Régression logistique et Analyse discriminante

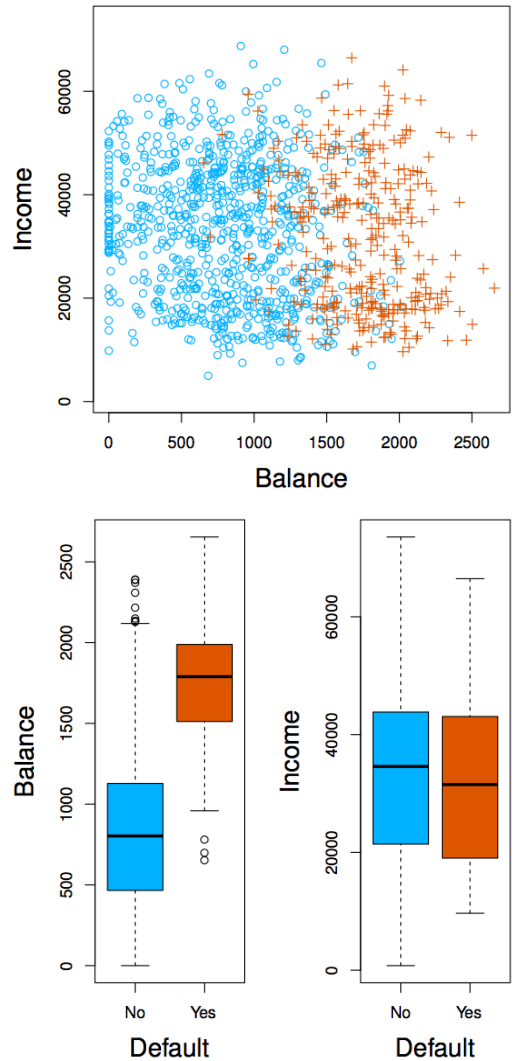
Joseph Salmon, Nicolas Verzelen

Université de Montpellier / INRA

Classification

- Expliquer une variable qualitative qui prend ses valeurs dans un ensemble non-ordonné $\mathcal{Y}$ tel que : `couleur.oeil`$\in\{$`brun`, `bleu`, `vert`$\}$
  `email`$\in\{$`spam`,`ham`$\}$
- **Objectif** : Etant donné un n-échantillon D_1^n, on veut construire un une règle de discrimination $\phi : \mathcal{X} \to \mathcal{Y}$.
- Parfois, on est plus intéressé par l'estimation des *probabilités* de chaque classe en x : $\mathbb{P}(y = 1|x)$

Exemple. Défaut sur carte de crédit



Plan

Régression logistique

Par convention, on prend $\mathcal{Y} = \{0, 1\}$ ici.
Notons $\eta^*(x) = \mathbb{P}(y|x) = \mathbb{E}[y|x]$ et supposons que $\mathcal{X} = \mathbb{R}$.

Régression logistique : cherche $\eta^*(\cdot)$ dans la classe $\mathcal{F}_{\log}$ de fonctions type

$$p(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

où β_0 et β_1 sont deux paramètres réels. On a alors

$$\log\left(\frac{p(x)}{1 - p(x)}\right) = \beta_0 + \beta_1 x \quad \text{(transformation logit)}$$

Maximum de vraisemblance

La vraisemblance : $L(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i:y_i=0} (1 - p(x_i))$

que l'on maximise pour obtenir β_0 et β_1 (pas de solution close ; algorithme itératifs de maximisation)

Packages : *cf.* `sklearn` en `Python` ou `glm` en `R`

Données de défaut de crédit

	coeff	Std.Err	Z-stat	p-value
Intercept	-10.6513	0.3612	-29.5	<0.0001
`balance`	0.0055	0.0002	24.9	<0.0001

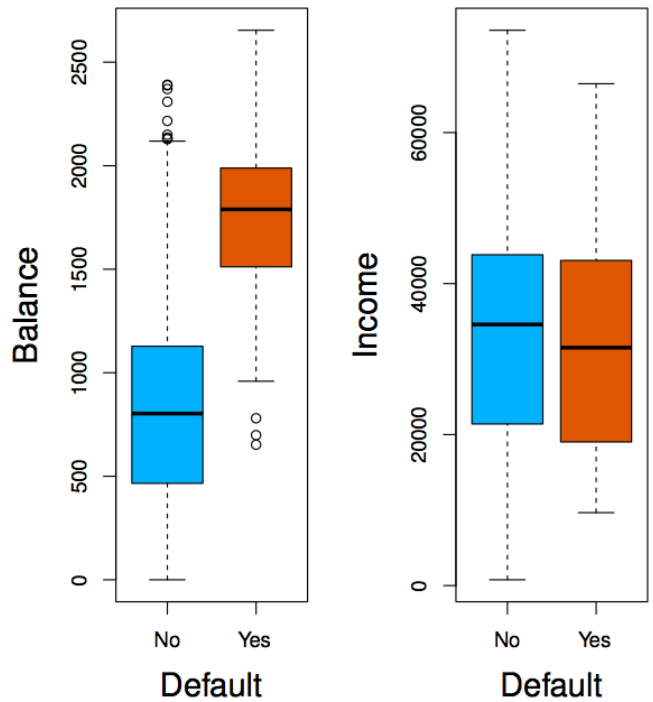
Faire des prédictions ?

Si `balance`$= 1000$?
$\widehat{p}(x) \approx 0.006$

Si `balance`$= 2000$?
$\widehat{p}(x) \approx 0.586$

Pourquoi ?

Remarque : Les écarts-types calculés et les p-valeurs reposent sur l'hypothèse $\eta^* \in \mathcal{F}_{log}$, i.e. le modèle de régression logistique est vrai.



Régression logistique à plusieurs co-variables

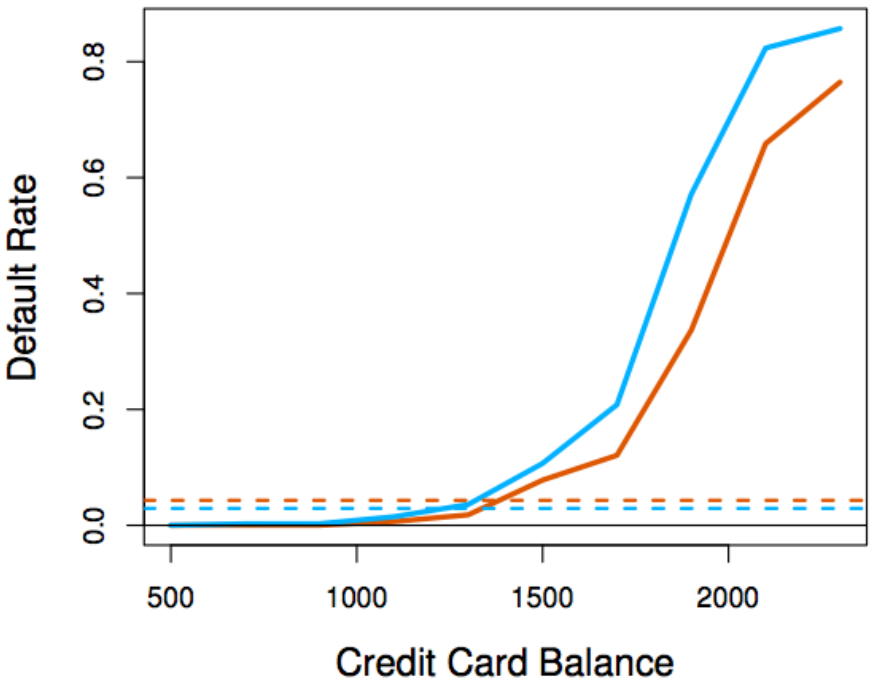
$$\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x^{(1)} + \cdots + \beta_p x^{(p)}$$

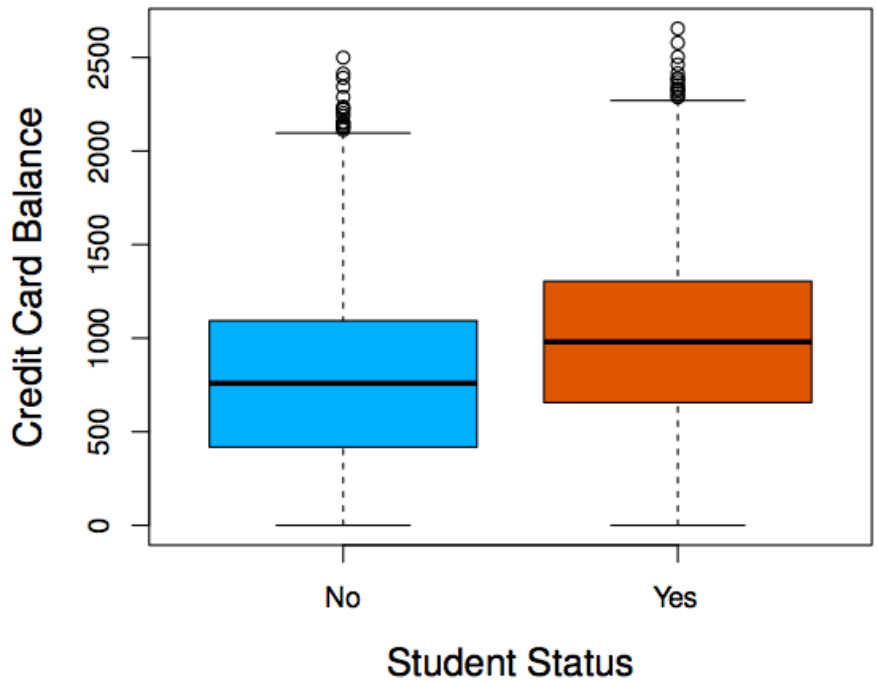
	coeff	Std.Err	Z-stat	p-value
Intercept	-10.8680	0.4923	-22.08	<0.0001
`balance`	0.0057	0.0002	24.74	<0.0001
`income`	0.0030	0.0082	0.37	0.7115
`student[Yes]`	-0.6468	0.2362	-2.74	0.0062

Si on n'utilise que `student` comme prédicteur, le coefficient est positif. Pourquoi ?

	coeff	Std.Err	Z-stat	p-value
Intercept	-3.5041	0.0707	-49.55	<0.0001
`student[Yes]`	0.4049	0.1150	3.52	0.0004

Effets confondants

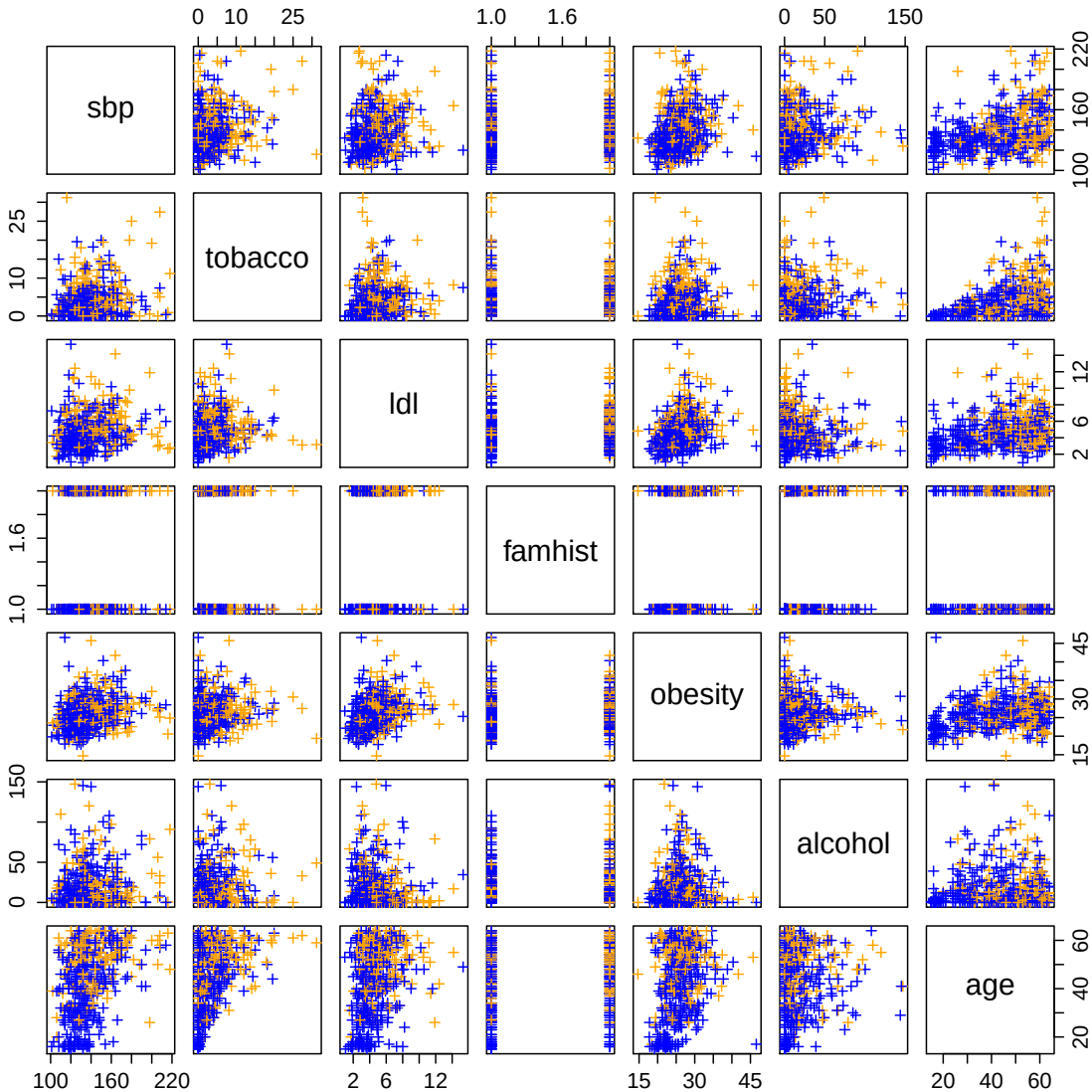




- Les étudiants tendent à avoir un solde (`balance`) plus élevé
- Mais à niveau de solde égal, les étudiants font moins défaut
- La régression logistique le montre.

Exemple : maladie cardiaque en Afrique du Sud

- 160 cas d'infarctus du myocarde (MI) et 302 cas de contrôle (homme entre 15-64 ans), de la province de Cap-Occidental en Afrique du Sud, au début des années 80
- Prévalence très élevée dans cette région : 5.1 %
- Mesure de 7 prédicteurs (facteurs de risque), montrés dans la page suivante
- Le but est d'identifier l'influence et la force relative des facteurs de risque
- Cette étude fait partie d'un programme de santé publique dont le but était de sensibiliser la population sur une régime plus équilibré



orange : MI

bleu : contrôle

`famhist` : 1 si antécédents familiaux

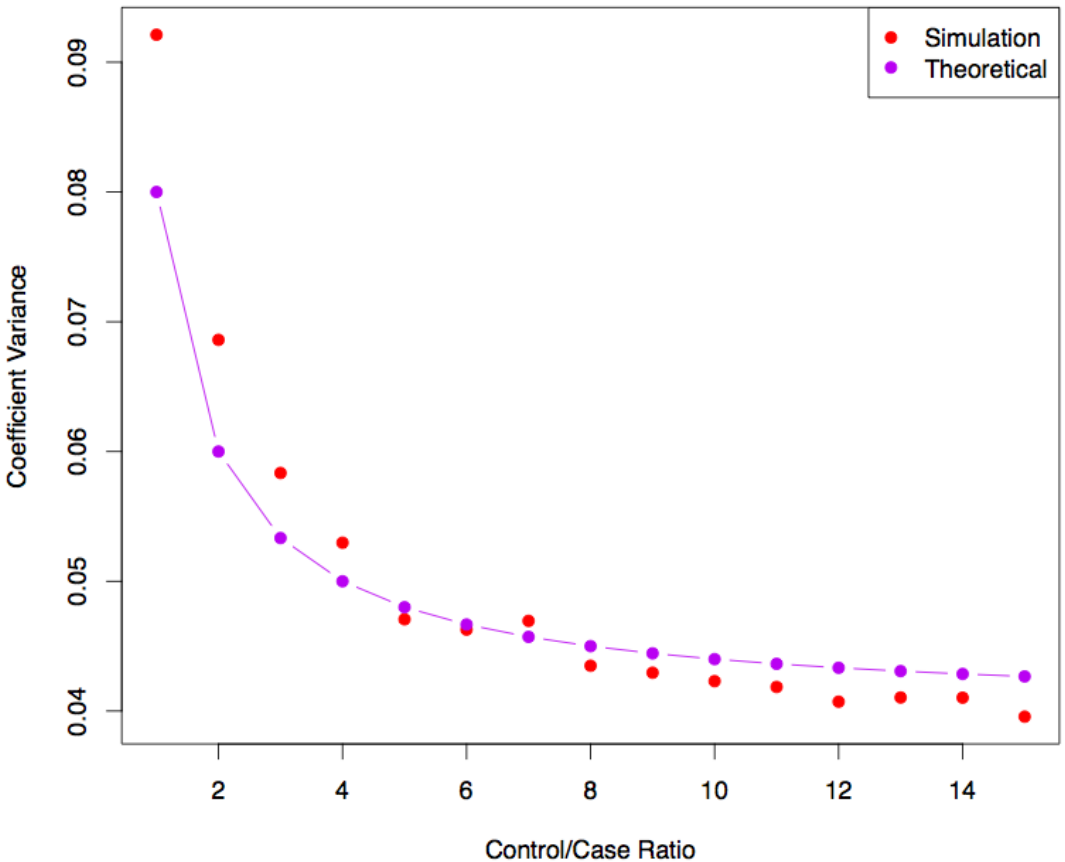
Échantillonnage du contrôle et régression logistique

- Dans les données d'Afrique du Sud, il y a 160 MI et 302 contrôle — $\tilde{\pi} = 0.35$ des cas. Cependant, la prévalence des MI dans la région est de $\pi = 0.05$.
- Ce biais d'échantillonnage permet d'estimer les β_j, $j \neq 0$, avec plus de précision (si modèle correct). Mais l'estimation de β_0 doit être corrigée.
- Une simple transformation permet de le faire :

$$\widehat{\beta_0}^* = \widehat{\beta_0} + \log\left(\frac{\pi}{1-\pi}\right) - \log\left(\frac{\tilde{\pi}}{1-\tilde{\pi}}\right)$$

- Souvent, les cas pathologiques sont rares et on les prend tous. On peut sur-échantillonner jusqu'à 5 fois plus que les cas témoins. Au delà, peu de gain dans la variance d'erreur d'échantillonnage.

Gain de variance par biais d'échantillonnage de données binaires



Au delà d'un facteur 5 de sur-représentation des cas pathologiques, le gain n'est plus intéressant.

Régression logistique à plus de deux modalités

Jusqu'à maintenant, nous avons discuté de régression logistique pour expliquer un y à deux modalités. Il est facile de généraliser à plus de deux classes. Une possibilité est la forme symétrique

$$\mathbb{P}(y = k|x) = \frac{\exp(\beta_{0k} + \beta_{1k}x^{(1)} + \cdots + \beta_{pk}x^{(p)})}{\sum_{\ell=1}^{K} \exp(\beta_{0\ell} + \beta_{1\ell}x^{(1)} + \cdots + \beta_{p\ell}x^{(p)})}$$

Il y a donc une fonction linéaire par classe ou modalité.

En fait, ce modèle est sur-paramétré, et comme dans le cas de 2 classes, on peut supprimer l'une des fonctions linéaires et seules $(K-1)$ sont utiles. *Le vérifier !*

La régression logistique multi-classe porte plusieurs noms. On parle parfois de régression multinomiale.

Plan

Analyse discriminante

Ici, $\mathcal{Y} = \{1, \dots, K\}$. L'idée est de modéliser la loi de x dans chaque classe séparément, et d'utiliser le *théorème de Bayes* pour obtenir $\eta_k^*(x) = \mathbb{P}(y = k|x)$.

Lorsque l'on utilise des lois gaussiennes pour chaque classe, cela donne l'analyse discriminante linéaire ou quadratique.

Notons

- $P_{x,k}$ la distribution de x conditionnellement à $y = k$;
- π_k la probabilité (marginale) de la classe k, i.e. $\mathbb{P}(y = k)$.

Le théorème de Bayes entraîne alors

$$\mathbb{P}(y = k|x = x_0) = \pi_k \frac{dP_{x,k}}{\sum_{\ell=1}^{K} \pi_\ell dP_{x,\ell}}(x_0).$$

Cette décomposition ne repose sur aucune hypothèse sur la distribution P des données !

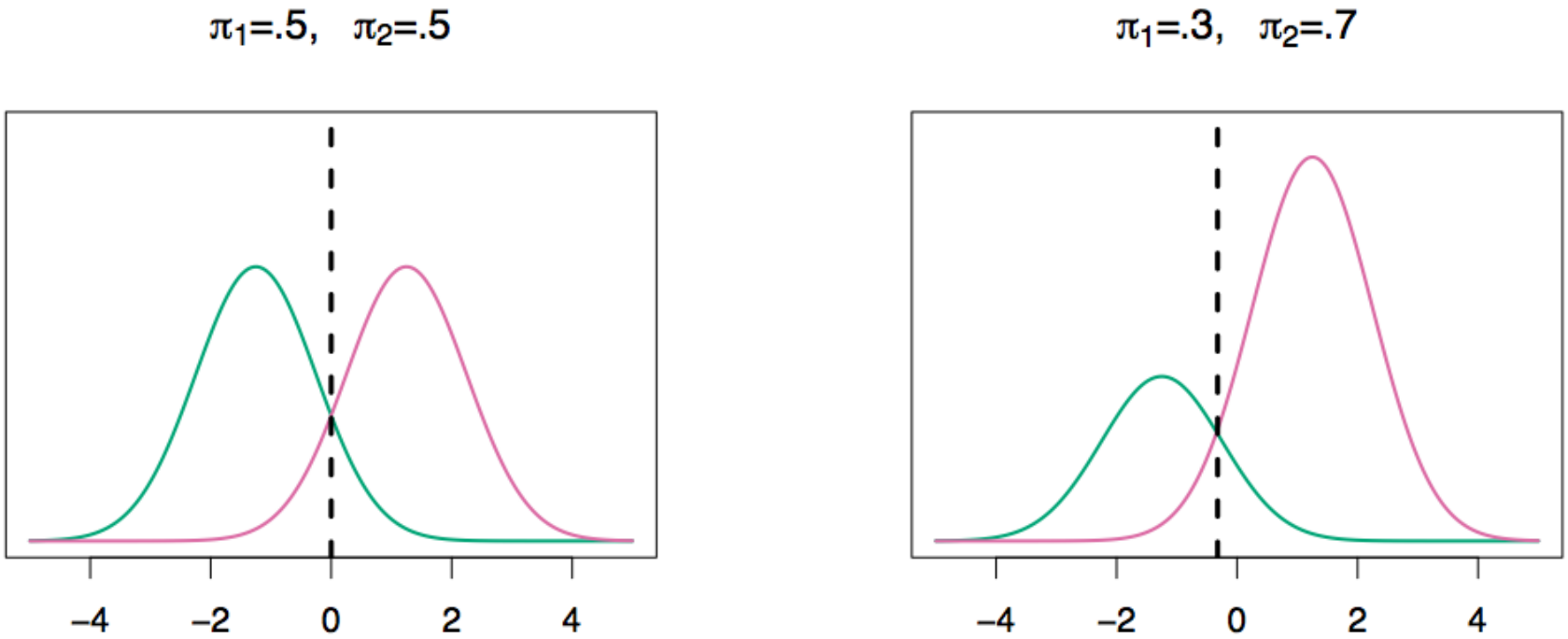
Analyse discriminante linéaire et quadratique

L'approche repose sur l'hypothèse que pour tout $k \in \mathcal{Y}$, $P_{x,k}$ est une distribution gaussienne (on notera leur densité f_k), de telle sorte qu'on cherche à estimer les fonctions $\eta_k^* = \mathbb{P}(y = k|x)$ par des fonctions de la forme

$$p_k(x) = \frac{\pi_k f_k(x)}{\sum_{\ell=1}^{k} \pi_\ell f_\ell(x)} \ ,$$

où (π_k) vérifie $\sum_k \pi_k = 1$ et les f_k sont des densités gaussiennes.

Classifier dans la classe de densité la plus élevée



On classifie une nouvelle observation dans la classe de densité la plus élevée.

Lorsque les fréquences (marginales) des classes sont différentes, il faut prendre en compte cette différence, et l'on compare les $\pi_k f_k(x)$.

Sur l'exemple ci-dessus, la classe des roses a une probabilité marginale plus élevée. La frontière entre les deux décisions s'est déplacée sur la gauche.

Pourquoi utiliser l'analyse discriminante linéaire ?

- Lorsque les classes sont bien séparées, l'estimation des paramètres de la régression logistique devient instable. L'analyse discriminante linéaire (LDA) ne souffre pas de ce problème.
- Lorsque n est petit et la distribution de x est à peu près gaussienne dans chaque classe, LDA est plus stable que la régression logistique.
- LDA est aussi populaire lorsqu'il y a plus de deux modalités pour y car elle permet de projeter les données dans des plans séparent les groupes.

En dimension 1

On cherche à estimer $\mathbb{P}(y = k|x)$ par $p_k(x)$ avec f_k de la forme

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

LDA suppose que $\sigma_1 = \cdots = \sigma_K = \sigma$.

Ce qui donne

$$p_k(x) = \frac{\pi_k e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma}\right)^2}}{\sum_{\ell=1}^{K} \pi_\ell e^{-\frac{1}{2}\left(\frac{x-\mu_\ell}{\sigma}\right)^2}}$$

après avoir simplifier par $1\Big/\sqrt{2\pi}\sigma$ en facteur partout.

En dimension 1

On cherche $p_k(x) = \mathbb{P}(y = k|x)$ avec

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

LDA suppose que $\sigma_1 = \cdots = \sigma_K = \sigma$.

Ce qui donne

$$p_k(x) = \frac{\pi_k e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma}\right)^2}}{\sum_{\ell=1}^{K} \pi_\ell e^{-\frac{1}{2}\left(\frac{x-\mu_\ell}{\sigma}\right)^2}}$$

après avoir simplifier par $1\Big/\sqrt{2\pi}\sigma$ en facteur partout.

Pour construire une régle de discrimination ϕ en x, il faut maintenant chercher quel $p_k(x)$ est le plus grand (Pourquoi ? Quelle est la fonction de perte associée ?)

Noter que les dénominateurs de dépendent pas de k.

En passant au $\log$, il suffit de comparer les

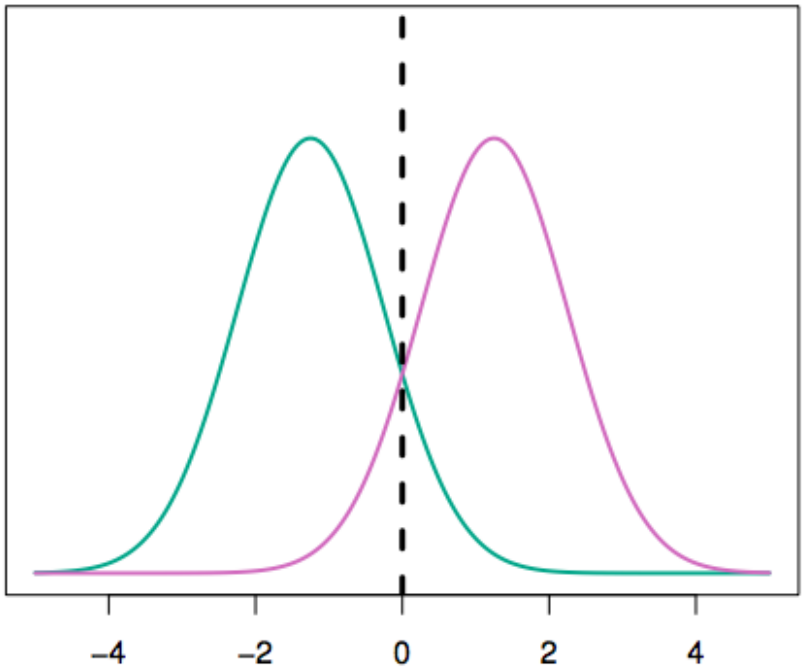
$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k).$$

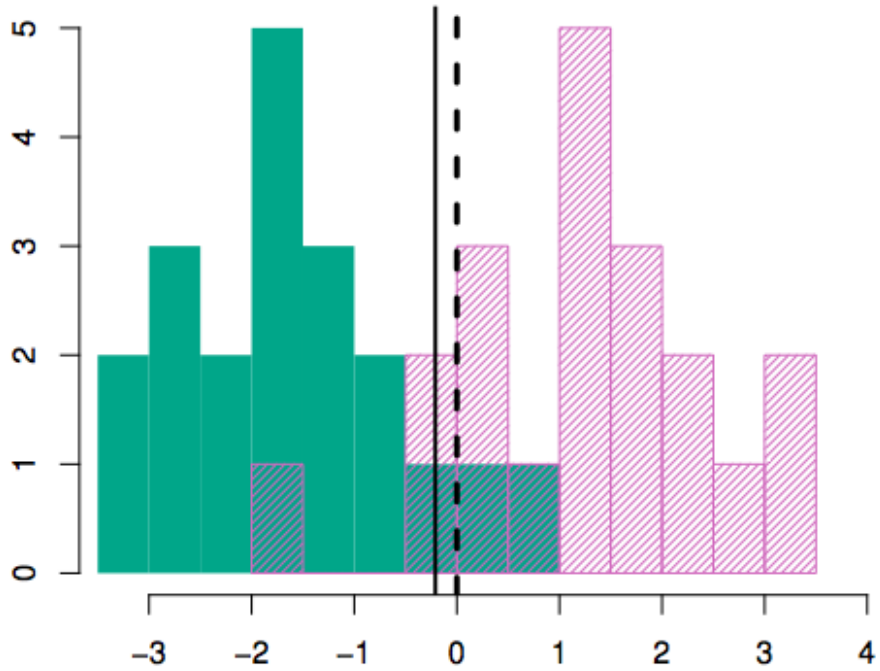
Ces fonctions **linéaires** de x s'appellent les scores de Fisher.

Si $K = 2$, et $\pi_1 = \pi_2 = 0.5$, vérifier que la frontière de décision est en

$$x = \frac{\mu_1 + \mu_2}{2}.$$

Analyse discriminante : estimation





Exemple simulé avec $\mu_1 = -1.5$, $\mu_2 = 1.5$, $\pi_1 = \pi_2 = 0.5$, et $\sigma^2 = 1$.

Typiquement, on ne connait pas ces paramètres et on doit les apprendre sur des données. Dans ce cas, on estime simplement les paramètres et on « *plug* » les estimations dans les formules théoriques.

Analyse discriminante linéaire (suite)

$$\widehat{\pi_k} = n_k/n$$

$$\widehat{\mu_k} = \sum_{i:y_i=k} x_i \Big/ n_k$$

$$\widehat{\sigma_k^2} = \sum_{i:T_i=k} \left(x_i - \widehat{\mu_k}\right)^2 \Big/ (n_k - 1)$$

$$\widehat{\sigma^2} = \sum_{k=1}^{K} \frac{n_k - 1}{n - K} \cdot \widehat{\sigma_k^2}$$

En dimension $p > 1$, les formules se généralisent.

Rappelons

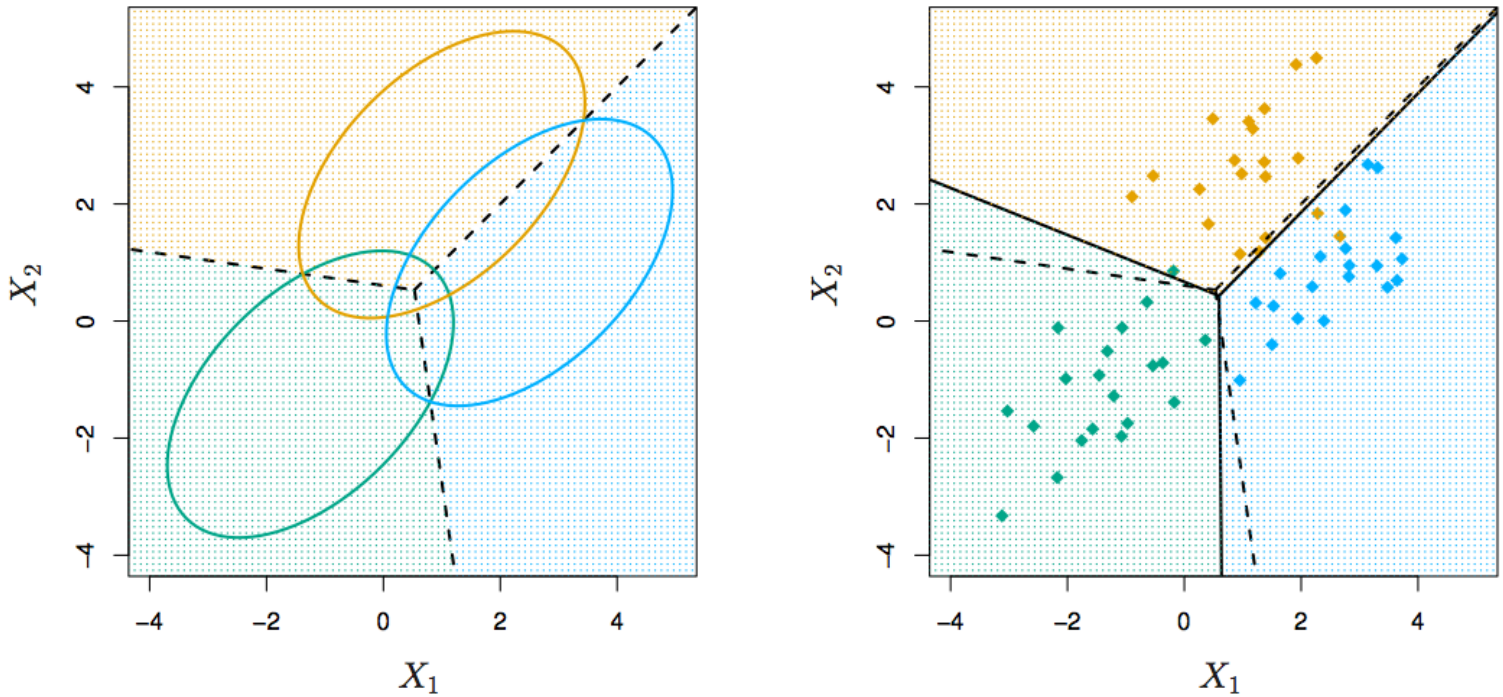
$$f(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$$

Les scores deviennent

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

qui est toujours une fonction linéaire en x.

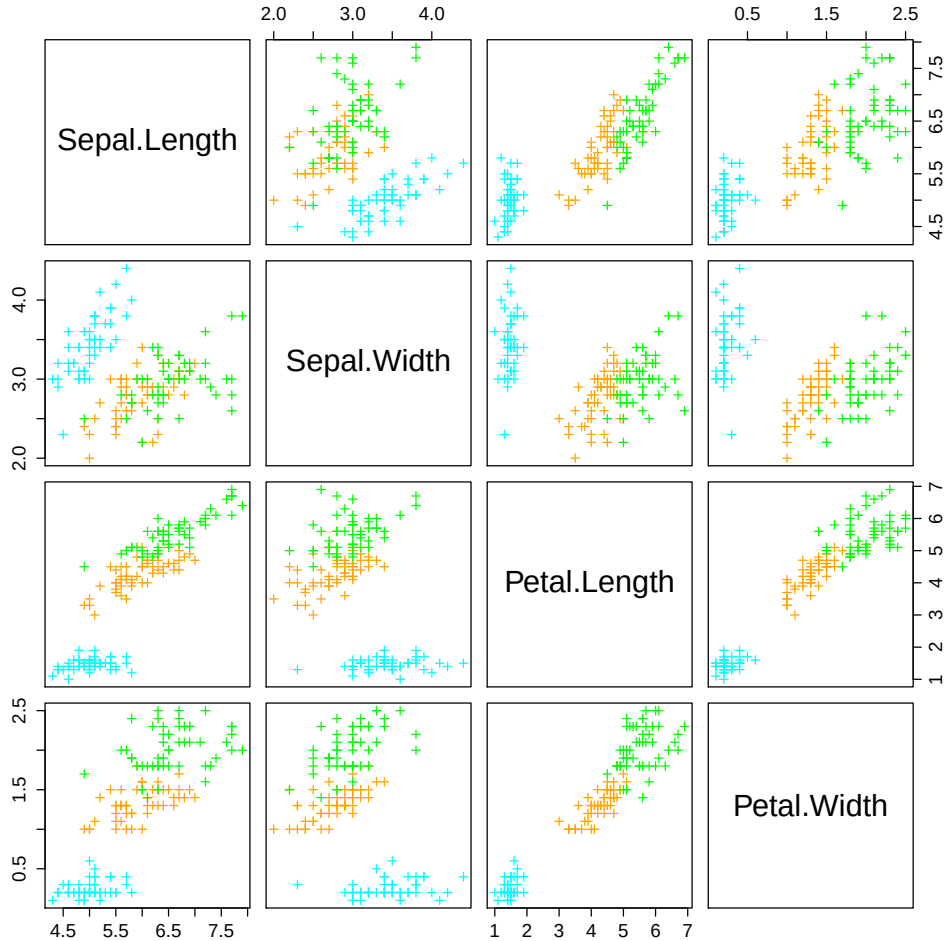
Exemple simulé en dimension 2



Exemple simulé $\pi_1 = \pi_2 = \pi_3 = 1/3$.

Les lignes pointillées : frontières théoriques
Les lignes pleines : frontières estimées par LDA

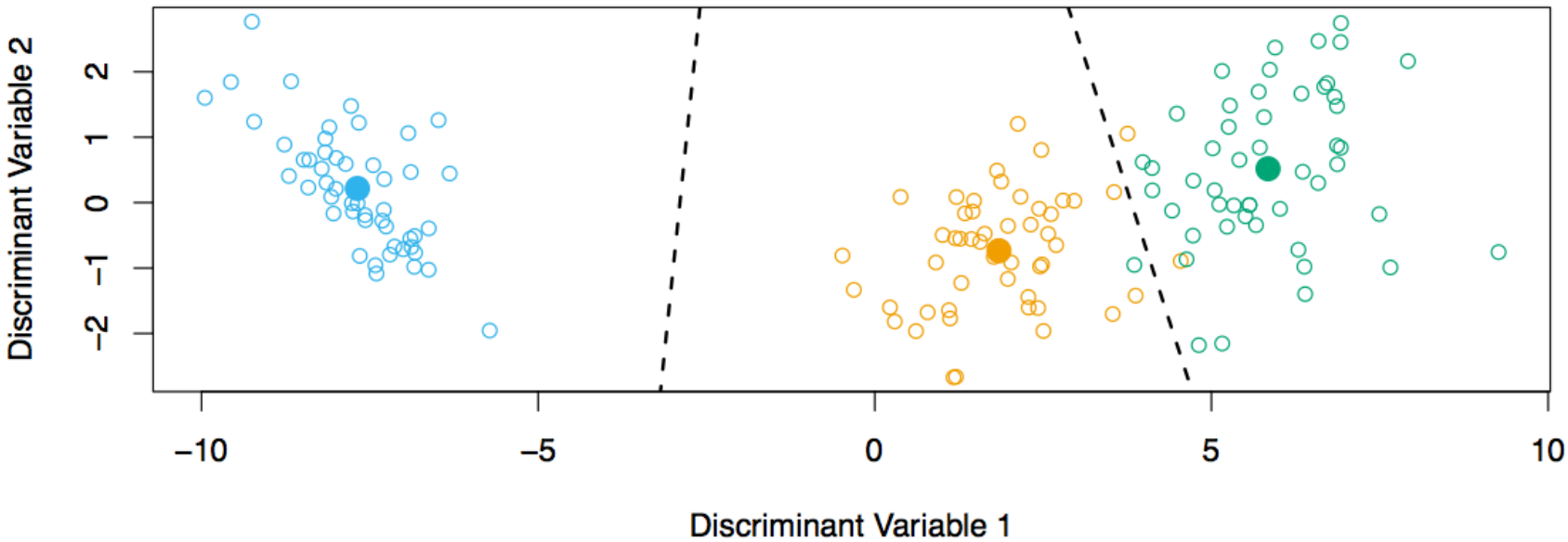
Autre exemple : les Iris de Fisher



4 variables explicatives
3 espèces
50 observations/classes

bleu : « setosa »
orange : « versicolor »
vert : « virginica »

Plan discriminant des Iris



Lorsqu'il y a K groupes, LDA fournit une projection de dimension $(K-1)$ qui sépare au mieux les nuages de points.

$(K-1)$? Essentiellement parce les centres des classes μ_1, μ_2 et μ_3 (ou plutôt leurs estimations) définissent un plan dans l'espace.

LDA sur les défauts de cartes de crédit

Matrice de confusion

		Vrai		
		`No`	`Yes`	Total
Prédit	`No`	9644	252	9896
	`Yes`	23	81	104
	Total	9667	333	10000

Soit $(23 + 252)/10000$ erreur — un taux de mauvaise classif de 2.75%.

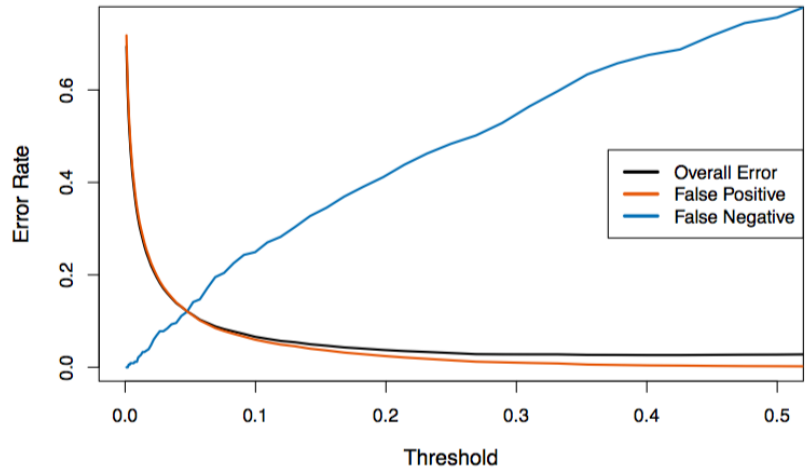
Quelques avertissements :

- C'est l'erreur sur l'échantillon d'apprentissage, et il peut y avoir du sur-apprentissage !
  Pas un gros problème ici car $n = 10000$ et $p = 4$.
- Si on tire au hasard suivant les fréquences dans la population, on ferait une erreur d'environ $333/10000 = 3.33\%$.
- Parmi les vrais `No`, on fait une erreur de $23/9667 = 0.2\%$.
  Parmi les vrais `Yes`, on fait une erreur de $252/333 = 75.7\%$! ! !

Types d'erreurs et coût

Taux de faux positifs : ici 0.2%
Taux de faux négatifs : ici 75.7%

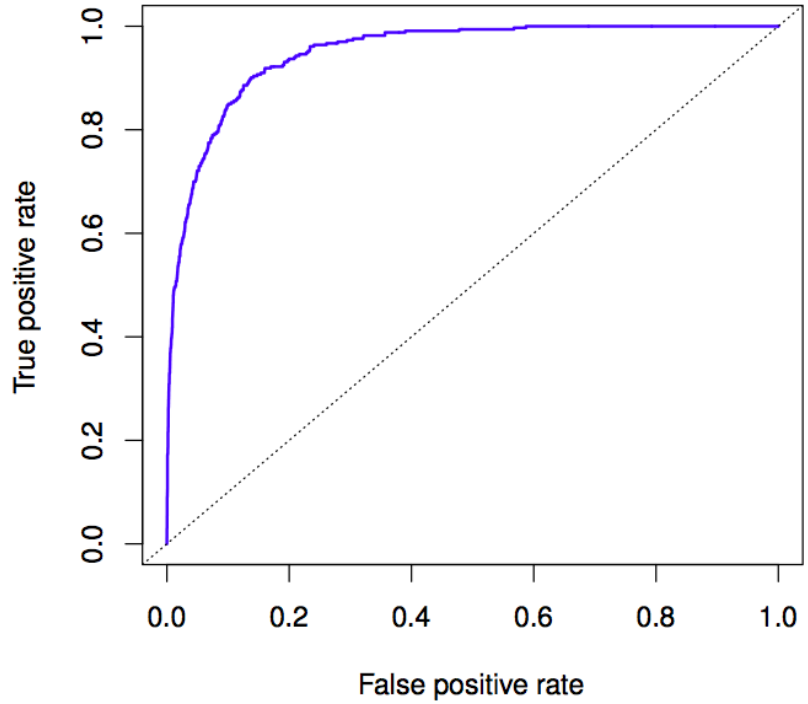
La règle de décision a décidé pour `Yes` si
$\widehat{\mathbb{P}}(y = \texttt{Yes} | \texttt{Balance}, \texttt{Student}) \geq 0.5$



Voici comment évolue le taux de faux positifs et de faux négatifs lorsque ce seuil varie.

On peut remplacer ce 0.5 par n'importe quelle *seuil* entre 0 et 1 : on décide maintenant pour `Yes` si
$\widehat{\mathbb{P}}(y = \texttt{Yes} | \texttt{Balance}, \texttt{Student}) \geq \texttt{SEUIL}$

Courbes ROC



La courbe ROC représente le taux de vrais positifs en fonction du taux de faux positifs lorsque le SEUIL varie.

Question : *Peut-on utiliser l'échantillon d'apprentissage pour construire la courbe ROC ?*

Parfois, on utilise l'aire totale sous la courbe (AUC) pour résumer la performance globale.
Plus cette aire est grande, plus les performances sont bonnes.

Autres formes d'analyse discriminante

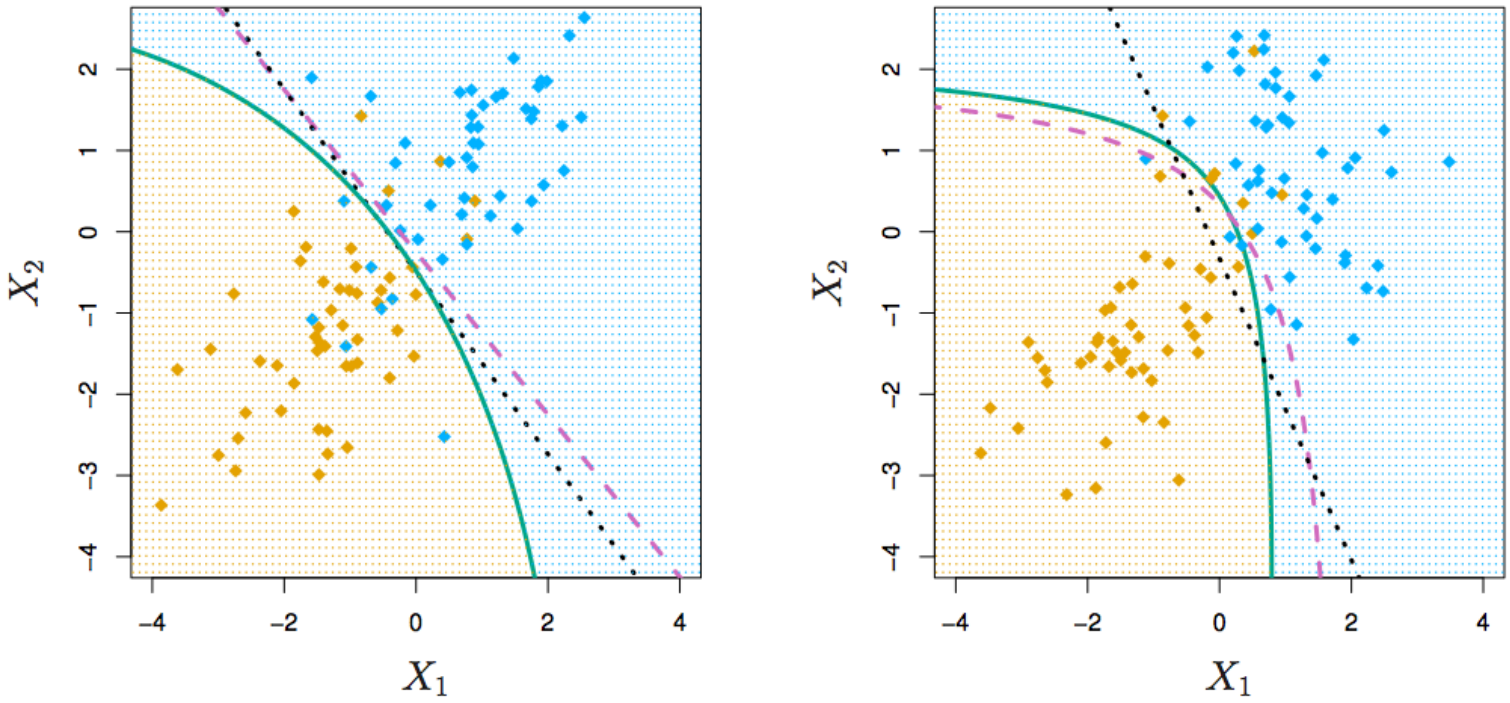
$$p_k(x) = \frac{\pi_k f_k(x)}{\sum_{\ell=1}^{K} \pi_\ell f_\ell(x)}$$

Lorsque les $f_k(x)$ sont des gaussiennes de même variance, LDA.

En changeant la forme des $f_k(x)$, on obtient d'autres classifieurs.

- Avec des gaussiennes, mais de variances distinctes, on obtient l'*analyse discriminante quadratique*
- Avec des $f_k(x) = \prod_{j=1}^{p} f_{jk}(x^{(j)})$ qui se factorisent par co-variables, on obtient *naive Bayes*
- Beaucoup d'autres cas, y compris non-paramétriques.

Analyse discriminante quadratique



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

Comme les Σ_k sont différents, des termes quadratiques apparaissent.

Naive Bayes

Suppose que les co-variables sont indépendantes conditionnellement à chaque classe.

Utile quand p grand et que des méthodes comme QDA ou LDA tombent.

- Le cas gaussien revient à supposer que les Σ_k sont diagonales

$$\delta_k(x) \propto \log\left(\pi_k \prod_{j=1}^{p} f_{kj}(x^{(j)})\right) = -\frac{1}{2}\sum_{j=1}^{p} \frac{(x^{(j)} - \mu_{jk})^2}{\sigma_{kj}^2} + \log(\pi_k)$$

- peut servir dans le cas mixte où certaines co-variables sont qualitatives. Dans ce cas, on remplace les densités intra-classes correspondantes par des fréquences intra-classes.

Même si l'hypothèse de départ semble très forte, mérite d'être testé car donne souvent de bons résultats.

Régression logistique ou LDA ?

Pour un problème à deux classes, LDA vérifie

$$\log\left(\frac{p_1(x)}{1-p_1(x)}\right) = \log\left(\frac{p_1(x)}{p_2(x)}\right) = c_0 + c_1 x^{(1)} + \cdots + c_P x^{(p)}$$

Même forme que pour la régression logistique.

Différence : méthode d'estimation des paramètres

- La régression logistique utilise des vraisemblances conditionnelles et ne modélise que $\mathbb{P}(y = k|x)$ — on parle d'*apprentissage discriminant*
- LDA utilise une vraisemblance complète basée sur la loi jointe de (x, y) — on parle d'*apprentissage génératif*
- Malgré ces différences, en pratique, les résultats sont souvent similaires.

Conclusion et...

- La régression logistique est très populaire pour la classification, surtout lorsqu'il n'y a que $K = 2$ classes
- LDA est utile, même lorsque n est petit, ou lorsque les classes sont bien séparées et que l'hypothèse gaussienne est raisonnable. Et lorsque $K > 2$.
- *Naive Bayes* est utile lorsque p est grand
- En petite dimension ($p < 4$), et avec beaucoup d'observations, on peut faire de l'analyse discriminante non-paramétrique, en remplaçant les gaussiennes par des densités intra-classes estimées (par une méthode à noyau)
- Autres méthodes (modèles additifs généralisés, SVM, random forest, etc.) dans les cours suivants.