Apprentissage supervisé
Présentation générale.

Marie Chavent

Université de Bordeaux

3 septembre 2019

Plan

1. Références et exemples.
2. Une méthode simple : les k plus proches voisins
3. Cadre général : la théorie bayésienne de la décision et la règle de Bayes.
4. Evaluer les performances d'une règle de décision.

Ressources

- Livres de référence

Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani (2013).

An Introduction to Statistical Learning

Springer. `http://www-bcf.usc.edu/~gareth/ISL/`

Trevor Hastie, Robert Tibshirani et Jerome Friedman (2009).

The Elements of Statistical Learning

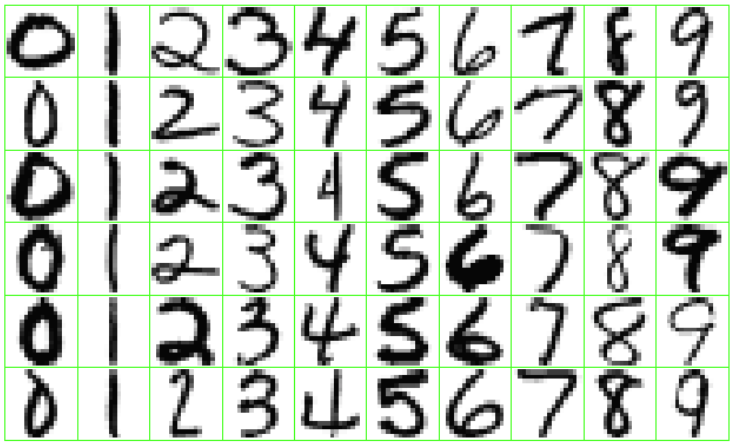
Springer Series in Statistics.
`http://statweb.stanford.edu/~tibs/ElemStatLearn/`

- Cours en ligne
  - R for Data Science, Garrett Grolemund and Hadley Wickham : `http://r4ds.had.co.nz/`
  - WikiStat, Philippe Besse : `http://wikistat.fr/`
  - Hugo Larochelle : `https://www.youtube.com/user/hugolarochelle`

- Logiciels
  - R + RStudio + Rmarkdown : `http://www.rstudio.com/`
  - Python + scikit-learn : `http://scikit-learn.org/`
- Technologies big data
  - MapReduce avec RHadoop
  - Spark+MLlib : `http://spark.apache.org/mllib/`
- Jeux de données
  - UC Irvine Machine Learning Repository : `http://archive.ics.uci.edu/ml/`
- Challenges industriels
  - Kaggle : `https://www.kaggle.com/`
  - Datascience.net : `https://datascience.net/`

Exemples

Reconnaissance de caractères manuscripts



- Objectif : Prédire la classe de chaque image (0,. . . ,9) à partir d'une matrice de 16×16 pixel, chaque pixel ayant une intensité de 0 à 255
- Taux d'erreur doit être faible afin d'éviter les mauvaises attributions du courrier

[Classification supervisée]

Exemples

Reconnaissance automatique de spams

Spam

```
WINNING NOTIFICATION
We are pleased to inform you of the result of the Lottery Winners
International programs held on the 30th january 2005. [...] You
have been approved for a lump sum pay out of 175,000.00 euros.
CONGRATULATIONS!!!
```

No Spam

```
Dear George,
Could you please send me the report #1248 on the project
advancement?
Thanks in advance.

Regards,
Cathia
```

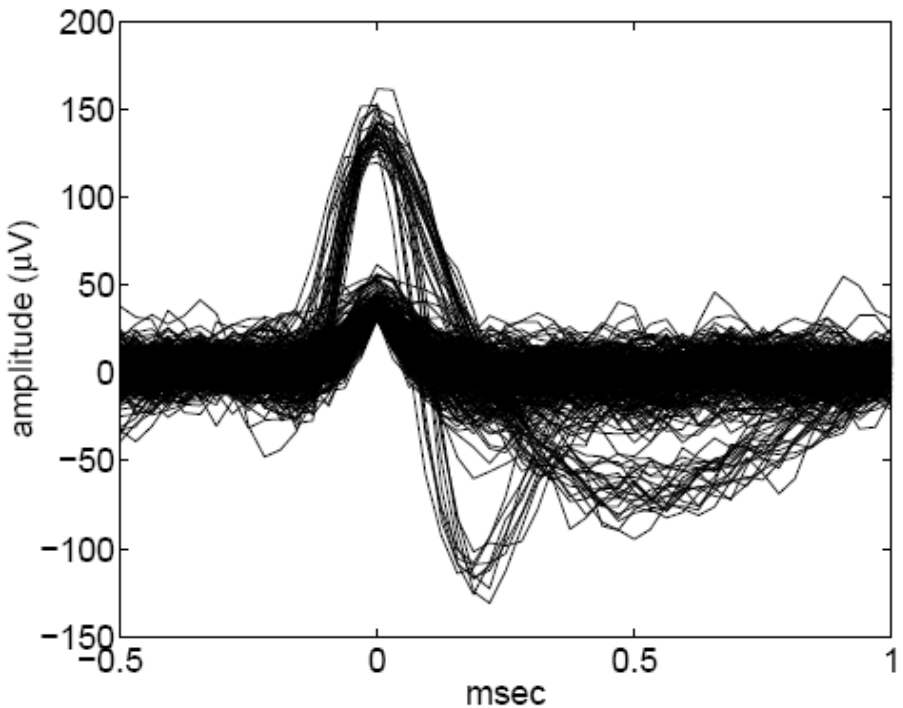
- Base de données de courriel identifiés ou non comme spam
- Objectif : Définir un modèle de prédiction permettant, à l'arrivée d'un courriel de prédire si celui-ci est un spam ou non
- Taux d'erreur doit être faible afin d'éviter de supprimer des messages importants ou d'inonder la boite au lettre de courriers inutiles

[Classification supervisée]

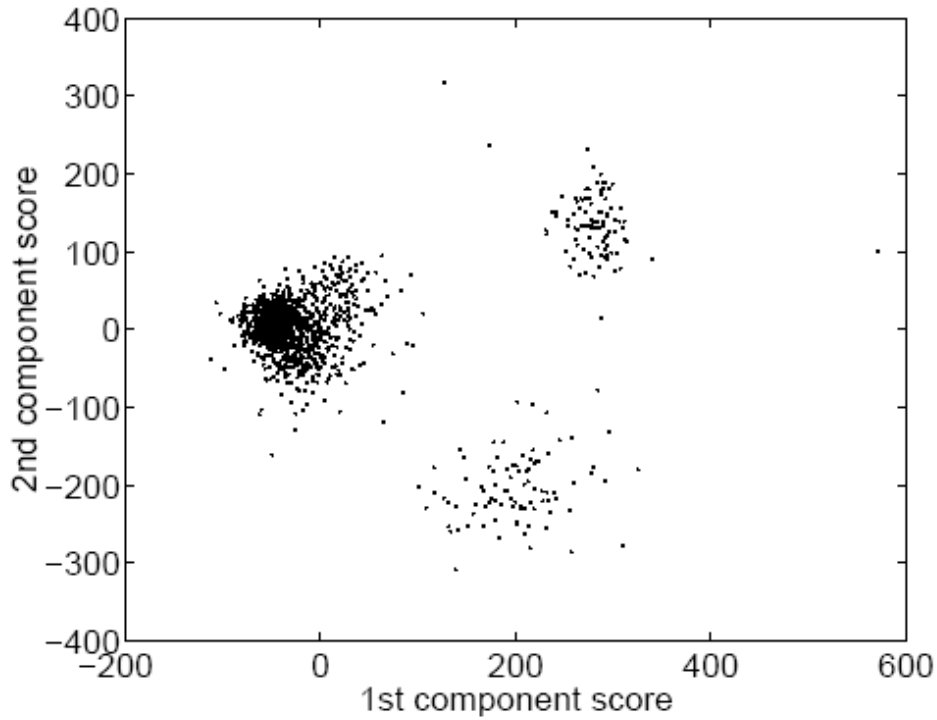
Exemples

Clustering de signaux neuronaux

- Signaux enregistrés par une micro-électrode
- Objectif : Trouver des classes de signaux qui se ressemblent.



- Représentation des signaux sur les deux premières composantes principales de l'ACP.
- Trois classes (clusters) semblent pouvoir être trouvées.



[Classification non supervisée]

- Variable de sortie Y qualitative à K modalités.
- Variables d'entrée réelles $X = (X^1, \ldots, X^p) \in \mathbb{R}^p$.
- Règle de classification $g : \mathbb{R}^p \to \{1, \ldots, K\}$:
  - dans le cas d'une réponse binaire $Y \in \{0, 1\}$

$$g(X) = 1 \text{ si } \sum_{X_i \in N_k(X)} Y_i > \sum_{X_i \in N_k(X)} (1 - Y_i)$$
$$= 0 \text{ sinon}$$

    où $N_k(X)$ est un voisinage de X contenant uniquement les k plus proches voisins de l'ensemble d'apprentissage
  - Dans le cas général, vote à la majorité parmi les k plus proches voisins du point X à classer.

- Pour appliquer cette méthode il faut fixer un paramètre : le nombre de voisins k
- Une stratégie de choix du paramètre k.
  - Choisir un échantillon d'apprentissage et un échantillon test
  - Faire varier k
  - Pour chaque valeur de k calculer le taux d'erreurs de prédiction des individus de l'échantillon test.
  - Retenir k qui minimise cette erreur test.

On considère :

- Une variable de sortie Y qualitative à valeur dans $\mathcal{Y} = \{1, \ldots, K\}$ (ensemble de classes).
- Une variable d'entrée $X = (X^1, \ldots, X^p)$ à valeurs dans $\mathcal{X}$, par exemple $\mathcal{X} = \mathbb{R}^p$.

La règle de classification de Bayes est une fonction $g : \mathcal{X} \to \mathcal{Y}$ optimale au sens de la minimisation d'un risque (un coût de mauvaise classification).

En pratique :

- La fonction de décision g sera estimée sur des données d'apprentissage i.e. sur un échantillon de couples de variables aléatoires i.i.d. $(X_1, Y_1), \ldots, (X_n, Y_n)$ de même loi que (X, Y).
- Le risque théorique de cette règle de classification sera estimé par un risque empirique, calculé sur des données test, i.e. un autre échantillon de couples de variables aléatoires i.i.d. de même loi que (X, Y).

Pour définir le risque théorique d'une régle de classification g, on doit définir le coût d'une prédiction.

- Soit $\hat{Y} = g(X)$ la classe prédite pour la variable d'entrée X et Y la vraie classe de cette variable d'entrée.
- Le coût de prédire la classe $\hat{Y}$ alors que la vraie classe est Y est donné par $L(Y, \hat{Y})$, où L est la fonction de coût (L pour LOSS).
- La fonction de coût peut être résumée par une matrice de coûts C de taille $K \times K$ telle que $C_{k\ell} = L(k, \ell)$.

vraie classe \ classe prédite	$\ell = 1$	$\ell = 2$	...	$\ell = K$
$k = 1$	$C_{11} \leq 0$	$C_{12} \geq 0$	...	$C_{1K} \geq 0$
$k = 2$	$C_{21} \geq 0$	$C_{22} \leq 0$	...	$C_{2K} \geq 0$
...	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$k = K$	$C_{K1} \geq 0$	$C_{K2} \geq 0$	...	$C_{KK} \leq 0$

Le risque théorique d'une règle de classification g est alors :

$$E(g) = \mathbb{E}_{X,Y}[L(Y, g(X))].$$

Lorsque la fonction de coût 0-1 est utilisée, on parle de taux d'erreur théorique :

$$E(g) = \mathbb{P}(g(X) \neq Y).$$

Le risque empirique (moyen) associé à l'échantillon $(X_1, Y_1), \ldots, (X_n, Y_n)$ est :

$$\widehat{E}(g) = \frac{1}{n} \sum_{i=1}^{n} L(g(X_i), Y_i)$$

Lorsque la fonction de coût 0-1 est utilisée, on parle de taux d'erreur empirique :

$$\widehat{E}(g) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(g(X_i) \neq Y_i).$$

La règle de classification de Bayes est la fonction g qui minimise le risque théorique $E(g)$.

On peut montrer que :

$$g(x) = \underset{\ell \in \{1,\dots,K\}}{\arg\min} \; \mathbb{E}\left[L(Y,\ell)|X = x\right],$$

où $\mathbb{E}\left[L(Y,\ell)|X = x\right]$ est le risque conditionnel associé à la prédiction ℓ pour la variable d'entrée $X = x$.

Le risque conditionnel s'interpète comme le coût moyen de mauvaise classification dans la classe ℓ sachant x.

Ainsi, la règle de classification de Bayes s'écrit aussi :

$$\begin{aligned} g(x) &= \underset{\ell \in \{1,\dots,K\}}{\arg\min} \sum_{k=1}^{K} L(k,\ell)\mathbb{P}(Y=k|X=x) \\ &= \underset{\ell \in \{1,\dots,K\}}{\arg\min} \sum_{k=1}^{K} C_{k\ell}\, \mathbb{P}(Y=k|X=x) \end{aligned}$$

Lorsque la fonction de coût 0-1 est utilisée cette règle s'écrit

$$g(x) = \underset{\ell \in \{1,\dots,K\}}{\arg\max}\, \mathbb{P}(Y=\ell|X=x).$$

Pour pouvoir appliquer cette règle, il faut disposer des probabilité à posteriori $\mathbb{P}(Y=k|X=x)$.

- $K = 2$ classes et la matrice de coûts est $C = \begin{pmatrix} 0 & C_{12} \\ C_{21} & 0 \end{pmatrix}$
- Les risques conditionnels associés aux deux prédictions possibles pour une valeur d'entrée x sont :

$$R(1|x) = \mathbb{E}\left[L(Y,1)|X=x\right] = \sum_{k=1}^{2} C_{k1}\mathbb{P}(Y=k|X=x)$$
$$= C_{21}\mathbb{P}(Y=2|X=x)$$
$$R(2|x) = \mathbb{E}\left[L(Y,2)|X=x\right] = \sum_{k=1}^{2} C_{k2}\mathrm{Pr}(Y=k|X=x)$$
$$= C_{12}\mathrm{Pr}(Y=1|X=x)$$

- Quelle décision prendre pour x ?

La règle de classification de Bayes consiste à affecter à la classe la moins risquée à posteriori :

- Affecter x à la classe 1 si $R(1|x) < R(2|x)$,
- Sinon affecter x à la classe 2.

c'est à dire :

- Affecter x à la classe 1 si

$$C_{21}\mathbb{P}(Y = 2|X = x) \leq C_{12}\mathbb{P}(Y = 1|X = x).$$

- Sinon affecter x à la classe 2.

Dans le cas d'une fonction de coût 0-1 on retrouve la règle de classification qui consiste à affecter x à la classe la plus probable à posteriori.

- Affecter x à la classe 1 si

$$\mathbb{P}(Y = 1|X = x) \geq \mathbb{P}(Y = 2|X = x).$$

- Sinon affecter x à la classe 2.

- Soit un échantillon $(Y_i, \hat{Y}_i)$, $i = 1, ..., n$ de n couples de vraies classes et de classes prédites avec la fonction de classification g.
- La matrice de confusion comptabilise les occurences des prédictions en fonction des vraies valeurs

vraie classe \ classe prédite	$\ell = 1$	$\ell = 2$	...	$\ell = K$
$k = 1$	n_{11}	n_{12}	...	n_{1K}
$k = 2$	n_{21}	n_{22}	...	n_{2K}
...	⋮	⋮	⋱	⋮
$k = K$	n_{K1}	n_{K2}	...	n_{KK}

où $n_{k\ell}$ est le nombre d'observations de classe k auxquels on a prédit la classe ℓ :

$$n_{k\ell} = \sum_{i=1}^{n} \mathbb{1}(Y_i = k \text{ et } \hat{Y}_i = \ell)$$

Le risque empirique (coût moyen de mauvaise classification) de la règle de classification g est :

$$\begin{aligned}
\widehat{E}(g) &= \frac{1}{n}\sum_{i=1}^{n} L(Y_i, \hat{Y}_i) \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{\ell=1}^{K} C_{kl}\mathbb{1}(Y_i = k \text{ et } \hat{Y}_i = \ell) \\
&= \frac{1}{n}\sum_{k=1}^{K}\sum_{\ell=1}^{K} C_{k\ell} n_{k\ell}
\end{aligned}$$

Dans le cas d'une fonction de coût 0-1, on retrouve le taux d'erreur empirique.

$$\widehat{E}(g) = \frac{1}{n}\sum_{k=1}^{K}\sum_{\ell=1,\ell\neq k}^{K} n_{k\ell}$$

On prend le cas particulier de la classification binaire.

- La variable à expliquer à deux modalités $\{1, 2\}$.
- Vocabulaire : "prédire 1"=positif et "prédire 2"=négatif (par exemple).
- La matrice de confusion :

		prédictions	
		1 (positif)	2 (négatif)
vraies	1	VP	FN
valeurs	2	FP	VN

avec

- VP = vrais positifs,
- VN = vrai négatifs,
- FP = faux positifs,
- FN = faux négatifs.

Cas d'une fonction de coût 0-1 avec $C = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$

- La matrice de coût ne distingue pas les deux types d'erreurs FP et FN
- L'expression du risque empirique (coût moyen) devient

$$\begin{aligned} \hat{E}(g) &= \frac{1}{n} \sum_{k=1}^{2} \sum_{\ell=1}^{2} C_{k\ell} n_{k\ell} \\ &= \frac{C_{11} n_{11} + C_{21} n_{21} + C_{12} n_{12} + C_{22} n_{22}}{n} \\ &= \frac{n_{21} + n_{12}}{n} \\ &= \frac{FP + FN}{n} \end{aligned}$$

ce qui correspond au taux d'erreur standard.

Cas d'une fonction de coût pondérée avec $C = \begin{pmatrix} 0 & 1 \\ 5 & 0 \end{pmatrix}$

- On associe un coût cinq fois plus important aux faux positifs
- L'expression du risque empirique (coût moyen) devient

$$\begin{aligned} \hat{E}_{(g)} &= \frac{C_{11}n_{11} + C_{21}n_{21} + C_{12}n_{12} + C_{22}n_{22}}{n} \\ &= \frac{5 \times n_{21} + n_{12}}{n} \\ &= \frac{5 \times FP + FN}{n} \end{aligned}$$

ce qui correspond à un taux d'erreur pondéré.

Evaluer la performance d'une règle de classification

Exemples de critères de performance :

- le risque empirique, le taux d'erreur empirique,
- la sensiblité, la spécificité, l'air sous la courbe roc (AUC), l'indice de Gini, la F-mesure (dans le cas binaire).

Ces critères doivent être toujours estimés sur des données test.

Deux approches :

- la méthode de l'échantillon test : partager aléatoirement les données en un échantillon d'apprentissage (80% par exemple des données) et un échantillon test. En effectuant plusieurs tirages aléatoires on peut obtenir plusieurs vecteurs de prédictions et donc plusieurs estimations du critère de performance choisi (par exemple le taux d'erreur).
- la validation croisée (utile pour les petits échantillons).

Deux approches de valisation croisée :

- la validation croisée leave one out :
  - Pour $i = 1, \ldots, n$,
    - estimer la règle sur les données privées de la $i$ème donnée,
    - prédire cette donnée i avec cette règle.
  - Calculer le critère de performance sur ces n prédictions.

  $\Rightarrow$ un seul vecteur de prédictions possible.

- la validation croisée K-folds :
  - Découper les données en K sous-échantillons de même taille (en respectant si possible les proportions des classes).
  - Pour $k = 1, \ldots, K$,
    - estimer la règle sur les données privées de l'échantillon k,
    - prédire les données de l'échantillon k avec cette règle.
  - Calculer le critère de performance sur ces n prédictions.

  $\Rightarrow$ plusieurs vecteurs de prédictions possibles.

Critères de performance dans le cas binaire

On se place dans le cadre de la classification binaire.

- Variable à expliquer à deux modalités $\{1, 2\}$.
- Vocabulaire : "prédire 1"=positif et "prédire 2"=négatif.
- La matrice de confusion :

		prédictions	
		1 (positif)	2 (négatif)
vraies	1	VP	FN
valeurs	2	FP	VN

avec VP = vrais positifs, VN = vrai négatifs, FP = faux positifs et FN = faux négatifs.

$\Rightarrow$ Pourcentages lignes et colonnes.

- Les pourcentages lignes donnent :
  - le taux de vrais positifs : TVP=VP/(VP+FN)
    - proportion d'entrées appartenant à la classe 1 qui ont été bien prédites.
    - aussi appellé le rappel (recall) ou encore la sensibilité (sensitivity).
  - le taux de vrais négatifs : TVN=VN/(FP+VN)
    - proportion d'entrées appartenant à la classe 2 qui ont été bien prédites.
    - aussi appellé la spécificité (spécificity).
- Les pourcentages colonnes donnent :
  - la précision : VP/(VP+FP)
    - proportion des prédictions 1 qui sont effectivement des entrées de la classe 1.
    - cette proportion se compare à la prévalence qui est la proportion de la classe 1 dans les données.

Un bon modèle est à la fois :

- sensible et spécifique. Cela se mesure avec :
  - la courbe ROC et l'auc.
  - la courbe lift et l'indice de Gini.
- sensible et précis. Cela se mesure avec la F-mesure

$$F = 2\,\frac{\textit{rappel} \times \textit{précision}}{\textit{rappel} + \textit{précision}}.$$

  - Moyenne harmonique du rappel (la sensibilité, le TVP) et de la précision.
  - Mesure la capacité de la règle de classification à bien prédire les entrées de la classe 1 et à ne pas prédire 1 des entrées de la classe 2.

Mesure de performance d'un score

- On se place dans d'une variable de sortie binaire à 2 modalités $\{1, 2\}$.
- La classe 1 sera la "classe d'intéret", l'événement dont la prédiction sera condidérée comme "positive".
- On associe souvent un score (une note) à une entrée x qui sera d'autant plus grand que la probabilité qu'elle appartienne à la classe 1 est grande. Ce score est donc souvent défini par :
  - la probabilité à posteriori d'appartenir à la classe 1 sachant x :

$$p = \mathbb{P}(Y = 1|X = x),$$

  - ou encore le logit de p qui transforme les probabilités sur $]0, 1[$ en évidence sur $\mathbb{R}$:

$$logit(p) = ln\frac{p}{1-p}.$$

- Pour définir une règle de classification à partir du score, il faut fixer un seuil. Par exemple :
  - affecter x à la classe 1 si $p \geq 0.5$,
  - ou encore affecter x à la classe 1 si $logit(p) \geq 0$.
- Si on modifie le seuil, on modifie la règle de classification, la matrice de confusion, et donc tous les indicateurs présentés précédemment (taux d'erreur, spécificité, sensibilité...).
- On mesure souvent visuellement et numériquement de l'efficacité d'un score indépendamment du choix du seuil :
  - à partir de la courbe ROC (Receiver Operating Characteristic) et de l'AUC (area under the curve),
  - à partir de la courbe LIFT et de l'indice de Gini.

Question : Comment évoluent le TVP et le TVN lorsque le seuil augmente ?

La courbe ROC et le AUC

Construction de cette courbe.

- Le TVP et le TFP dépendent du seuil s choisi :
  - $TVP(s)$ donne le taux de vrais positifs obtenu avec un seuil s,
  - $TFP(s)$ donne le taux de faux positifs obtenu avec le même seuil.
- La courbe ROC relie les points $(TFP(s), TVP(s))$ obtenus en faisant varier s.
- Pour construire cette courbe, on construit une grille de seuils et on calcule pour chaque seuil s le TVP et le TFP.

Interprétation de cette courbe.

- Si cette courbe coïncide avec la diagonale, c'est que le score n'est pas plus performant qu'un modèle aléatoire (où on attribue la classe au hasard).

- Plus la courbe ROC s'approche du coin supérieur gauche, meilleur est le modèle, car il permet de capturer le plus possible de vrais positifs avec le moins possible de faux positifs.
- En conséquence, l'aire sous la courbe ROC, appellée critère AUC, peut être vu comme une mesure de la qualité du score. Ce critère AUC varie entre 0 (cas le pire) et 1 (cas le meilleure...).

Choix d'un seuil.

- On utilise parfois la courbe ROC pour choisir un seuil. En pratique, on peut prendre le seuil correspondant au point de la courbe la plus éloigné de la première bissectrice et le plus prêt du point supérieur gauche $(0, 1)$. Ou encore le seuil correspondant au point où la pente de la courbe est la plus proche de 0.
- Mais on peut également choisir le seuil qui optimise un critère de performance comme le taux d'erreur, le risque empirique, la F-mesure... C'est alors une approche différente de la version empirique de la règle de classification de Bayes souvent utilisée par défaut.