Datamining 3a: Classification supervisée Partie 1 Mesure de l'erreur Coube Roc et plus proches voisins

M2 STD

September 29, 2016

But

On a des individus i caractérisés par p variables quantitatives représentées par le vecteur X_i et une variables qualitative Y_i On cherche a prédire Y en fonction des X (Dans cet exposé on prendra toujours un Y "binaire" ($Y \in \{-1, 1\}$)) On cherche aussi a estimer (dans la mesure du possible) $P(Y = +1|X)$ (on va voir l'utilité de cela tout de suite)

Choix d'un modèle le principe

Supposons qu'on ait construit un classifieur (sur une base d'apprentissage)
On voudrait en estimer ses performances (sur une base test de taille n)
La première idée est d'introduire la matrice de confusion (tableau de contingence):

	$\hat{Y} = 1$	$\hat{Y} = -1$
$Y = 1$	a	b
$Y = -1$	c	d

Vocabulaire : vrai positifs= a; faux positifs= c; taux d'erreur= $(b + c)/n$;

Choix d'un modèle le principe

Attention le taux d'erreur est un bon critère... mais pas tous le temps !

- Sa lecture seule n'est pas suffisante dans le cas de classes de tailles disproportionnées
- En fonction du probleme pratique minimiser $\alpha c + \beta d$ (i.e. affecter des poids différents aux deux erreurs) peut etre pertinent (ex mailing)

Choix d'un modèle le principe

Classes disproportionnées : exercice

Supposons que la classe $+1$ représente p pourcents de la population.

1. quel est le taux d'erreur du classifieur deterministe qui affecte la classe 1 a tout le monde ?
2. quelle est l'espérance du taux d'erreur du classifieur aléaroire qui répartit p pourcent des individus dans la classe $+1$ (et $100 - p$ dans la classe -1)

Etudier les résultats lorsque p est très proche de 1

Choix d'un modèle la courbe ROC

Le principe de la courbe roc est de donner un indicateur qui permet de prendre en compte les deux problèmes précédents

Principe de la courbe ROC

Si on a une estimation $\hat{P}(Y = 1|X)$ (ou quelque chose de raisonnablement lié) pour chaque valeur de X:

- on affecte une regle de décision "raisonnable" du type $\hat{Y} = 1$ si $\hat{P}(Y = 1|X) \geq t$ ce qui donne lieu au calcul d'une matrice de confusion
- on reporte sur un graphique le point $(a/(a + b), c/(c + d))$

Exercices

1. Montrer que la courbe ROC passe toujours par les points $(0, 0)$ et $(1, 1)$.
2. Que vaut la courbe ROC pour le classifieur parfait (i.e.) qui donne la probabilité 1 d'appartenir a la classe 1 lorsque c'est vrai et 0 sinon ? Tracer la courbe roc associée.
3. Que vaut la courbe ROC (moyenne) pour un classifieur aléatoire (qui tire aléatoirement les probabilités dans $[0, 1]$)? Tracer la courbe ROC associée.
4. Pour une abscice t donnée est ce mieux d'avoir une grande ou une petite valeur pour la courbe ROC, quelle est la meilleur ordonnée associée possible $y(t)$? tracer la courbe $(t, y(t))$

Aire Sous la courbe interprétations

L'aire sous la courbe ROC, notée AUC représente la probabilité que $\hat{P}(Y = 1|X)$ soit supérieure pour un individu $+1$ que pour un individu -1.

comparaison de deux indicateurs

La courbe roc présente le Rapport "Faux positifs" sur nombre de négatifs en fonction du taux de vrai positifs (vrais positifs sur positifs) fixés. Si pour la courbe roc d'un classifieur est tout le temps au dessus de celle d'un autre ce classifieur est meilleur quelque soit la regle de décision "raisonnable" choisie.

comparaison de deux indicateurs

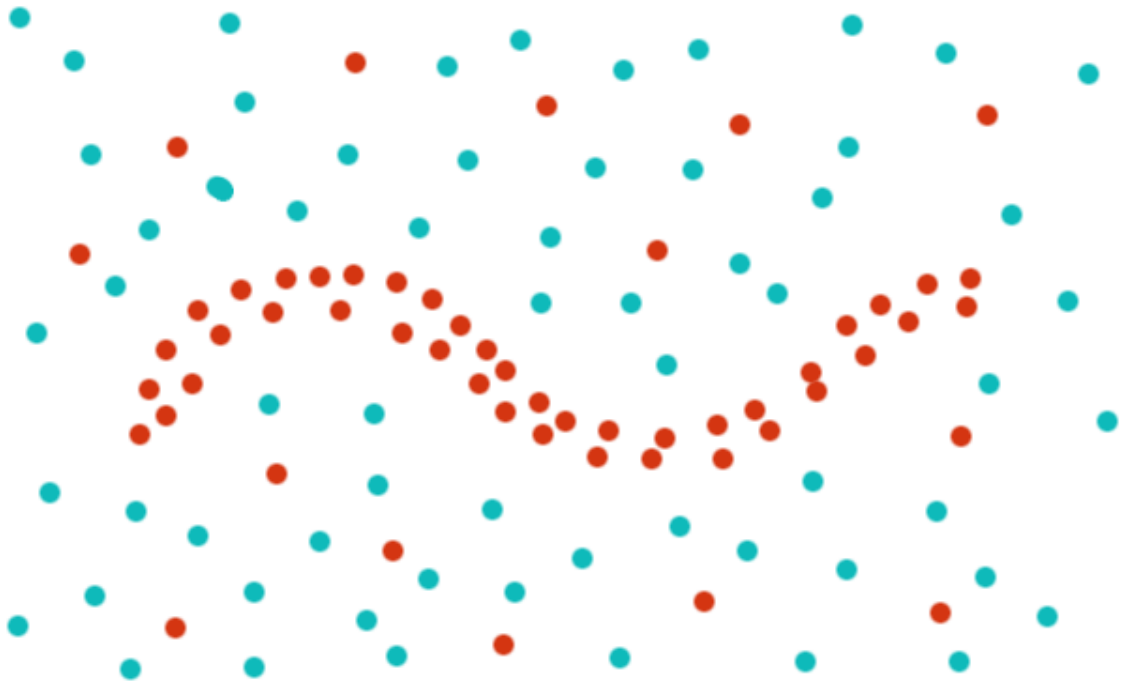
Si deux courbes ROC se croisent... le choix du modèle se fera en fonction des priorités et fonctions couts

Conclusion

La courbe roc apporte plus d'informations sur la qualité d'un classifieur que le seul taux d'erreur mais :

- Il faut avoir un classifieur permettant d'obtenir un indicateur de $P(Y = 1|X)$
- ATTENTION : pour bien faire on doit construire le modèle sur une base d'apprentissage et évaluer la courbe ROC sur une base test

Illustration générale de la classification

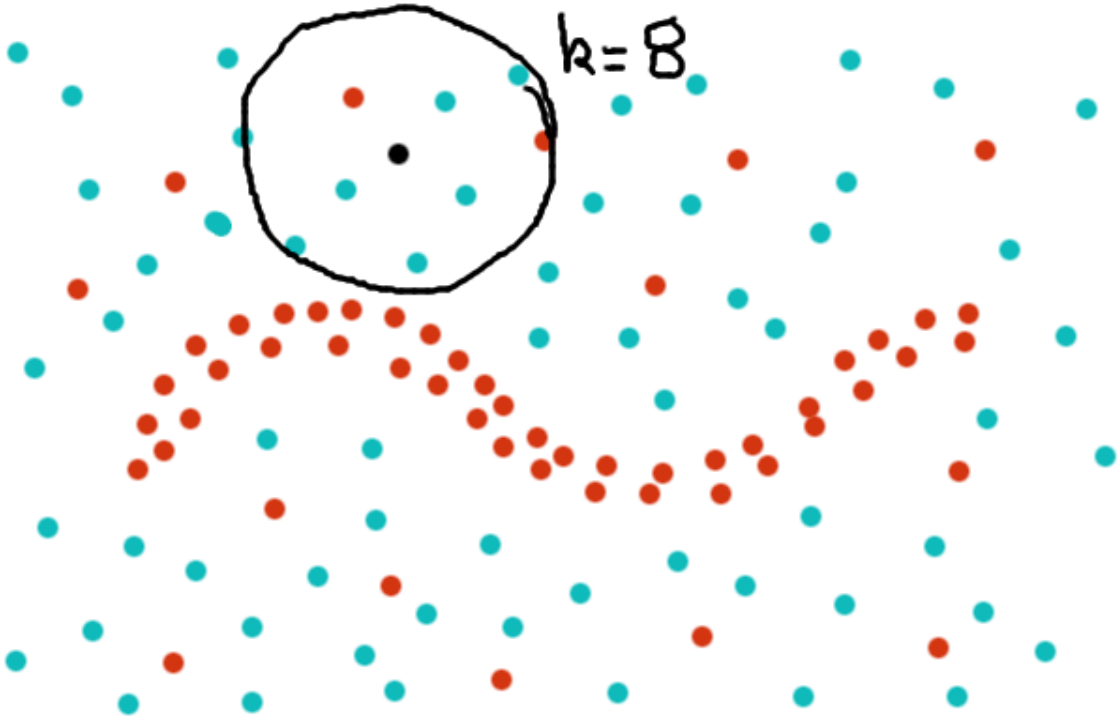


L'estimateur des plus proches voisins

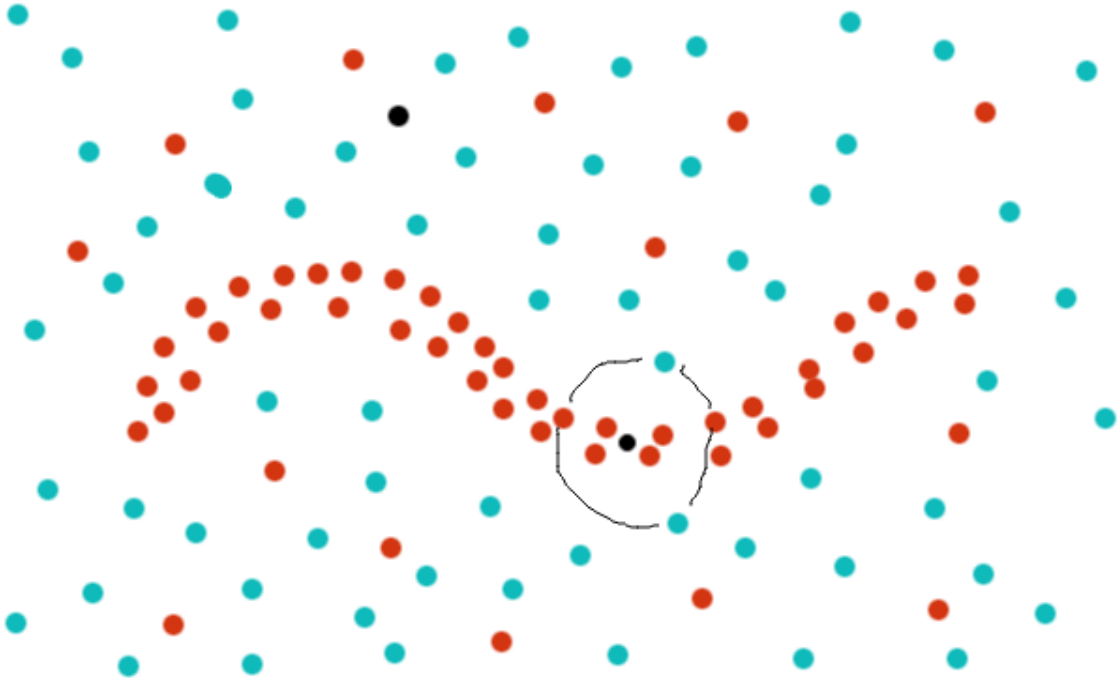
Pour n taille de la base donnée on se fixe un nombre de voisin k_n et pour un nouveau point X_0 :

- On cherche r_{k_n} la distance a son $k$eme voisin dans la base
- On "compte" n_+ le nombre d'observations dans $\mathcal{B}(X_0, r_{k_n})$ de label $+1$
- Si $n_+ >= k_n/2$ on estime $\hat{Y}_0 = +1$ sinon -1.
- On peut estimer $\hat{P}(Y_0 = 1|X_0)$ par n_+/k_n

L'estimateur des plus proches voisins



L'estimateur des plus proches voisins



La théorie

Pour n taille de la base donnée on se fixe un nombre de voisin k_n et pour un nouveau point X_0 :

- On cherche r_{k_n} la distance a son $k$eme voisin dans la base
- On "compte" n_+ le nombre d'observations dans $\mathcal{B}(X_0, r_{k_n})$ de label $+1$
- Si $n_+ >= k_n/2$ on estime $\hat{Y}_0 = +1$ sinon -1 (quelle condition a votre avis ?).
- On peut estimer $\hat{P}(Y_0 = 1|X_0)$ par n_+/k_n (quelle condition a votre avis ?).

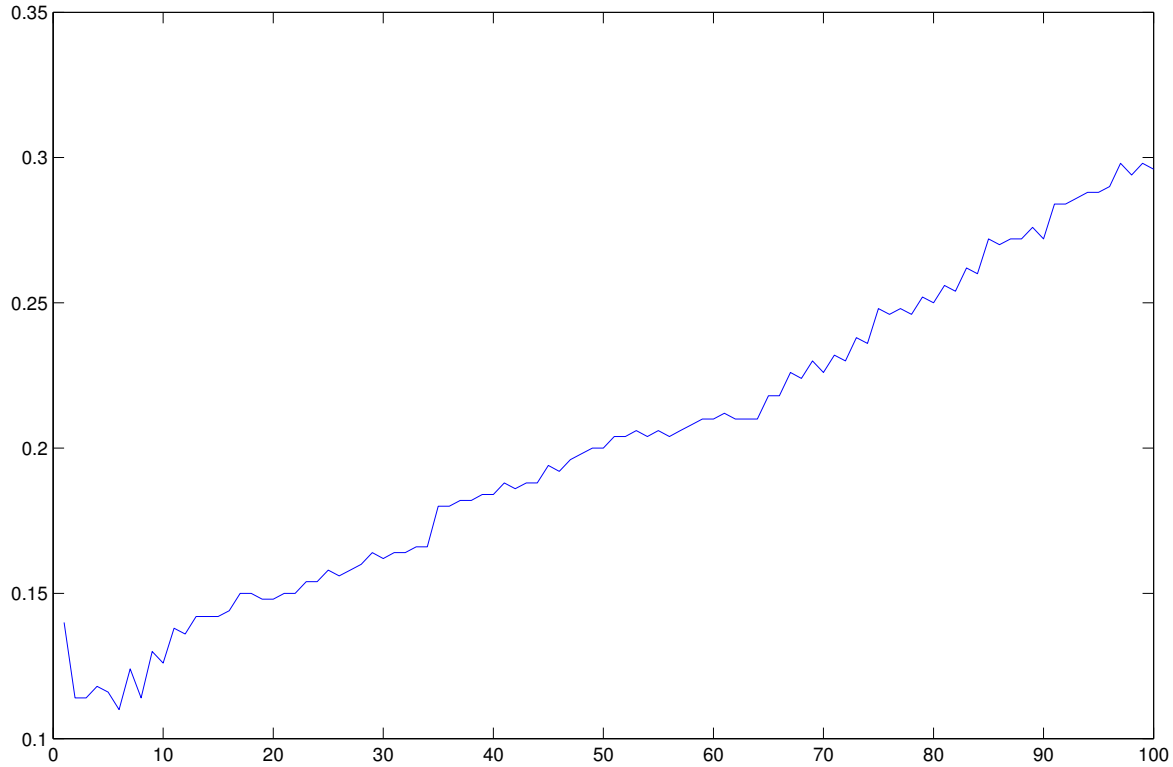
La théorie

Si $k_n \to \infty$ et $k_n/n \to 0$ alors "la méthode converge"... ca c'est la théorie... c'est bien... mais comment choisir pratiquement k_n ?

La pratique

Le k_n optimal dépend a priori de la taille de l'échantillon... donc on va préferer le leave one out !

Exemple



Exemple

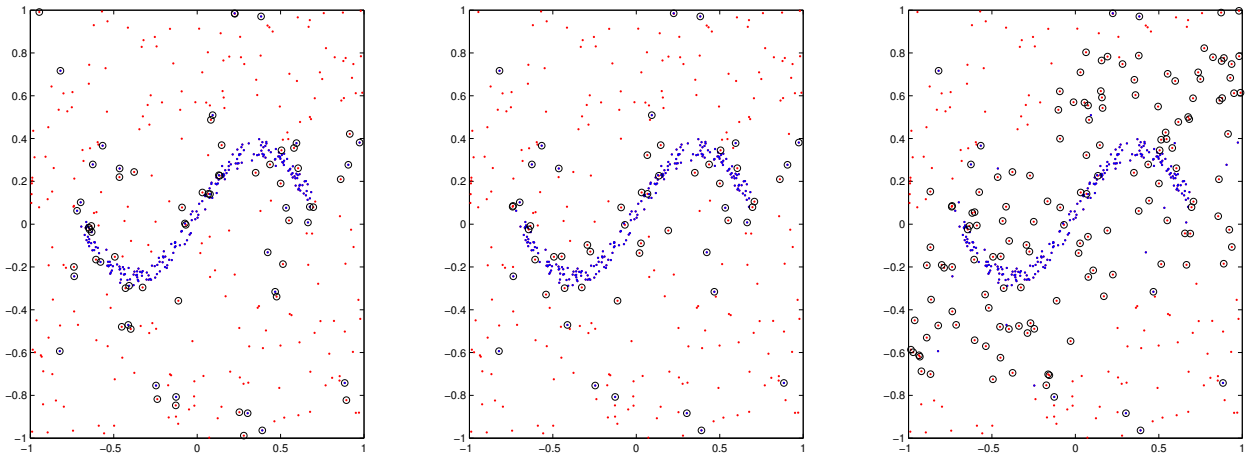


Figure: $k = 1$, $kopt = 6$ et $k = 100$ (en noir les mal classés)

Les limites

discussion : pouvez vous imaginer les Limites d'une telle méthode ?