features
- How to choose parameters ω?



[1] check: https://shinyserv.es/shiny/log-maximum-likelihood/, by Eduardo García Portugués

Cost function of simple linear regression

- Model: $f_\omega(x) = \omega_0 + \omega_1 x = \omega' x$
- Parameters: ω_0 and ω_1
- Cost function: $J(\omega_0, \omega_1) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} \left(f_\omega\left(x^{(i)}\right) - y^{(i)}\right)^2$
- Goal: $\min_{\omega_0, \omega_1} J(\omega_0, \omega_1)$

Non-convex in case of logistic regression !

Logistic Regression: how to estimate the parameters

- How to choose parameters ω?
- $y \in \{0, 1\}$, Let's assume:

$$p(y = 1 \mid x, \omega) = f_\omega(x)$$
$$p(y = 0 \mid x, \omega) = 1 - f_\omega(x)$$

- We represent $y \mid x, \omega \sim \mathcal{B}(f_\omega(x))$
- We can write:

$$p(y \mid x, \omega) = (f_\omega(x))^y \, (1 - f_\omega(x))^{1-y} \qquad y \in \{0, 1\}$$

- Given the n observations and assuming independance, we estimate ω by maximizing the likelihood:

$$\mathcal{L}(\omega) = \prod_{i=1}^{n} p\left(y^{(i)} \mid x^{(i)}, \omega\right)$$

- The likelihood:

$$\mathcal{L}(\omega) = \prod_{i=1}^{n} p\left(y^{(i)} \mid x^{(i)}, \omega\right)$$
$$= \prod_{i=1}^{n} \left(f_\omega\left(x^{(i)}\right)\right)^{y^{(i)}} \left(1 - f_\omega\left(x^{(i)}\right)\right)^{1-y^{(i)}}$$

- Maximizing the likelihood is same as maximizing its log:

$$\ell(\omega) = \log\left(\mathcal{L}(\omega)\right)$$
$$= \sum_{i=1}^{n} y^{(i)} \log f_\omega\left(x^{(i)}\right) + \left(1 - y^{(i)}\right) \log\left(1 - f_\omega\left(x^{(i)}\right)\right)$$

- Maximizing $\ell(\omega)$ is same as minimizing: $-\frac{1}{n}\ell(\omega)$
- Let $J(\omega) = -\frac{1}{n}\ell(\omega)$, a convex cost function for the logistic regression model (known as *binary cross entropy*).

Logistic Regression: optimization of the cost function

- **Goal**: Find ω s.t. $\omega = \text{argmin}_{\omega} J(\omega)$
- $J(\omega) = -\frac{1}{n}\sum_{i=1}^{n} y^{(i)} \log f_{\omega}\left(x^{(i)}\right) + \left(1 - y^{(i)}\right) \log\left(1 - f_{\omega}\left(x^{(i)}\right)\right)$
- Contrary to the linear regression, this cost function **does not** have an analytical solution. We need an optimization technique.

GD for logistic regression

- `initialize` ω 'randomly"
- `repeat until convergence`{

$$\omega_i^{new} = \omega_i^{old} - \alpha \frac{\partial J(\omega)}{\partial \omega_i}$$

`simultaneously for` $i = 0, \dots, p$ }

- Recall that $g(z) = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}}$
- Notice that $g'(z) = g(z)(1-g(z))$
- $\frac{\partial J(\omega)}{\partial \omega_i} = (y - f_\omega(x))x_i$

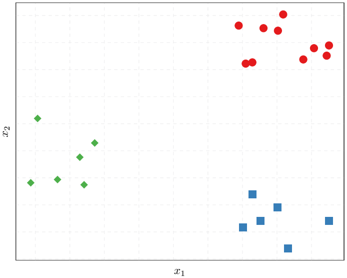
GD for logistic regression

- `initialize` ω `randomly`
- `repeat until convergence`$\{$

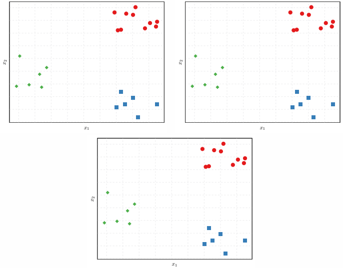
$$\omega_i^{new} = \omega_i^{old} - \alpha \frac{1}{n} \sum_{i=1}^{n} \left(f_\omega\left(x^{(i)}\right) - y^{(i)}\right) . x_i^{(i)}$$

`simultaneously for` $i = 0, \dots, p$ $\}$

- Weather: Sunny, Cloudy, Rain, Snow
- Medical diagrams: Not ill, Cold, Flu
- News articles: Sport, Education, Technology, Politics



- $f_\omega^{(i)}(x) = P(y = 1|x, \omega)$ for $i = 1, 2, 3$
- Train a logistic regression classifier for each class i to predict the probability that $y = i$
- On a new input x, to make a prediction, pick the class i that maximizes $f_\omega^{(i)}(x)$



Logistic Regression: Remarks & summary

- Very famous method and maybe the most used
- Adapted for a binary y
- Relation with linear regression
- Linear decision boundary, but can be non linear using other hypothesis
- Direct calculation of $p(y = 1 \mid x)$

Disciminant Analysis

- L'analyse discriminante est une famille méthode de classification qui cherche à prédire avec quelle probabilité un individu appartient à une classe
- Au lieu de calculer directement $p(y \mid x)$, comme dans la régression logistique, on modélise $p(x \mid y)$ et $p(x)$
- Ensuite, on applique la formule de Bayes pour calculer $p(y \mid x)$
- Les méthodes présentées dans ce chapitre sont appelées méthodes génératives

Bayes

$$p(y \mid x) = \frac{p(y)p(x \mid y)}{p(x)}$$

- $p(y \mid x)$ probabilité "a posteriori"
- $p(y)$ probabilité "a priori"
- $p(x \mid y)$ distribution dans les classes
- $p(x)$ vraisemblance

Pour prédire la classe associée à une observation x:

$$\begin{aligned} \arg\max_y p(y|x) &= \arg\max_y \frac{p(x|y)p(y)}{p(x)} \\ &= \arg\max_y p(x|y)p(y) \end{aligned}$$

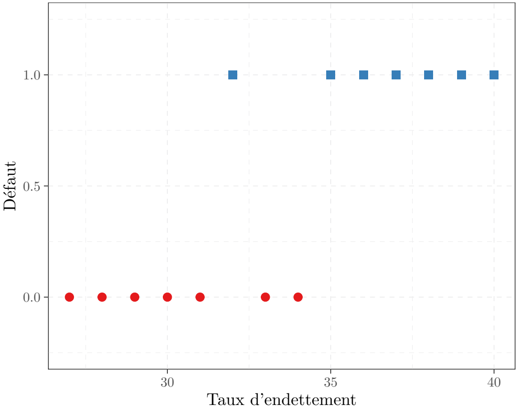
Analyse discriminante: intuition ($p = 1$)

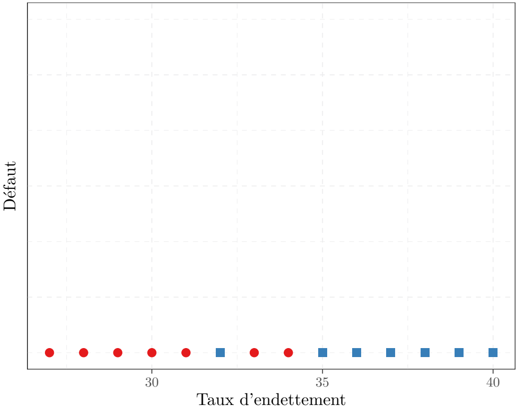
$$p(y \mid x) = \frac{p(y)p(x \mid y)}{p(x)}$$

Soit

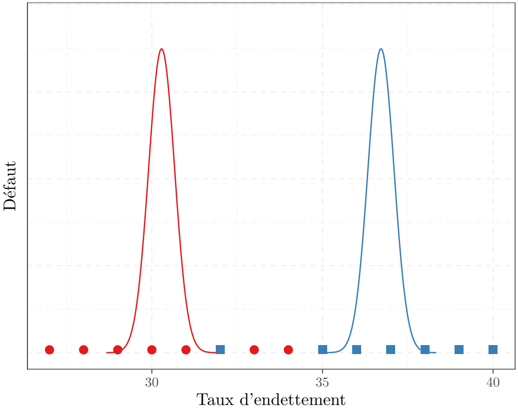
- $y \sim \mathcal{B}(\phi)$ donc $p(y) = \phi^y(1-\phi)^{1-y}$
- $x \mid y = 0 \sim \mathcal{N}\left(\mu_0, \sigma_0^2\right)$
- $x \mid y = 1 \sim \mathcal{N}\left(\mu_1, \sigma_1^2\right)$

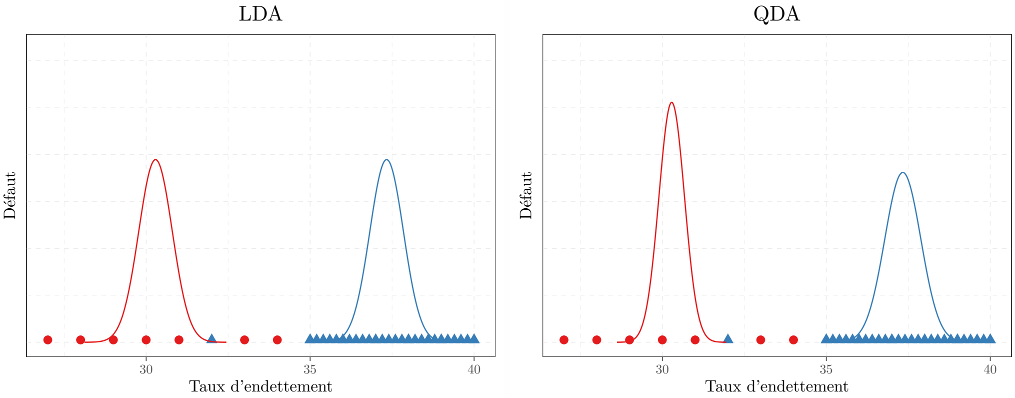
- Si $\sigma_0 = \sigma_1 \Rightarrow$ anaylse disiminante linéaire (LDA)
- Si $\sigma_0 \neq \sigma_1 \Rightarrow$ anaylse disiminante quadratique (QDA)


1.0
0.5
0.0
Défaut
30
35
40
Taux d'endettement


Défaut
30
35
40
Taux d'endettement

Analyse discriminante: intuition $(p = 1)$




LDA
QDA
Défaut
Défaut
30
35
40
30
35
40
Taux d'endettement
Taux d'endettement

$$p(y \mid x) = \frac{p(y)p(x \mid y)}{p(x)}$$

Soit

- $y \sim \mathcal{B}(\phi)$ donc $p(y) = \phi^{y}(1-\phi)^{1-y}$
- $x \mid y = 0 \sim \mathcal{N}(\mu_0, \Sigma_0)$
- $x \mid y = 1 \sim \mathcal{N}(\mu_1, \Sigma_1)$

Analyse discriminante: $(p \geqslant 1)$

$$p(y \mid x) = \frac{p(y)p(x \mid y)}{p(x)}$$

Soit

- $y \sim \mathcal{B}(\phi)$ donc $p(y) = \phi^y(1-\phi)^{1-y}$
- $x \mid y = 0 \sim \mathcal{N}(\mu_0, \Sigma_0)$
- $x \mid y = 1 \sim \mathcal{N}(\mu_1, \Sigma_1)$

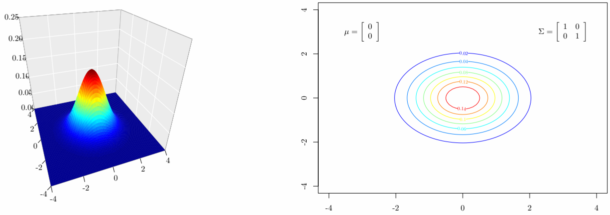
Ici, les gaussiennes sont multi-dimensionnelles. La densité de $\mathcal{N}(\mu, \Sigma)$ se définit comme suit:

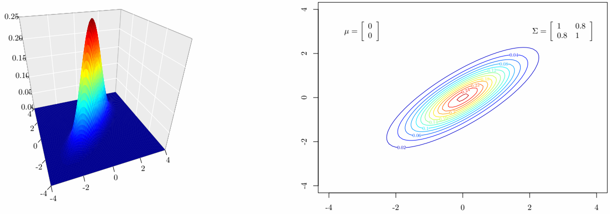
$$p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

Où :

- μ est le vecteur des moyennes, $\mu \in \mathbb{R}^d$
- Σ est la matrice de covariance, $\Sigma \in \mathbb{R}^{d \times d}$ avec Σ une matrice semi-définie positive

La loi Normale multi-dimensionnelle: exemples ($d = 2$)





- Les paramètres à estimer sont: ϕ, μ, et Σ

$$\phi = \begin{pmatrix} \phi_0 \\ \phi_1 \\ \vdots \\ \phi_k \end{pmatrix} \quad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix}$$

$$\Sigma = Cov(x) = \begin{pmatrix} \sigma_1^2 & Cov(x_1, x_2) & \dots & Cov(x_1, x_p) \\ Cov(x_2, x_1) & \sigma_2^2 & \dots & Cov(x_2, x_p) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(x_p, x_1) & Cov(x_p, x_2) & \dots & \sigma_p^2 \end{pmatrix}$$

- Étant donné les n observations $\{(x^{(i)}, y^{(i)})\}$ et en assumant qu'elles sont indépendantes, on estime ϕ, μ, et Σ qui maximisent la vraisemblance:

$$\begin{aligned}\mathcal{L}(\phi, \mu_0, \mu_1, \Sigma) &= \prod_{i=1}^{n} p\left(x^{(i)}, y^{(i)}; \phi, \mu_0, \mu_1, \Sigma\right) \\ &= \prod_{i=1}^{n} p(y^{(i)}; \phi) p\left(x^{(i)} | y^{(i)}; \mu_0, \mu_1, \Sigma\right) \\ &= \prod_{i=1}^{n} \phi^{y^{(i)}} (1-\phi)^{1-y^{(i)}} \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \\ &\quad \exp\left(-\frac{1}{2}(x^{(i)} - \mu_{y^{(i)}})^T \Sigma^{-1} (x^{(i)} - \mu_{y^{(i)}})\right)\end{aligned}$$

- Maximiser la vraisemblance revient à maximiser son log:

$$\ell(\phi, \mu, \Sigma) = \sum_{i=1}^{n} \Big(y^{(i)} \log \phi + (1 - y^{(i)}) \log(1 - \phi) - \frac{1}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma| - \frac{1}{2} (x^{(i)} - \mu_{y^{(i)}})^T \Sigma^{-1} (x^{(i)} - \mu_{y^{(i)}}) \Big)$$

- Maximiser ℓ possède une solution analytique. On résout le système:

$$\frac{\partial \ell}{\partial \phi} = 0, \quad \frac{\partial \ell}{\partial \mu} = 0, \quad \text{et} \quad \frac{\partial \ell}{\partial \Sigma} = 0$$

Analyse discriminante: estimation des paramètres

- En maximisant la log vraisemblance ℓ, on trouve (pour LDA):

$$\phi = \frac{1}{n}\sum_{i=1}^{n} 1_{\{y^{(i)}=1\}}$$

$$\mu_0 = \frac{\sum_{i=1}^{n} 1_{\{y^{(i)}=0\}} x^{(i)}}{\sum_{i=1}^{n} 1_{\{y^{(i)}=0\}}} \text{ et } \mu_1 = \frac{\sum_{i=1}^{n} 1_{\{y^{(i)}=1\}} x^{(i)}}{\sum_{i=1}^{n} 1_{\{y^{(i)}=1\}}}$$

$$\Sigma = \frac{1}{n}\sum_{i=1}^{n} \left(x^{(i)} - \mu_{y^{(i)}}\right)\left(x^{(i)} - \mu_{y^{(i)}}\right)^T$$

- En maximisant la log vraisemblance ℓ, on trouve (pour LDA):

$$\phi = \frac{1}{n}\sum_{i=1}^{n} 1_{\{y^{(i)}=1\}}$$

$$\mu_0 = \frac{\sum_{i=1}^{n} 1_{\{y^{(i)}=0\}} x^{(i)}}{\sum_{i=1}^{n} 1_{\{y^{(i)}=0\}}} \text{ et } \mu_1 = \frac{\sum_{i=1}^{n} 1_{\{y^{(i)}=1\}} x^{(i)}}{\sum_{i=1}^{n} 1_{\{y^{(i)}=1\}}}$$

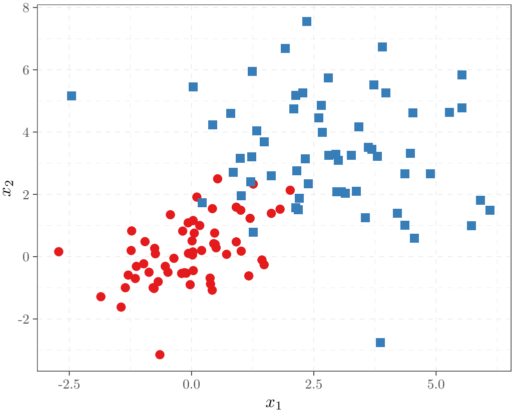
$$\Sigma = \frac{1}{n}\sum_{i=1}^{n}\left(x^{(i)} - \mu_{y^{(i)}}\right)\left(x^{(i)} - \mu_{y^{(i)}}\right)^T$$

Dans le cadre de la QDA:

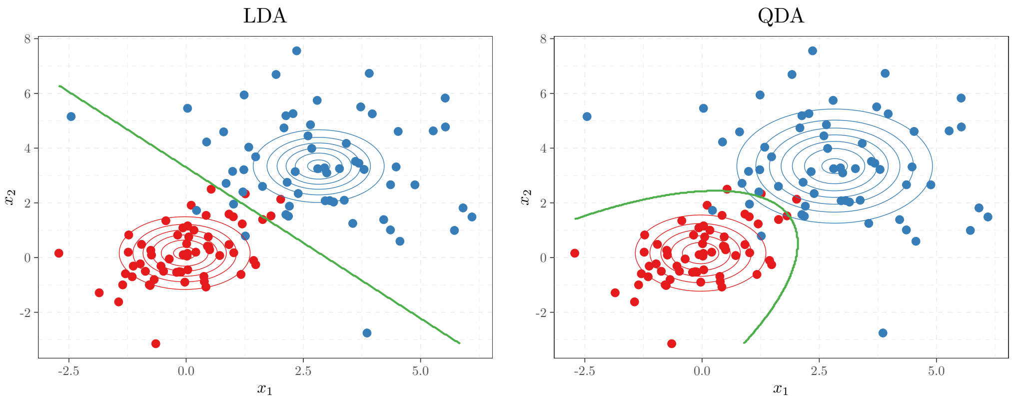
$$\Sigma_0 = \frac{1}{n_0}\sum_{i=1}^{n_0}(x^{(i)} - \mu_0)(x^{(i)} - \mu_0)^T 1_{\{y^{(i)}=0\}}$$

$$\Sigma_1 = \frac{1}{n_1}\sum_{i=1}^{n_1}(x^{(i)} - \mu_1)(x^{(i)} - \mu_1)^T 1_{\{y^{(i)}=1\}}$$

Analyse discriminante: frontières de séparation



- Les frontières de séparations sont définies par $p(y = 0|x) = p(y = 1|x)$


LDA
QDA
8
6
4
2
0
-2
-2.5
0.0
2.5
5.0
x_1
x_2

Analyse discriminante: Remarques & résumé

- LDA et QDA méthodes génératives
- Relation entre LDA et la régression logistique

$$\left.\begin{aligned} x \mid y = 0 &\sim \mathcal{N}(\mu_0, \Sigma) \\ x \mid y = 1 &\sim \mathcal{N}(\mu_1, \Sigma) \\ y &\sim \mathcal{B}(\phi) \end{aligned}\right\} \implies p(y = 1 \mid x) = \frac{1}{1 + e^{-\beta' x}}$$

- Lorsque l'hypothèse de normalité est satisfaite, la méthode générative est préférée