Module 9 Juin - Sélection et évaluation de modèles

Florence d'Alché-Buc,
florence.dalche@telecom-paristech.fr

Telecom Evolution, Paris, France

Objectifs

- Rappel sur la thorie de l'apprentissage statistique
- Sélection et évaluation de modèles

Rappel: apprentissage supervisé par appro régularisée

Construire une fonction $\hat{h}$ un classifieur de $\mathbb{R}^p$ vers $\{-1, 1\}$ (resp. $\{1 \ldots, C\}$) telle que:

$$\hat{h} = \arg\min_{h \in} \sum_{i=1}^{n} L(y_i, h(x_i)) + \lambda \Omega(h) \qquad (1)$$

- L est une **fonction de perte locale**: L mesure quel point $h(x_i)$ est proche de y_i la valeur de sortie désirée.
- $\sum_{i=1}^{n} L(y_i, h(x_i))$: terme d'attache aux données
- $\Omega(h)$: pénalité sur la fonction h, contrôle la complexité du modèle, par exemple la norme au carré du vecteur de paramètres de h.

NB: l'approche régularisée n'est pas la seule approche possible, mais c'est la plus courante !

Résoudre un problème d'apprentissage d'u classifieur

Méthodologie pour développer une approche discriminante

- Définir
  - l'**espace de représentation** des données

Résoudre un problème d'apprentissage d'u classifieur

Méthodologie pour développer une approche discriminante

- Définir
  - l'**espace de représentation** des données
  - la **classe des fonctions** de classification binaire considérées

Résoudre un problème d'apprentissage d'u classifieur

Méthodologie pour développer une approche discriminante

- Définir
  - l'**espace de représentation** des données
  - la **classe des fonctions** de classification binaire considérées
  - la **fonction de coût** à minimiser pour obtenir le meilleur classifieur dans cette classe

Résoudre un problème d'apprentissage d'u classifieur

Méthodologie pour développer une approche discriminante

- Définir
  - l'**espace de représentation** des données
  - la **classe des fonctions** de classification binaire considérées
  - la **fonction de coût** à minimiser pour obtenir le meilleur classifieur dans cette classe
  - l'**algorithme de minimisation** de cette fonction de coût

Résoudre un problème d'apprentissage d'u classifieur

Méthodologie pour développer une approche discriminante

- Définir
  - l'**espace de représentation** des données
  - la **classe des fonctions** de classification binaire considérées
  - la **fonction de coût** à minimiser pour obtenir le meilleur classifieur dans cette classe
  - l'**algorithme de minimisation** de cette fonction de coût
  - une **méthode de sélection de modèle** pour définir les hyperparamètres

Résoudre un problème d'apprentissage d'u classifieur

Méthodologie pour développer une approche discriminante

- Définir
  - l'**espace de représentation** des données
  - la **classe des fonctions** de classification binaire considérées
  - la **fonction de coût** à minimiser pour obtenir le meilleur classifieur dans cette classe
  - l'**algorithme de minimisation** de cette fonction de coût
  - une **méthode de sélection de modèle** pour définir les hyperparamètres
  - une méthode d'évaluation des performances

Sélection ou évaluation de modèle ?

- Estimer les performances de différents modèles afin de choisir le meilleur : sélection de modèle
- Ayant choisi un modèle, estimer son erreur en généralisation (le vrai risque) : évaluation de modèle

Aujourd'hui, nous nous concentrons sur la première de ces deux questions.

Premier exemple

Un classifier linéaire dans le plan:

$$h(x) = \text{signe}(\beta_1 x_1 + \beta_2 x_2 + \beta_0) \tag{2}$$

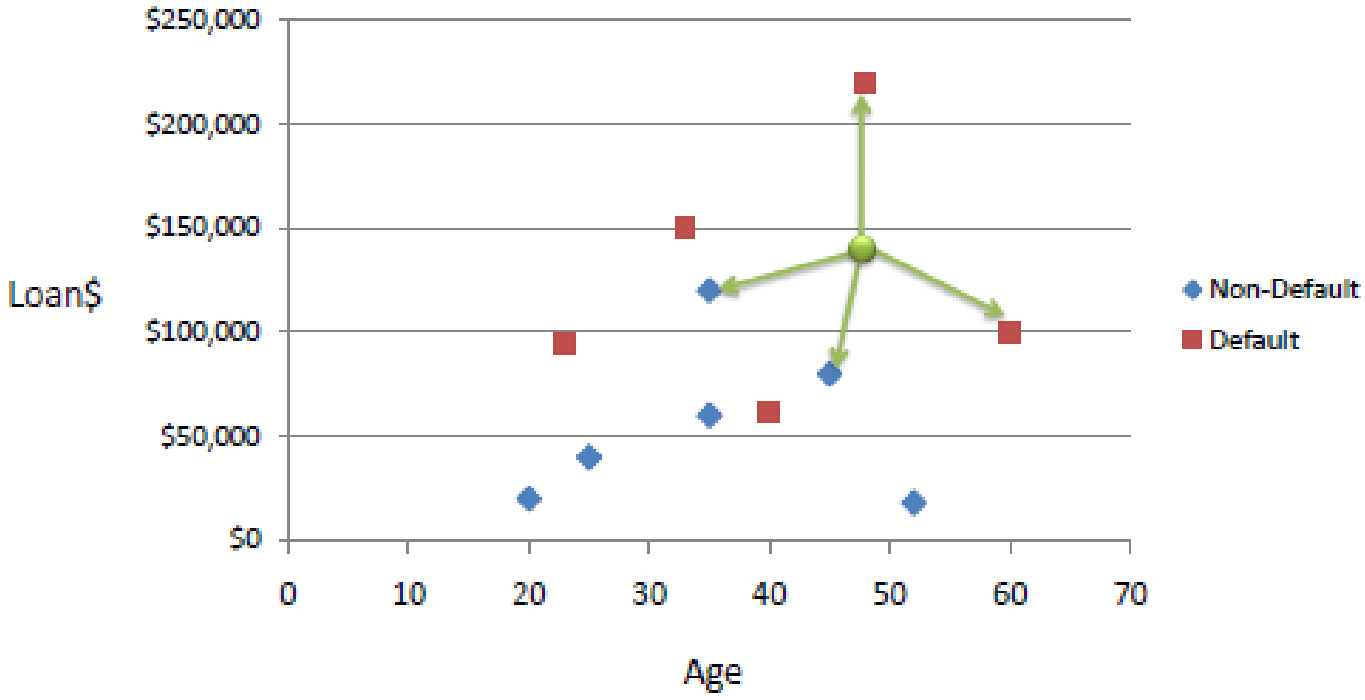
Apprendre h_β en minimisant: $\sum_{i=1}^{n} L(y_i, h(x_i)) + \lambda \|\beta\|_2^2$
OU $\sum_{i=1}^{n} L(y_i, h(x_i)) + \lambda \|\beta\|_1$ Quelle valeur de λ choisir ?

Premier exemple

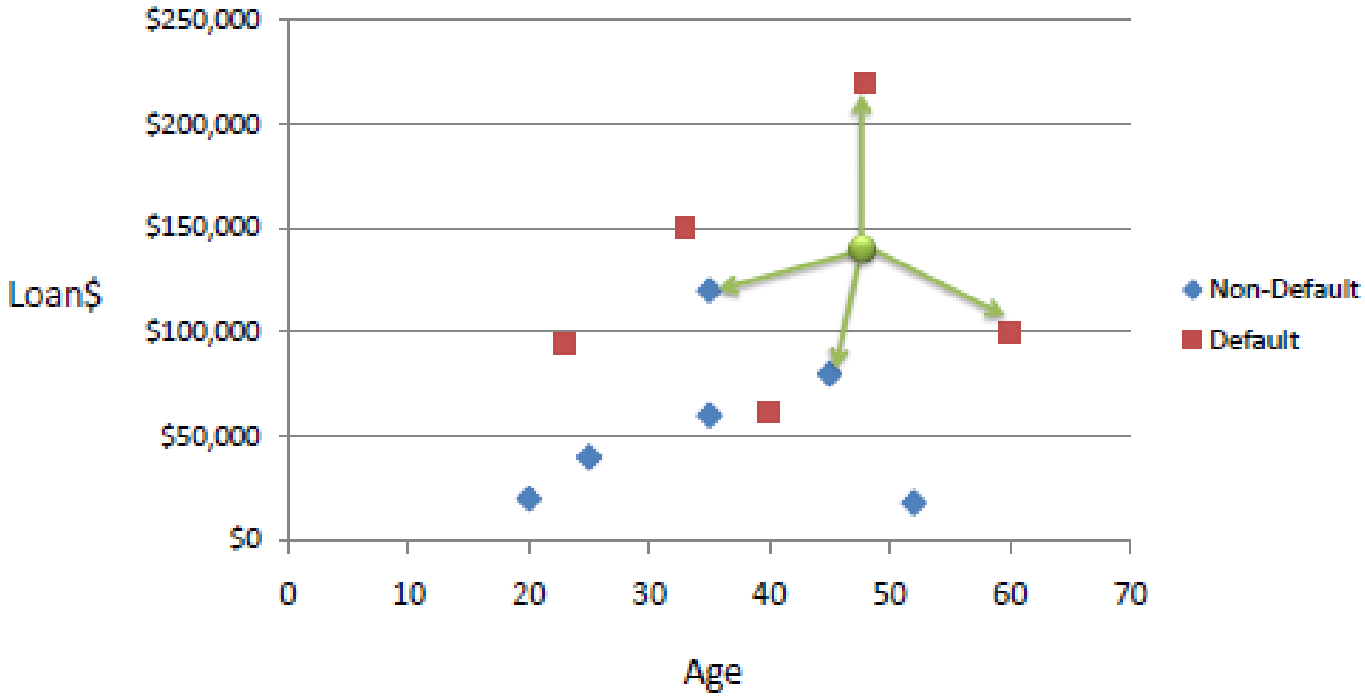
1. Quelle valeur de λ choisir ? Algorithme de sélection $\tilde{\lambda}$
2. Une fois que $\tilde{\lambda}$ est choisi, j'applique mon algorithme d'apprentissage et j'obtiens $\hat{h}_{\tilde{\lambda}}$: comment évaluer ses performances ?

Deuxième exemple

Le classifieur des k-plus-proches voisins:



Classifieur des K-plus-proches voisins



Classifieur des K-plus-proches voisins

K-PPV (en anglais K-Nearest neighbors: K-NN)

Cas 2 classes:

$$h_{KNN}(x) = \arg \max_{y \in \{-1,1\}} \frac{N_y^K(x)}{K},$$

avec :

- Soit K un entier strictement positif.
- Soit d une métrique définie sur $\times$
- $S = \{(x_i, y_i), i = 1, \ldots, n\}$
- Pour une donnée x, on définit la permutation d'indices $(\cdot)$ dans $\{1, \ldots, n\}$ telle que:
  - $d(x, x_{(1)}) \leq d(x, x_{(1)}) \ldots \leq d(x, x_{(n)})$
- $S_x^K = \{x_{(1)}, \ldots, x_{(K)}\}$: K premiers voisins de x
- $N_y^K(x) = |\{x_i \in S_x^K, y_1 = y\}|$

Régression par K-plus-proches voisins

K-PPV (en anglais K-Nearest neighbors: K-NN)

$$\hat{f}_{KNN}(x) = \frac{1}{L} \sum_{\ell=1}^{K} y_{(\ell)}$$

avec :

- Soit K un entier strictement positif.
- Soit d une métrique définie sur $\times$
- $S = \{(x_i, y_i), i = 1, \ldots, n\}$
- Pour une donnée x, on définit la permutation d'indices $(\cdot)$ dans $\{1, \ldots, n\}$ telle que:
  - $d(x, x_{(1)}) \leq d(x, x_{(1)}) \ldots \leq d(x, x_{(n)})$
- $S_x^K = \{x_{(1)}, \ldots, x_{(K)}\}$: K premiers voisins de x

Le paramètre de lissage K

Comment choisir K ? K: trop petit : la fonction f est trop sensible aux données : large variance
K : trop large : la fonction f devient trop peu sensible aux donnes : biais important

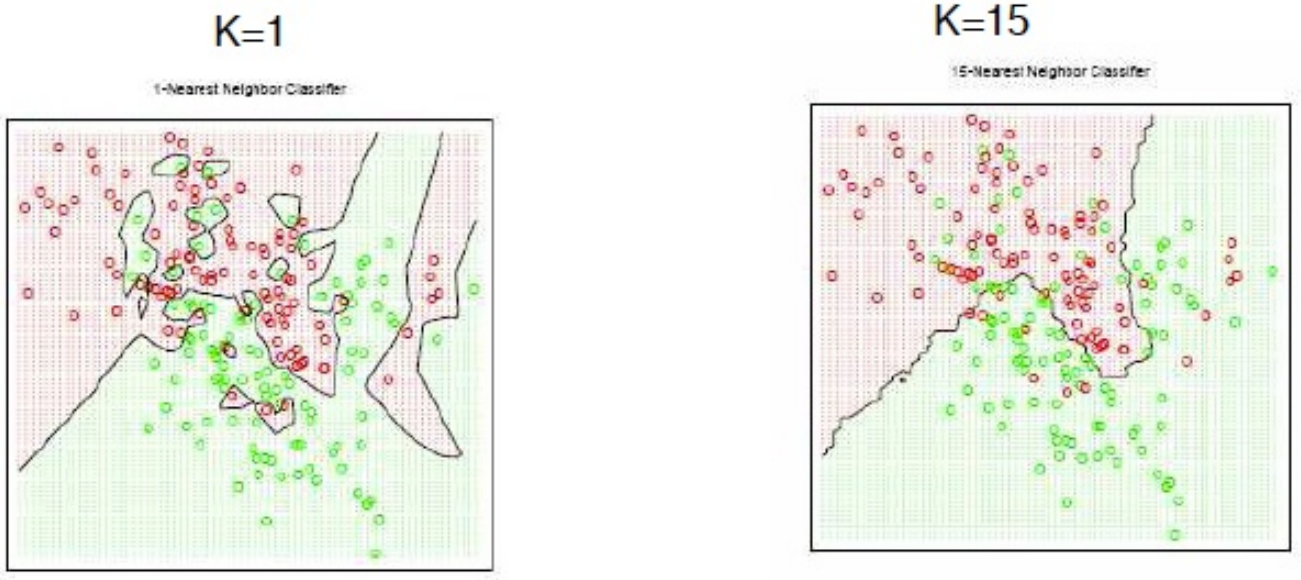


Fig 2.2, 2.3 of HTF01

Book of Hastie, Tibshirani and Friedman (The elements of statistical learning, Springer)

Question: Tracer la frontière de décision lorsque K = 50

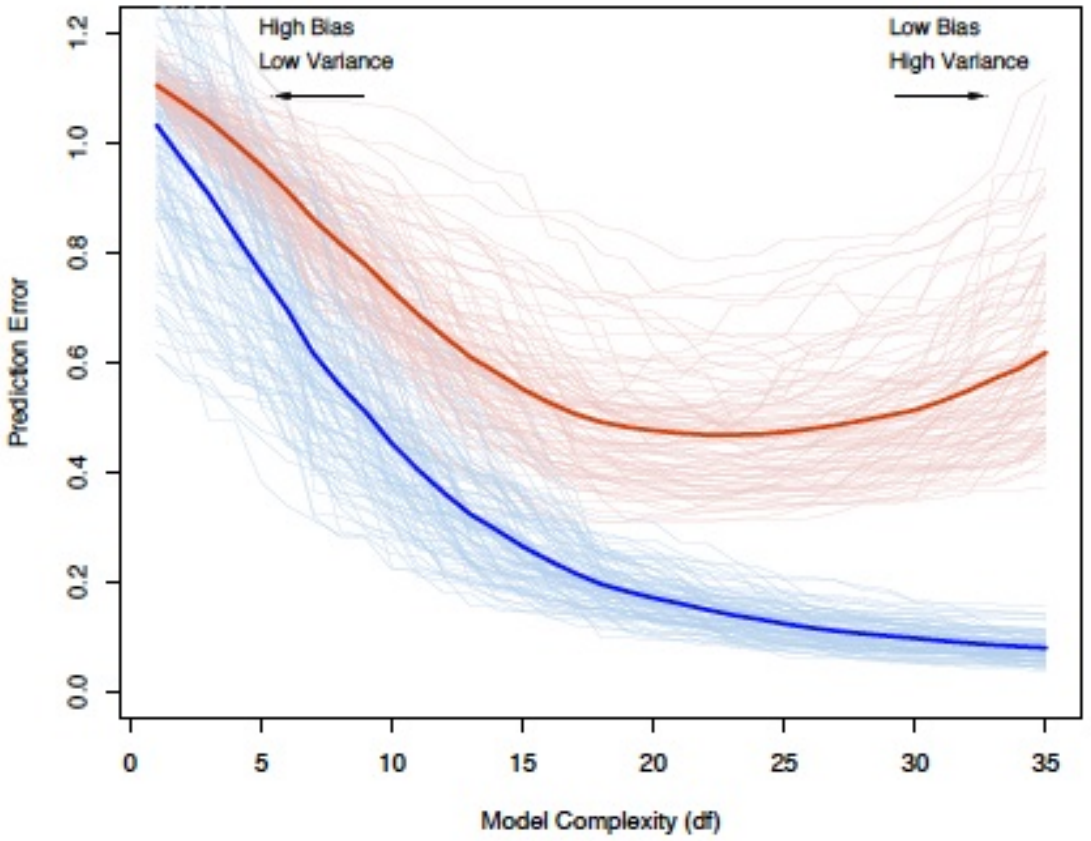
Décomposition biais-variance

Calcul du risque (de l'erreur en généralisation)
On suppose: $Y = f(X) +$ avec ϵ centré et de variance σ_ϵ^2.

$$
\begin{aligned}
E[(Y - \hat{f}(X))^2] &= E[Y^2 + \hat{f}(X) - 2Y\hat{f}(X)] \\
&= E[Y^2] + E[\hat{f}(X)^2] - 2E[Y\hat{f}] \\
&= VarY + E[Y]^2 + Var\hat{f}(X) + [f]^2 - 2E[f(X) + \epsilon]E[\hat{f}(X)] \\
&= \sigma_\epsilon^2 + E[f(X) + \epsilon]^2 + E[\hat{f}]^2 - 2E[f(X)]E[\hat{f}(X)] + Var\hat{f}(X) \\
&= \sigma_\epsilon^2 + E[\hat{f}(X) - f(X)]^2 + Var\hat{f}(X) \\
&= \sigma_\epsilon^2 + \text{Biais}^2 + \text{variance}
\end{aligned}
$$

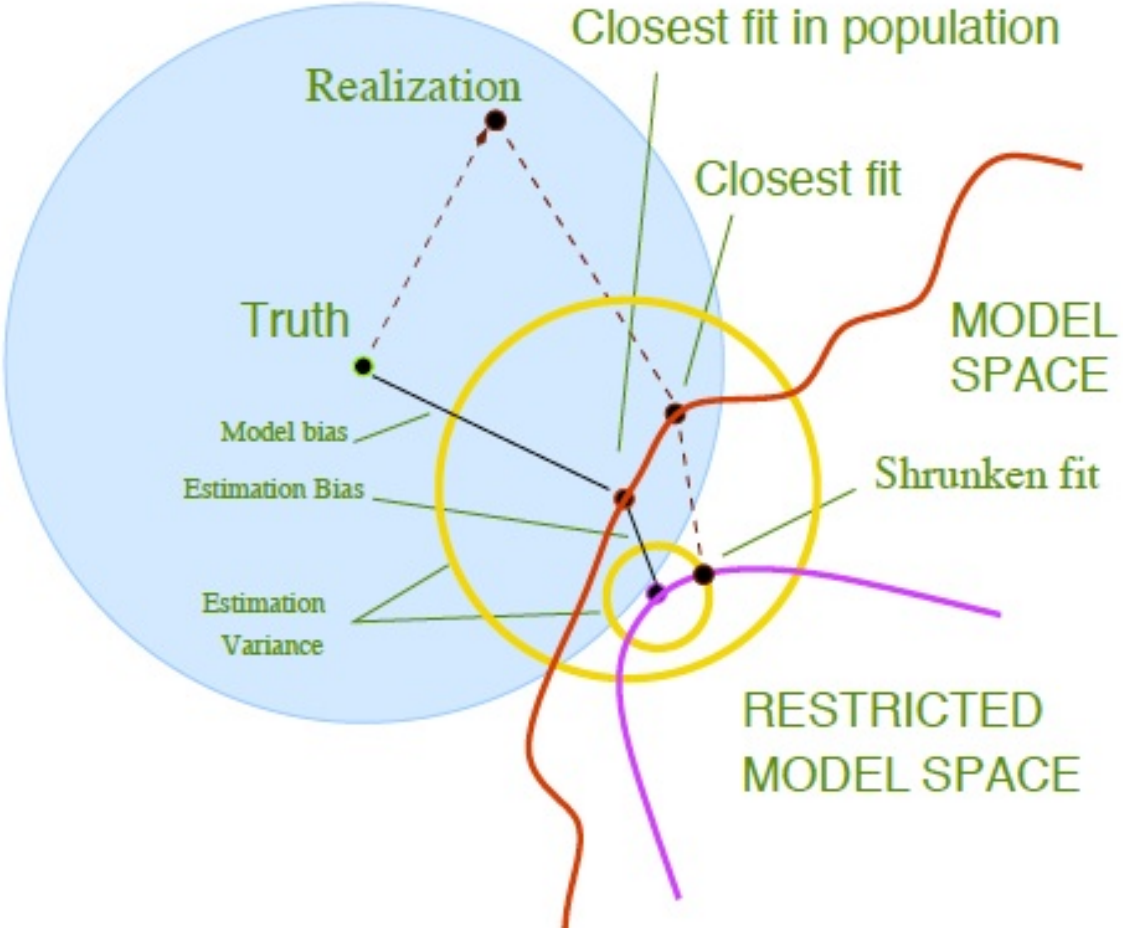
terme incompressible : bruit des données
Biais au carré: mesure à quel point $\hat{f}$ est loin de la cible
Variance de $\hat{f}(X)$: mesure à quel point $\hat{f}(X)$ est sensible aux données

Biais variance



Book of Hastie, Tibshirani and Friedman (The elements of statistical learning, Springer)

Biais variance



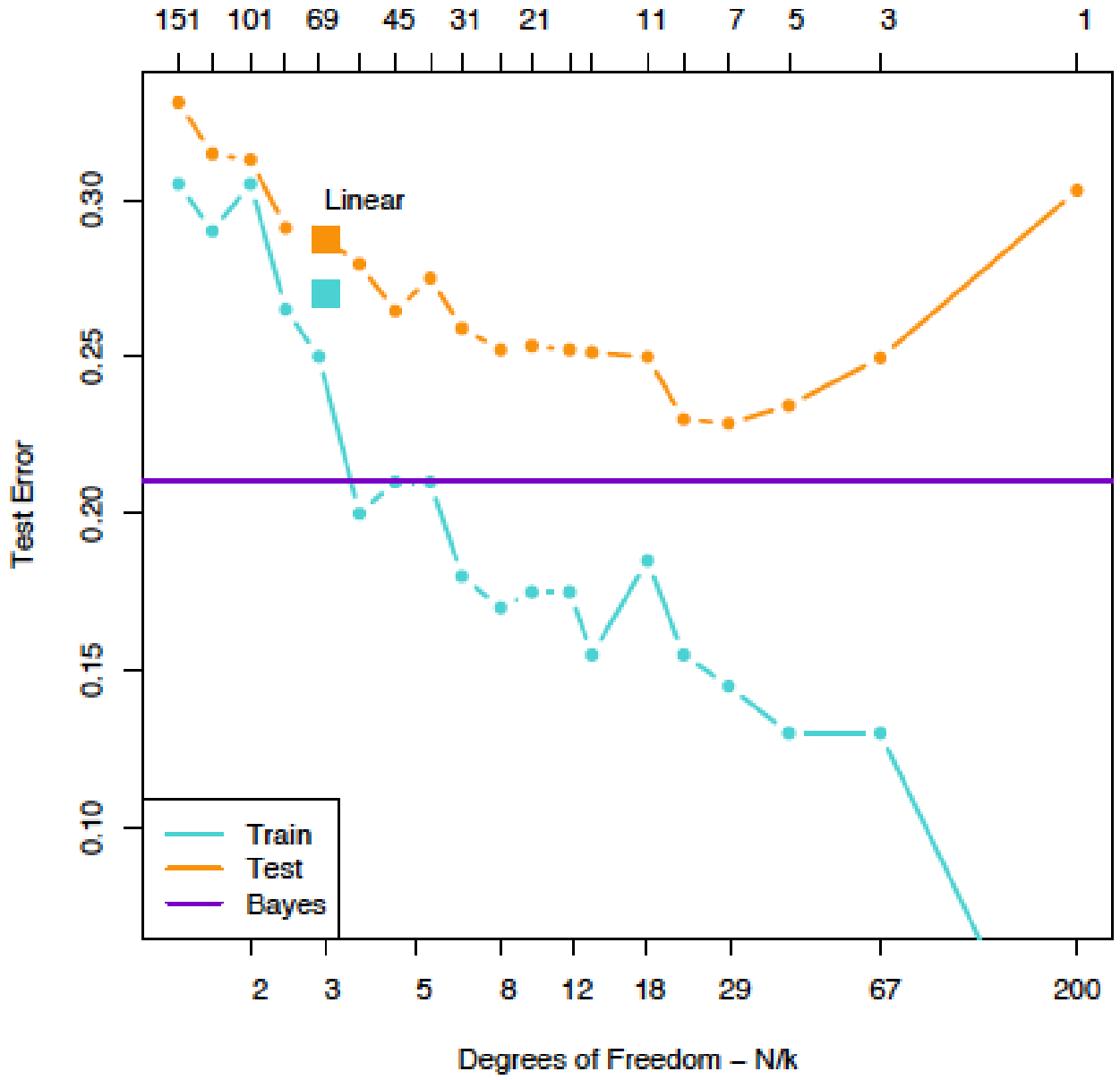
Décomposition biais-variance des k-plus-proches voisins

Posons x_0. Supposons que l'ala ne vient que des y et pas des x. On peut montrer que:

$$E[(Y - \hat{f}(x_0))^2] = \sigma_\epsilon^2 + (f(x_0) - \frac{1}{K}\sum_{\ell=1}^{K} f(x_{(\ell)})^2 + \frac{\sigma_\epsilon}{K}$$

K contrôle le terme de variance : plus grande est la valeur de K, plus la variance décroit; masi K contrôle aussi le biais, plus petite est la valeur de K, plus petit est le biais : nous sommes en plein *dilemne biais-variance*.

Erreur de test en fonction de $\frac{n}{K}$



Book of Hastie, Tibshirani and Friedman (The elements of statistical

Deuxième exemple

Le classifieur des k-plus-proches voisins: classifier paresseux : pas besoin d'algorithme d'apprentissage ! J'ai besoin des données dites d'apprentissage, d'une métrique et de la valeur de k.

- Comment choisir la valeur de K ?
- Ayant choisi $\tilde{K}$, comment estimer l'erreur en généralisation de ce K-NN ?

Estimation de l'erreur par validation croisé

- Méthode robuste pour estimer l'erreur en généralisation, i.e. le vrai risque d'un algorithme d'apprentissage
- Références
  - Allen 1977
  - Efron, Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation, JASA, June,1983
  - Surtout le chapitre 7 du livre The Iments of statistical learning, Hastie, Tibshirani, Frideman.

1. Diviser les données S en B parties de même taille et disjointes $S_{b=1}, \ldots, D_{b=B}$ avec $|D_b| = n/B$.
2. Pour $b \in \{1, \ldots, B\}$:
   - Entrainer le modèle $\mathcal{H}_\lambda$ sur toutes les données **sauf** D_b pour obtenir un estimateur $\hat{h}^b_{\lambda,n}$
   - Calculer **sur les données restantes** D_b (test) le risque empirique

$$R_{n,b}(\lambda) = \frac{1}{n/B} \sum_{j \in D_b} L(x_j, y_j, \hat{h}^b_{\lambda,n})$$

3. Calculer le risque empirique moyen de λ (dit 'de cross-validation')

$$R^B_{n,CV}(\lambda) = \frac{1}{B} \sum_{b=1}^{k} R_{n,b}(\lambda) \tag{3}$$

Trouver λ

Répéter cette procédure sur tous les $\lambda \in \Lambda$ considérés (ou sur une grille sur λ est un paramtre continu) et choisir

$$\hat{\lambda}_{n,B} = \arg\min_{\lambda \in \Lambda} R_{n,CV}(\lambda). \tag{4}$$

Ensembles de validation, apprentissage, tes

- On sélectionne sur $\mathcal{S}_{val}$
- On apprend sur $\mathcal{S}_{app}$ en utilisant $\tilde{\lambda}$
- On teste sur $\mathcal{S}_{test}$

$\text{Err}_{CV,val}$ nous dit à quel type d'erreur en généralisation nous attendre en apprenant sur un ensemble de taille $n_{val} - n_{val}/B$. $\text{Err}_{\mathcal{S}_{app}}$ nous dit à quel point le classifieur a bien réussi à approcher les données d'apprentissage
$\text{Err}_{\mathcal{S}_{test}}$ nous dit à quel point le classifieur a bien réussi à approcher les données (nouvelles) de test
NB: Dans de nombreuses études, on se contente de sélectionner les hyperparamètres sur $\mathcal{S}_{val} = \mathcal{S}_{app}$ et de ré-apprendre sur $\mathcal{S}_{app}$, l'essentiel étant de ne pas utiliser les données de test pendant la phase de sélection de modèlles et la phase d'apprentissage.