Analyse discriminante et règle de classification de Bayes naïve

masedki.github.io

27 novembre 2019

Analyse discriminante

- L'idée est modéliser la loi de X dans chaque classe séparément, et d'utiliser le *théorème de Bayes* pour obtenir $\mathbb{P}(Y = k|X)$.
- Lorsque l'on utilise des lois gaussiennes pour chaque classes, cela donne l'analyse discriminante linéaire ou quadratique.
- Cette approche est plus générale, et l'on pourrait utiliser d'autres distributions. Concentrons nous d'abord sur le cas gaussien.

Analyse discriminante

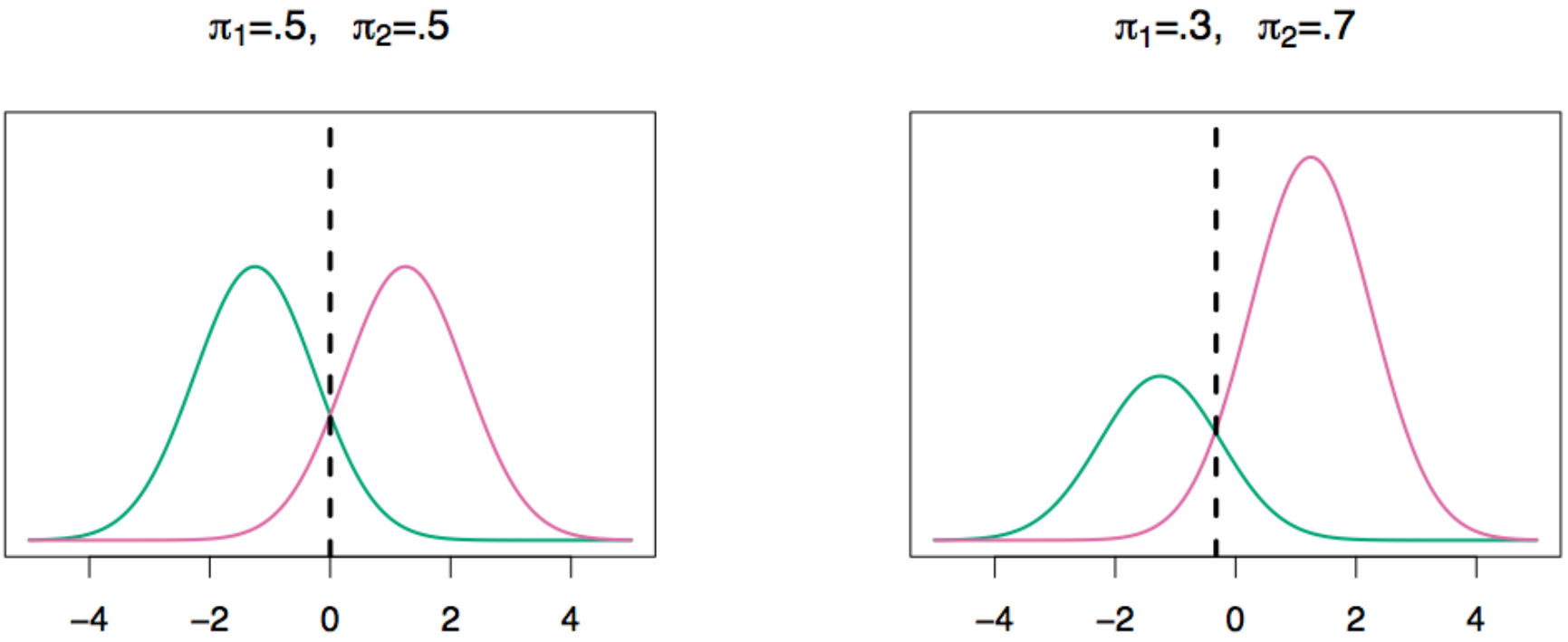
Notons

- $f_k(x)$ la *densité* de X sachant que l'on est dans la classe k ;
- π_k la probabilité (marginale) de la classe k

On a alors

$$\mathbb{P}(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{\ell=1}^{K} \pi_\ell f_\ell(x)}.$$

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 $p_k(x) = \mathbb{P}(Y = k | X = x)$ 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 $1/\sqrt{2\pi}\sigma$ en facteur partout.

En dimension 1

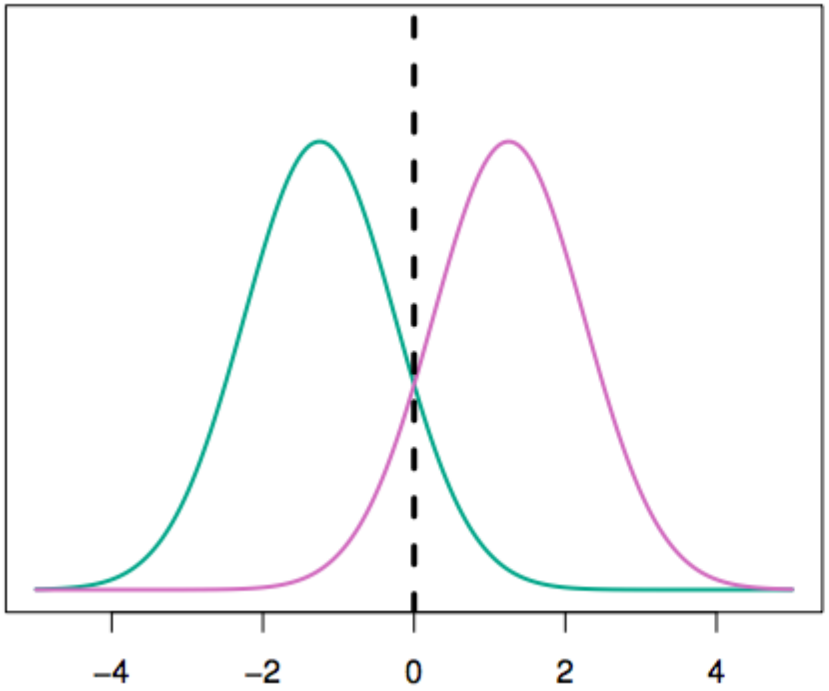
- Pour construire un classificateur en $X = x$, il faut maintenant chercher quel $p_k(x)$ est le plus grand.
- Noter que les dénominateurs de dépendent pas de k.
- En passant au log, il suffit de comparer les

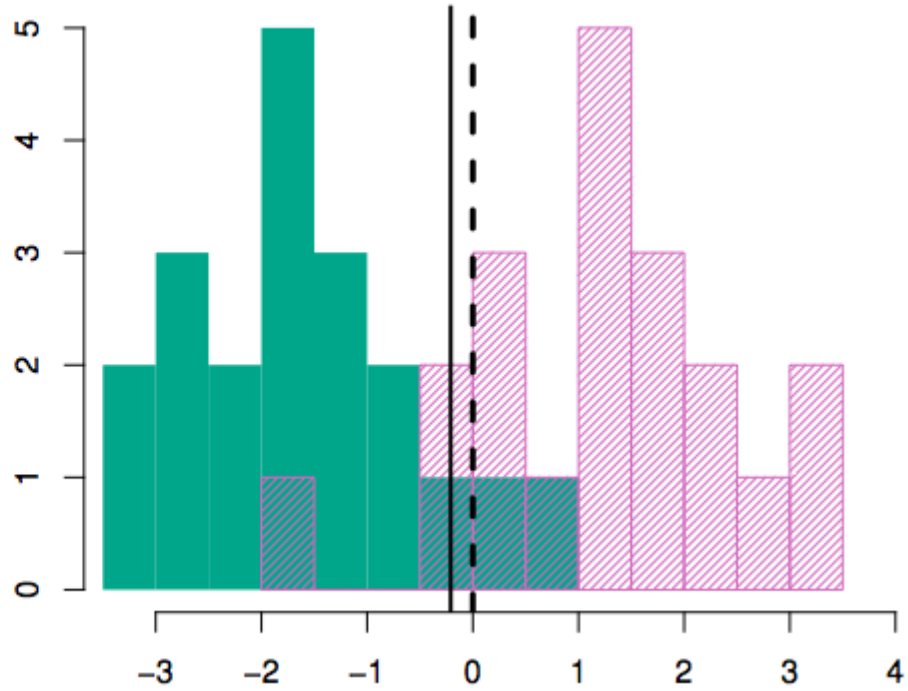
$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k).$$

- 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 « substitue » les estimations dans les formules théoriques.

Analyse discriminante linéaire (suite)

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

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

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

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

Analyse discriminante linéaire (suite)

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

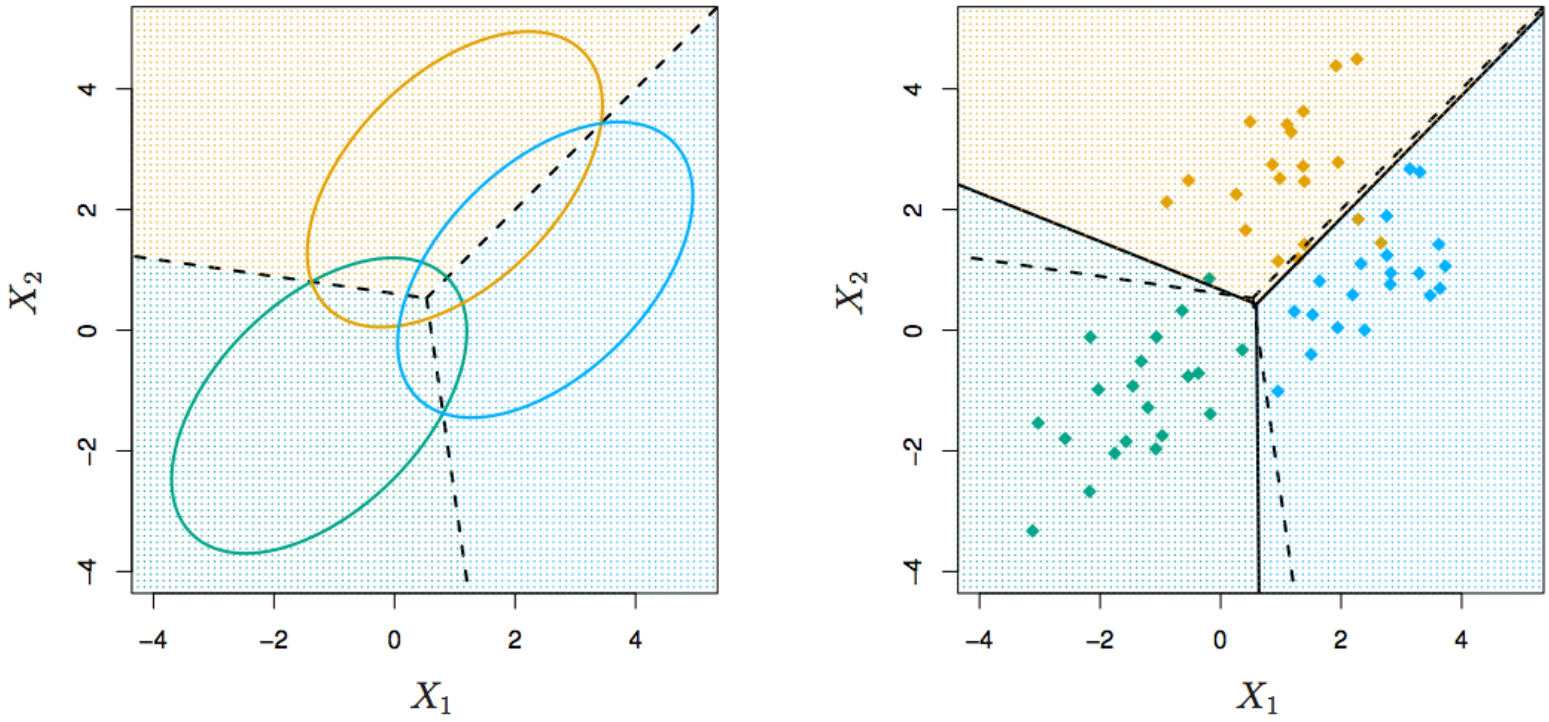
$$f(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$$

Les scores deviennent

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

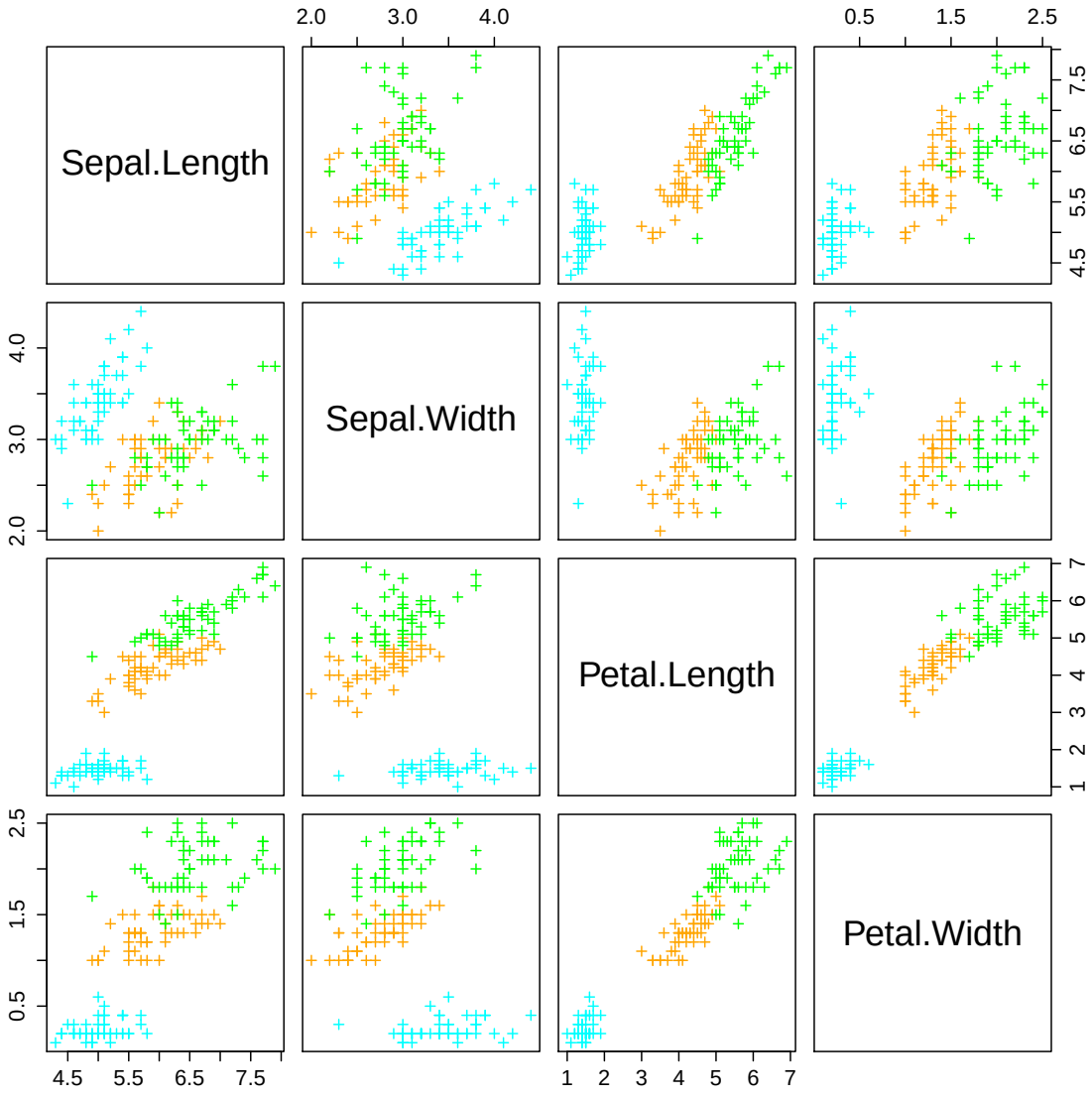
qui est toujours une fonction linéaire en x.

Exemple simulé en dimension 2



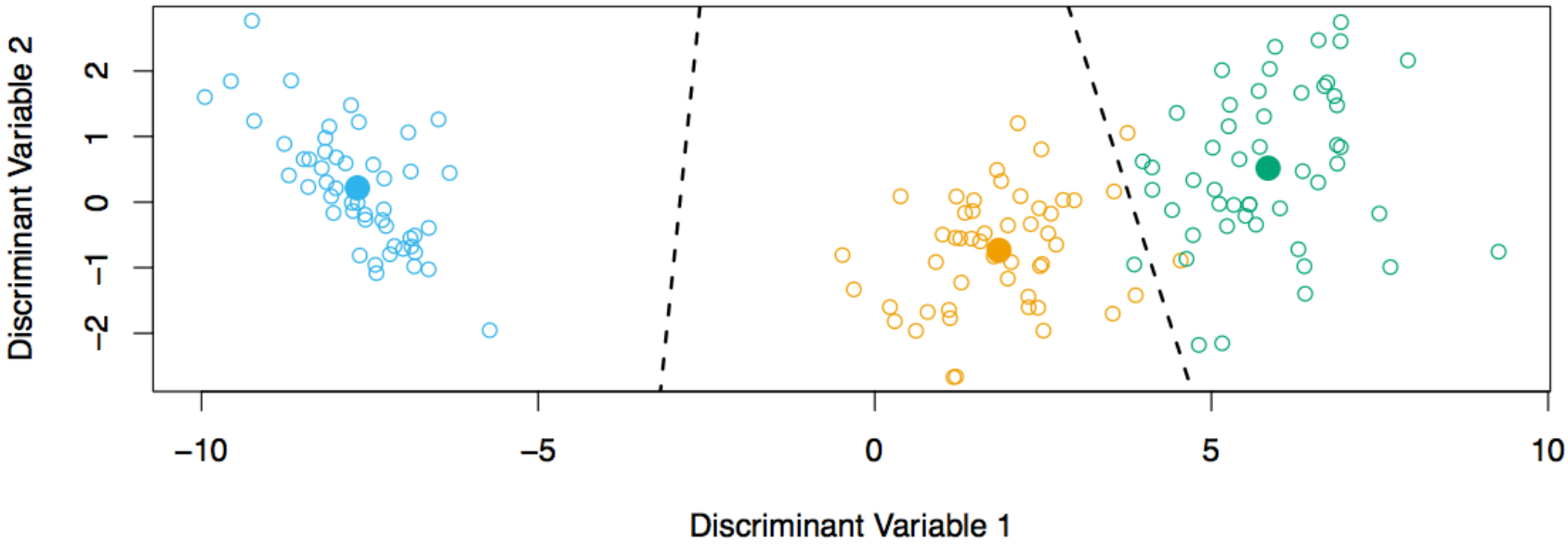
Exemple simulé $\pi_1 = \pi_2 = \pi_3 = 1/3$.
Les lignes pointillées : frontières théoriques
Les lignes pleines : frontières estimées par LDA

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.

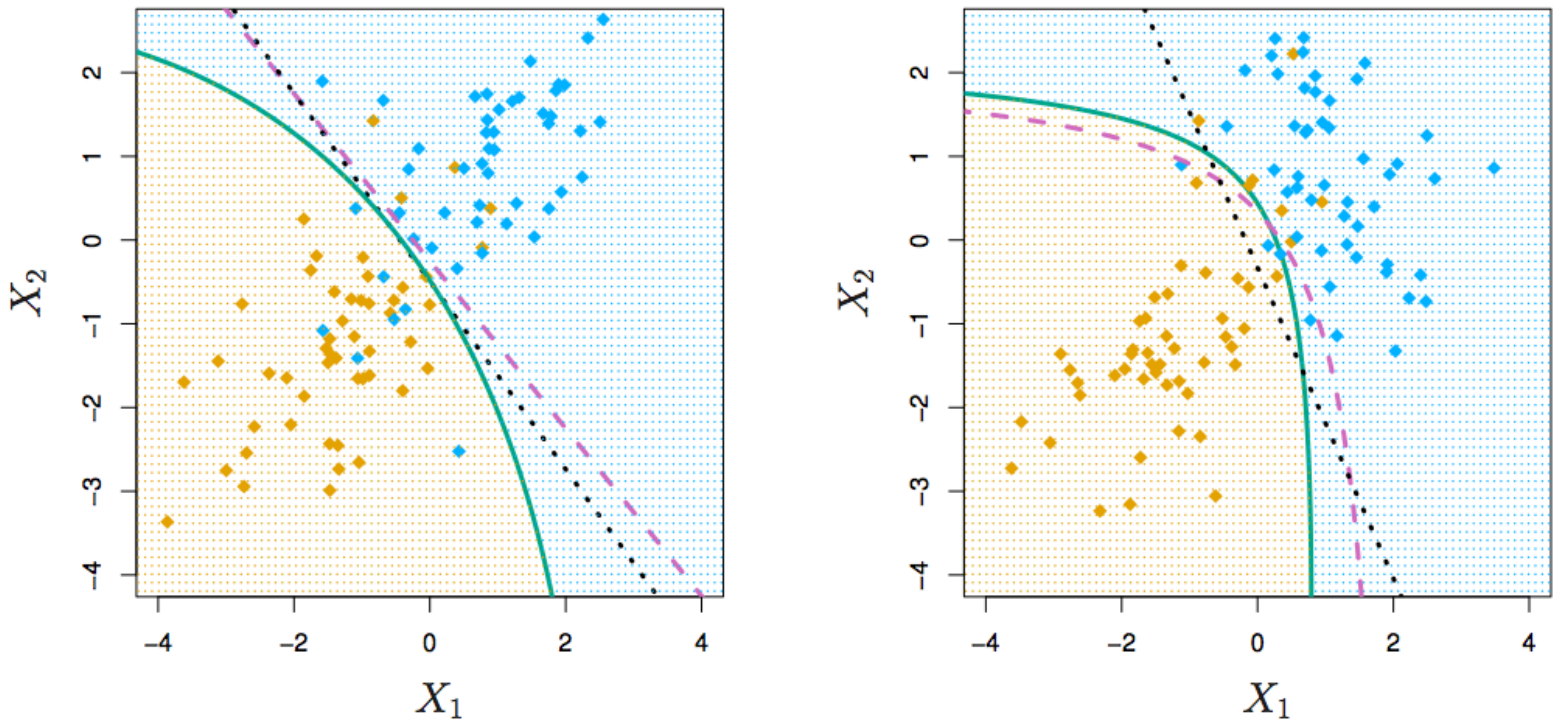
Autres formes d'analyse discriminante

$$\mathbb{P}(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{\ell=1}^{K} \pi_\ell f_\ell(x)}$$

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 discrimante 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 replace 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 a 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.
La différence est dans la façon dont sont estimés les 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 $K > 2$
- *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.