ESLII

Chapitre 2 apprentissage supervisé

La régression est lorsque nous cherchons a prédire une variable quantitative.
La classification est lorsque nous cherchons à prédire une variable qualitative.

Les variables qualitatives dans le cas où nous avons que deux catégories peuvent être représentées par 0 et 1 ou -1 et 1 (‘vivant’ ou ‘mort’ , ‘succès’ ou ‘échec’) .
Quand il y a plus de deux catégories, on utilise les ‘dummy varaibles’. On les représentes par un vecteur de K variables binaires.

Les variables quantitatives sont représentées par G (pour Groupe)
Les variables qualitatives sont représentées par Y

A-/ Deux approches simples de prédiction : Les moindres carrés et les plus proches voisins

On va étudier deux méthodes de prédictions simples (mais puissantes) : le modèle linéaire ajusté par les moindres carrés et la règle de prédiction des K plus proches voisins.

1) le modèle linéaire et les moindres carrés

On a des entrées sous forme de vecteur $X^T = (X_1, X_2, ...X_p)$
On prédit la sortie Y via le modèle :

$$\hat{Y} = \hat{\beta}_0 = \sum_{j=1}^{p} X_j \hat{\beta}_j$$

Il est souvent utile d’inclure la constante 1 à X pour écrire :

$$\hat{Y} = X^T \hat{\beta}$$

Il y a beaucoup de méthodes différentes pour faire un ajustement linéaire du modèle à un ensemble de données d’apprentissage. Mais la méthode la plus courante est la méthode des moindres carrés
Nous choisissons les coefficients β pour minimiser

$$RSS(\beta) = \sum_{i=1}^{N} (y_i - x_i^T \beta)^2$$

son minimum existe toujours mais peut ne pas être unique.
On peut réécrire

$$RSS(\beta) = (y - X\beta)^T (y - X\beta)$$

La dérivée est

$$X^T (y - X\beta) = 0$$

L’unique solution est

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Un Exemple du modèle linéaire dans un cas de classification :

FIGURE 2.1. *A classification example in two dimensions. The classes are coded as a binary variable (BLUE = 0, ORANGE = 1), and then fit by linear regression. The line is the decision boundary defined by* $x^T\hat{\beta} = 0.5$. *The orange shaded region denotes that part of input space classified as ORANGE, while the blue region is classified as BLUE.*

Les points de R^2 classifiés comme Orange correspondent à $\{x:x^T\beta>0.5\}$.
The decision boundary (la limite de décision) est $\{x:x^T\beta=0.5\}$

Une limite de décision est peu probable d'être optimal est en fait ne l'est pas.
La limite de décision optimale est non-linéaire.

2) La méthode des plus proches voisins

Le k plus proches voisins est définit par

$\hat{Y}(x)=\frac{1}{k}\sum_{x_i\in N_k(x)} y_i$ où $N_k(x)$ sont les k voisins plus proches de x dans les données d'entraînement
On considère que la loi métrique utilisée est la distance Euclidienne.
C'est donc la moyenne des valeurs des k plus proches points des données d'entraînement.

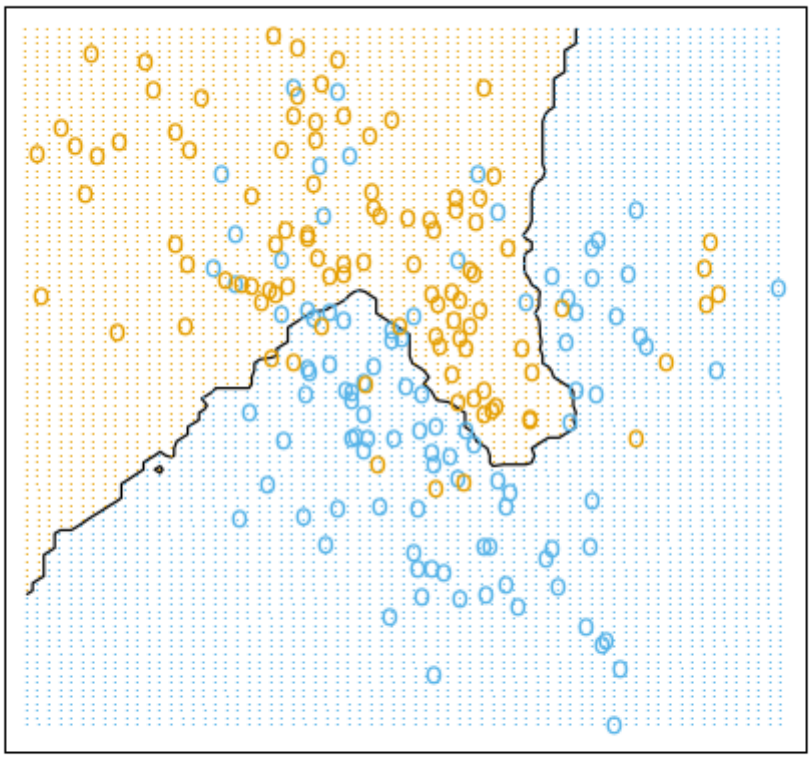

FIGURE 2.2. *The same classification example in two dimensions as in Figure 2.1. The classes are coded as a binary variable* (BLUE $= 0$, ORANGE $= 1$) *and then fit by* 15*-nearest-neighbor averaging as in (2.8). The predicted class is hence chosen by majority vote amongst the* 15*-nearest neighbors.*

Dans cette figure on a utilisé les mêmes données d'entraînement que pour la classification linéaire du 1). On a utilisé k =15 : les 15-plus-proches-voisins.

3) Des moindres carrés au plus proches voisins

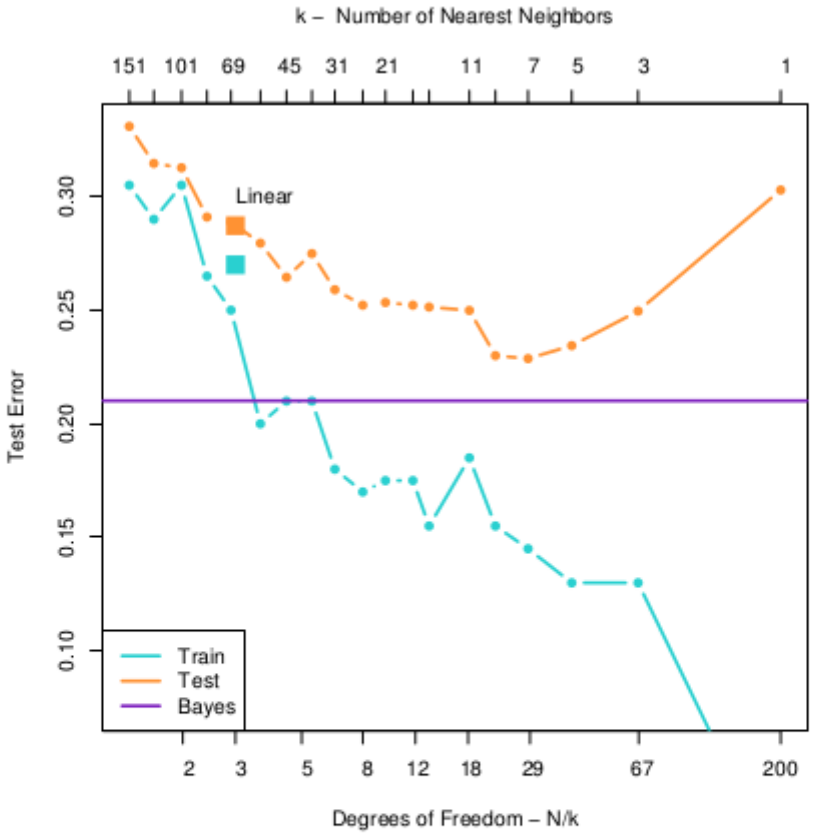

FIGURE 2.4. *Misclassification curves for the simulation example used in Figures 2.1, 2.2 and 2.3. A single training sample of size* 200 *was used, and a test sample of size* $10,000$. *The orange curves are test and the blue are training error for k-nearest-neighbor classification. The results for linear regression are the bigger orange and blue squares at three degrees of freedom. The purple line is the optimal Bayes error rate.*

B-/ Decision théorique statistique

On considère dans ce cas une sortie quantitative
On pose X € R^P un vecteur d'entrée et Y € R le scalaire de sortie.
On pose f(X) une fonction pour prédire Y
Cette théorie requiert une fonction de perte L(Y,f(X)). On utilisera Squared error loss
L(Y, f(X)) = (Y – f(X))^2
Cela nous amène a un critère pour choisir f : (EPE Expected (squared) Prediction Error)

$$EPE(f)=E(Y-f(X))^2=\int [y-f(x)]^2 Pr(dx,dy)$$

Qui nous mène à

$$EPE(f)=E_X E_{Y|X}([Y-f(X)]^2|X) \quad \text{car P(X,Y) = P(Y | X) P(X)}$$

pour que EPE(f) soit minimale

$$f(x)=argmin_c E_X E_{Y|X}([Y-c]^2|X)=argmin_c E_{Y|X}([Y-c]^2|X)$$

Comme on est à un minimum locale la dérivée est nulle

$$0=argmin_c E_{Y|X}(-2[Y-c]|X)$$

donc

$$f(x)=E_{Y|X}(Y|X=x)$$

On l'appelle espérance conditionnelle ou encore fonction de régression.
La meilleur prédiction de Y en tout point X=x est la moyenne conditionnelle lorsque nous utilisons l'erreur quadratique moyenne.

Pour les K plus proches voisins nous avons

$\hat{f}(x)=Average(y_i|x_i\in N_k(x))$ où $N_k(x)$ est l'ensemble des k plus proches voisins de x

Pour un ensemble de données d'entraînement grand N , les points voisins sont probables d'être proche de x. Et plus k devient grand, plus la moyenne devient stable.

Pour la fonction de régression linéaire on peut considérer que
$f(x)\approx x^T\beta$

on remplace cette dernière expression dans EPE(β)

$$EPE(\beta)=\int (y-x^T\beta)^2 Pr(dx,dy)$$

$$\frac{\partial EPE(\beta)}{\partial \beta}=0=\int 2(y-x^T\beta)(-1)x\,Pr(dx,dy)=-2\int (y-x^T\beta)x\,Pr(dx,dy)$$

Donc

$$E[yx]-E[xx^T\beta]=0 \quad \text{car } x^T\beta \text{ est un scalaire}$$

donc $\beta=E[xx^T]^{-1}E[yx]$

pour les deux (k plus proches voisins et les moindres carrés), les approximations conditionnelles sont des moyennes. Mais elle diffèrent :
- les moindres carrés supposent que f(x) est bien approximé par un fonction globale linéaire
- les k-plus-proches-voisins suppose que f(x) est bien approximé par un fonction constante locale

Si on remplace L_2 par L_1:E|Y – f(X)| , la solution dans ce cas est la médiane conditionnelle

$$\hat{f}(x)=median(Y|X=x)$$

Dans le cas d'une sortie G qualitative l' EPE est

$$EPE=E[L(G,\hat{G}(X))]$$

$$EPE=E_X\sum_{k=1}^{K}L[G_k,\hat{G}(X)]Pr(G_k|X)$$

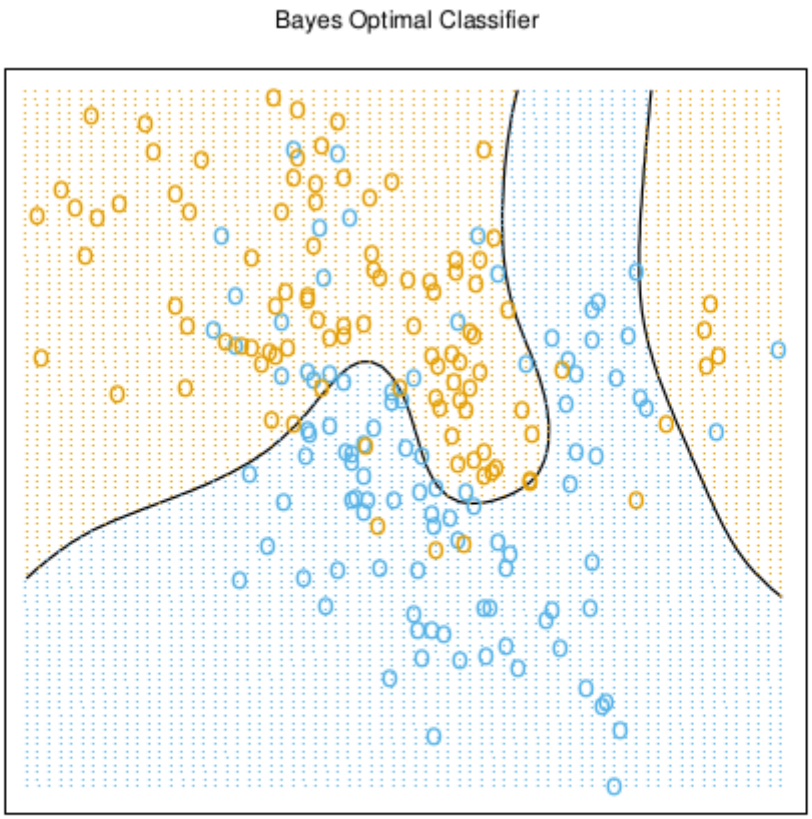

FIGURE 2.5. *The optimal Bayes decision boundary for the simulation example of Figures 2.1, 2.2 and 2.3. Since the generating density is known for each class, this boundary can be calculated exactly (Exercise 2.2).*

$$\hat{G}(x) = argmin_{g \in G} \sum_{k=1}^{K} L[G_k, g] Pr(G_k | X = x)$$

Avec la fonction de perte 0-1

$$\hat{G}(x) = argmin_{g \in G} [1 - Pr(G | X = x)]$$

$$\hat{G}(x) = G_k \, if \; Pr(G_k | X = x) = max_{g \in G} Pr(g | X = x)$$

Cette solution est connue sous le nom de Classifieur de Bayes.
Le taux d'erreur du classifieur de Bayes est appelé Bayes rate.

C-/ Méthode locale dans les hautes dimensions

Nous avons étudié deux techniques: le stable mais biaisé modèle linéaire et le modèle moins stable mais moins biaisé k plus proches voisins.

Nous rencontrons des problèmes pour les k plus proches voisins dans les hautes dimensions.
On suppose que en dimension p on a de manière répartie uniformément les valeurs des données.
Soit r la fraction des observations, la longueur de bord attendu moyenne entre chaque donné est de $e_p(r)=r^{1/p}$
Donc pour p=10 , pour 1% des données ont a une distance de $e_{10}(0.01)=63\%$ et pour 10% on a $e_{10}(0.1)=80\%$
Donc pour capturer 1 % ou 10 % des données pour former une moyenne locale, nous devons couvrir 63 % ou 80 % de la plage de chaque variable d'entrée. Les voisins ne sont donc plus 'locaux'. Réduire r n'aide pas non plus, car moins il y a d'observations moyennées plus la variance est grande.

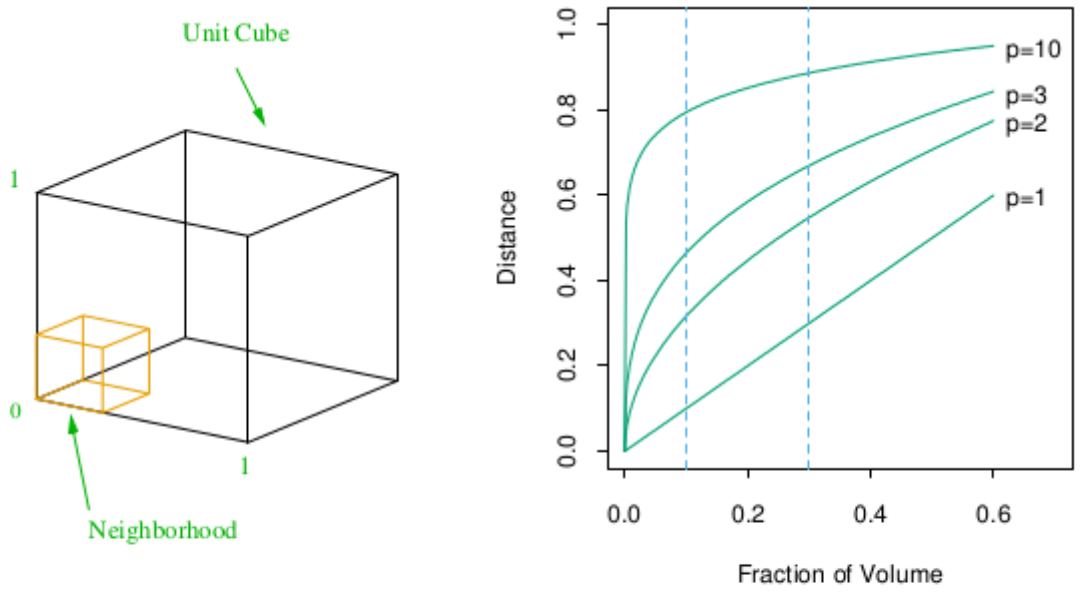

FIGURE 2.6. *The curse of dimensionality is well illustrated by a subcubical neighborhood for uniform data in a unit cube. The figure on the right shows the side-length of the subcube needed to capture a fraction r of the volume of the data, for different dimensions p. In ten dimensions we need to cover 80% of the range of each coordinate to capture 10% of the data.*

Considérons N points de données uniforméments répartis dans une boule unitaire de dimension p
Supposons que l'on se concentre au plus proche voisins de l'origine, la distance médiane de l'origine jusqu'au point de donnée le plus proche est donné par l'expression

$$d(p,N)=(1-\frac{1}{2}^{\frac{1}{N}})^{\frac{1}{p}}$$

Il existe une expression plus compliquée de la distance moyenne au plus proche voisin..
Pour N=500, p=10 d(p,N)=0,52 soit plus de la moitié de la limite du cercle !

Mean squard error (MSE)

$$MSE(x_0) = E_T[f(x_0) - \hat{y}_0]^2 = E_T[\hat{y}_0 - E_T(\hat{y}_0)]^2 + [E_T(\hat{y}_0) - f(x_0)]^2 = Var_T(\hat{y}_0) + Bias^2(\hat{y}_0)$$

On a divisé la MSE en deux parties : la variance et le biais au carré.
Une décomposition de ce type est toujours possible et souvent pratique.
Il est connu sous le non de décomposition biais-variance.

Pour être dans des dimensions importantes, il faut que le nombre de données augmente exponentiellement pour que le plus proche voisins de 0 soit assez proche.
Donc si la méthode des plus proches voisins en haute dimensions est inconcevable car les points seront beaucoup trop éloignés de par exemple 0.

Supposons que la relation entre Y et X est linéaire $Y = X^T\beta + \varepsilon$ avec $\varepsilon \approx N(0, \sigma^2)$
et que l'on utilise les moindres carrés

PAGE 44 5,76 % fait

βσε