Fouille de données

NGUYỄN Thị Minh Huyền ©2016

huyenntm@hus.edu.vn

Plan

Plan

Classification

Classification

- Classification binaire : positive ou négative.
- Classification probabiliste : $f(x) \in [0, 1]$, seuil t.
  $f(x) \geq t \Rightarrow x$ positive, sinon négative
  $\Rightarrow$ classification binaire en fonction de t.

Classification

Mesures de qualité

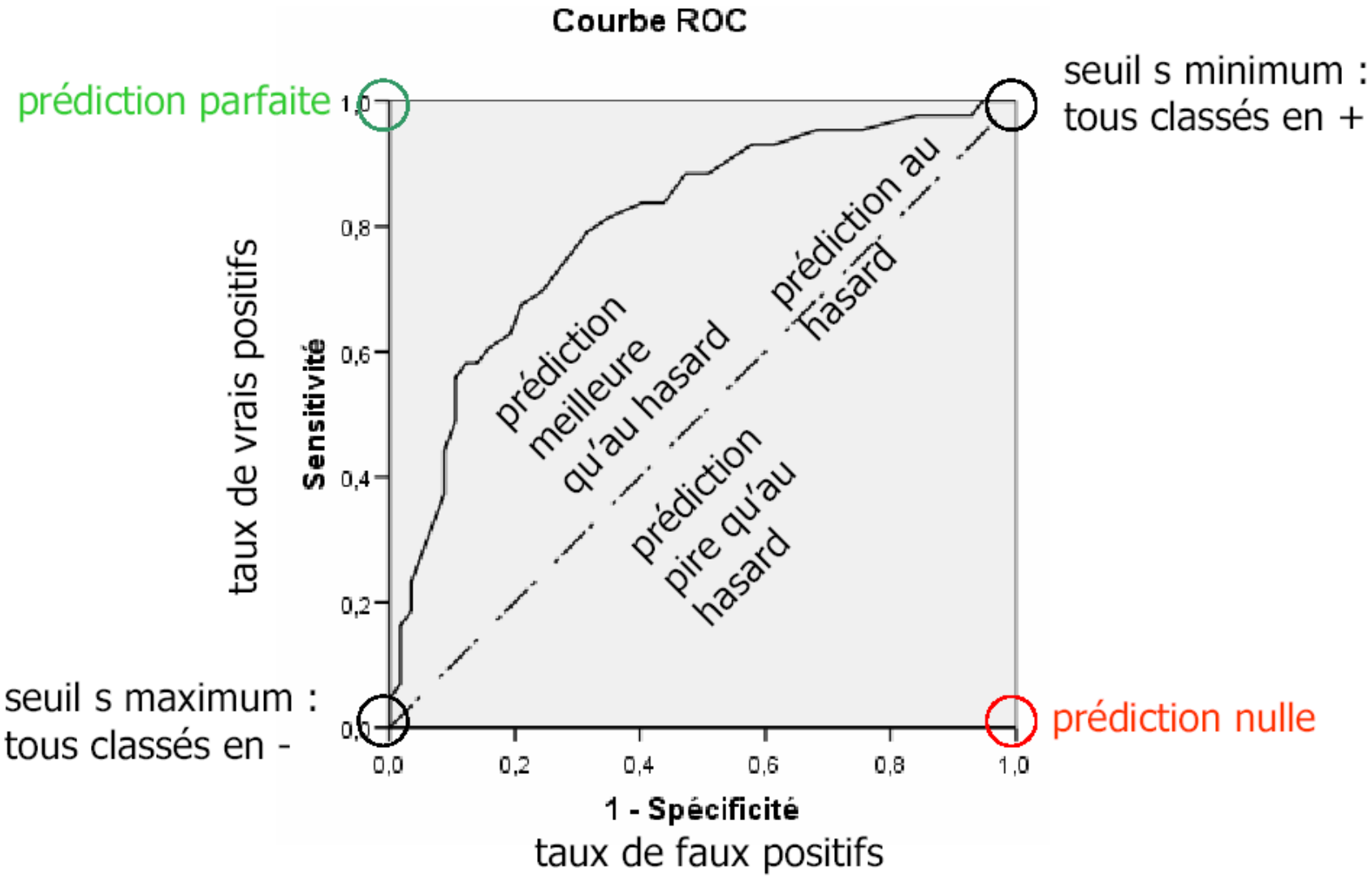
- Ensemble de test : P cas positifs, N cas négatifs.
- Valeurs $TP(t)$ (*true positive*), $FP(t)$, $TN(t)$, $TF(t)$
- $TPrate = TP/P$ (Recall), $FPrate = FP/N$,
  $YRate = (TP + FP)/(P + N)$
- Precision = $TP/(TP + FP)$, Accuracy = $(TP + TN)/(P + N)$
- $F - measure =$
  $Precision * Recall/(\alpha Precision + (1 - \alpha)Recall)$, $(\alpha \in [0, 1])$
  $\alpha = 0.5 \Rightarrow$
  $F1 = 2 * Precision * Recall/(Precision + Recall)$.

Plan

Courbe ROC (*receiver operating characteristics*)

- Pour chaque fonction f, la courbe ROC est définie par $x = FPrate(t)$, $y = TPrate(t)$ en variant le seuil t.
- Mesure AUC (*Area Under Curve*) : surface sous la courbe ROC = 1 pour un modèle idéal, = 0,5 pour un modèle aléatoire
  $\Rightarrow$ un bon modèle a une valeur AUC entre 0,5 et 1.

Courbe ROC


Courbe ROC
prédiction parfaite
seuil s minimum :
tous classés en +
taux de vrais positifs
Sensitivité
1,0
0,8
0,6
0,4
0,2
0,0
prédiction meilleure qu'au hasard
prédiction au hasard
prédiction pire qu'au hasard
seuil s maximum :
tous classés en -
prédiction nulle
0,0
0,2
0,4
0,6
0,8
1,0
1 - Spécificité
taux de faux positifs

Plan

Courbe de lift

- Pour chaque fonction f, la courbe de lift est définie par $x = Yrate(t)$, $y = TP(t)$ en variant le seuil t.
- Mesure AUC (*Area Under Curve*) : surface sous la courbe
  lift = P pour un modèle idéal, = $P/2$ pour un modèle aléatoire
  $\Rightarrow$ un bon modèle a une valeur AUC entre $P/2$ et P.

Courbe de lift : exemple