Introduction

masedki.github.io

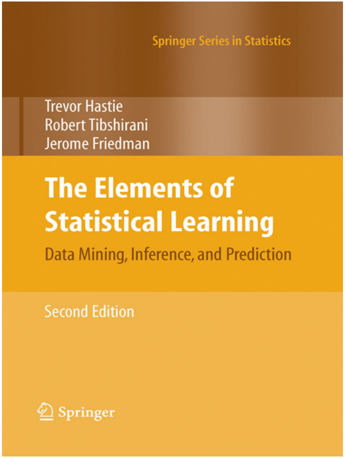
lun. 23 mars 2020

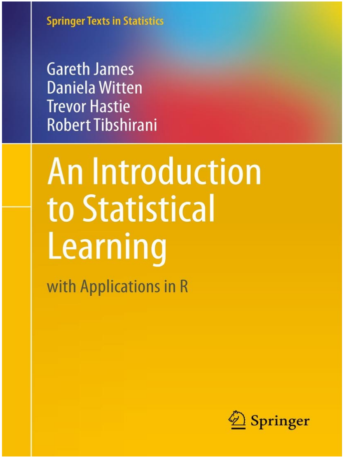
Outline

1. Introduction

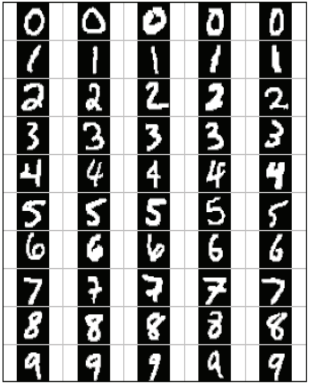
2. Régression vs classification supervisée

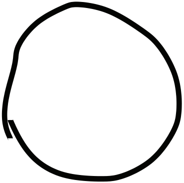
Références


Springer Series in Statistics
Trevor Hastie
Robert Tibshirani
Jerome Friedman
The Elements of
Statistical Learning
Data Mining, Inference, and Prediction
Second Edition
Springer


Springer Texts in Statistics
Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani
An Introduction
to Statistical
Learning
with Applications in R
Springer

Problème d'apprentissage





- Reconnaissance de chiffres manuscrits ? 0, 1, 2 ... ?

Problèmes d’apprentissage statistique

- Identifier les facteurs de risque du cancer de la prostate
- Classifier des phonèmes à partir de périodogrammes
- Prédire si une personne est sujette aux crises cardiaques, à partir de mesures cliniques, son régime et des données démographiques
- Personnaliser un système de détection de spam email
- Lecture de codes postals écrits à la main
- Classification d’échantillons de tissus dans différents types de cancer, en fonction de données d’expression de gènes
- Établir une relation entre salaires et variables démographiques
- Classifier les pixels d’une image satellite

Question

- Sur 4 601 mails, on a pu identifier 1813 spams.
- On a également mesuré sur chacun de ces mails la présence ou absence de 57 mots.

Peut-on construire à partir de ces données une méthode de détection automatique de spam ?

Outline

1. Introduction

2. Régression vs classification supervisée

Réprésentation du problème

- La plupart de ces problèmes peuvent être appréhendés dans un contexte de régression : on cherche à expliquer une variable Y par d'autres variables dites explicatives $X_1, \ldots, X_p$:

Y	X
Chiffre	image
Mot	courbe
Spam ou pas	présence/absence d'un ensemble mots
Type de leucémie	expressions de gênes

- Lorsque la variable à expliquer est quantitative, on parle de régression.
- Lorsqu'elle est qualitative, on parle de discrimination ou classification supervisée.

Régression

- Un échantillon i.i.d d'apprentissage $(X_1, Y_1), \ldots, (X_n, Y_n)$ d'une loi conjointe $\mathbb{P}$ **inconnue** sur $\mathbb{R}^p \times \mathbb{R}$.
- Objectif : Prédire ou expliquer la variable Y à partir d'une nouvelle observation X.
- Méthode : construire une règle de prédiction (ou régression)

$$m : \mathbb{R}^p \mapsto \mathbb{R}.$$

- Soit $l : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}_+$ une fonction de perte (i.e, $l(y, y') = 0$ et $l(y, y') > 0$ pour $y \neq y'$), par exemple

$$l(y, y') = |y - y'|^q$$

(perte absolue si $q = 1$ et perte quadratique $q = 2$).

Risque ou erreur de généralisation

- Le risque ou erreur de généralisation d'une règle de décision (ou prédiction) m est défini par

$$R_{\mathbb{P}}(m) = \mathbb{E}_{(X,Y)}\Big[l(Y, m(X))\Big].$$

La fonction de régression

- Un champion
$$m^*(x) = \mathbb{E}[Y|X = x]$$
appelé fonction de régression.
- Pour toute autre fonction m, on a
$$\mathbb{E}\left[\left(Y - m^*(X)\right)^2\right] \leq \mathbb{E}\left[\left(Y - m(X)\right)^2\right]$$

.

La fonction de régression

Nous avons

$$\mathbb{E}_{X,Y}\left[\left(Y - m(X)\right)^2\right] = \mathbb{E}_X \mathbb{E}_{Y|X}\left[\left(Y - m(X)\right)^2 \mid X\right]$$

Donc il suffit de minimiser cette erreur ponctuellement en X

$$m(x) = \text{argmin}_c \mathbb{E}_{Y|X}\left[\left(Y - c\right)^2 \mid X = x\right].$$

La solution est donnée par

$$m^*(x) = \mathbb{E}[Y \mid X = x]$$

La classification binaire

- Un échantillon i.i.d d'apprentissage $(X_1, Y_1), \ldots, (X_n, Y_n)$ d'une loi conjointe $\mathbb{P}$ **inconnue** sur $\mathbb{R}^p \times \{0, 1\}$.
- Objectif : Prédire ou expliquer la variable Y à partir d'une nouvelle observation X.
- Méthode : construire une règle classification (ou décision)

$$g : \mathbb{R}^p \mapsto \{0, 1\}.$$

- La fonction de perte binaire $l(y, y') = \mathbb{1}_{y \neq y}.$

- Risque associé à g : taux de mauvais classement

$$R_{\mathbb{P}}(g) = \mathbb{E}\Big[l(g(X), Y)\Big] = \mathbb{P}(g(X) \neq Y).$$

La règle de Bayes

- Un champion appelé règle de Bayes

$$g^*(x) = \begin{cases} 1 & \text{si} \quad \eta(x) \geq \frac{1}{2} \\ 0 & \text{sinon,} \end{cases}$$

où $\eta(x) = \mathbb{P}(Y = 1 | X = x)$.

- Quelque soit la règle de décision g, nous avons

$$R_{\mathbb{P}}(g^*) = \mathbb{P}(g^*(X) \neq Y) \leq \mathbb{P}(g(X) \neq Y) = R_{\mathbb{P}}(g).$$

Règle de Bayes : un théorème

Pour toute règle de classification $g : \mathcal{X} \mapsto \mathcal{Y}$, pour la fonction de perte binaire, nous avons

$$R(g) - R(g^*) = \mathbb{E}_X\left[\mathbb{1}\left\{g(X) \neq g^*(X)\right\}\left|2\eta(X) - 1\right|\right].$$

- Interpréter ce résultat lorsque

$$\eta(x) = \frac{1}{2}, \forall x \in \left\{x \in \mathcal{X} : g(x) \neq g^*(x)\right\}$$

Début de preuve

On remarque que :

$$\begin{aligned}\mathbb{E}_{Y|X=x}\Big[\mathbb{1}\{Y = g^*(x)\}\Big] &= \mathbb{P}_{Y|X}[Y = g^*(x)] \\ &= \begin{cases} \eta(x) & \text{si} \quad \eta(x) \geq \frac{1}{2} \\ 1-\eta(x) & \text{sinon} \end{cases} \\ &= \frac{1}{2} + \Big|\eta(x) - \frac{1}{2}\Big|.\end{aligned}$$

Rappel :

$$\mathbb{E}_{X,Y} h(X,Y) = \mathbb{E}_X \mathbb{E}_{Y|X} h(X,Y).$$

suite de la preuve : voir notes

Proposition

$$R^* = R(g^*) = \mathbb{E}_X\left[\text{min}\left\{\eta(X), 1 - \eta(X)\right\}\right]$$

Preuve : voir notes !!

Problème majeur !!

- Problème: m^* est inconnu en pratique. Il faut construire un régresseur $\widehat{m}_n$ à partir des données $(X_1, Y_1), \ldots, (X_n, Y_n)$, tel que
$$\widehat{m}_n(x) \approx m^*(x).$$
- Problème: g^* est inconnue en pratique. Il faut construire une règle $\hat{g}_n$ à partir des données $(X_1, Y_1), \ldots, (X_n, Y_n)$, tel que
$$\hat{g}_n(x) \approx g^*(x).$$

Un candidat naturel

- À partir des expressions de m^* et g^*, proposer deux estimateurs intuitifs.

Décomposition de l'erreur

Pour tout estimateur $\widehat{m}_n(x)$ de $m^*(x)$ à x fixé, nous avons

$$\begin{aligned}
\mathbb{E}\left[\left(m^*(x) - \widehat{m}_n(x)\right)^2\right] &= \left[m^*(x)\right]^2 - 2m^*(x)\mathbb{E}\left[\widehat{m}_n(x)\right] \\
&\quad + \mathbb{E}\left[\left(\widehat{m}_n(x)\right)^2\right] \\
&= \left[m^*(x) - \mathbb{E}(\widehat{m}_n(x))\right]^2 \\
&\quad + \mathbb{E}\left[\left(\widehat{m}_n(x)\right)^2\right] - \left[\mathbb{E}(\widehat{m}_n(x))\right]^2 \\
&= \left(\text{biais}\right)^2 + \text{Var}\left[\widehat{m}_n(x)\right]
\end{aligned}$$

Notations

- On s'intéresse au cas où on cherche à expliquer une variable qualitative Y par p variables explicatives $X_1, \ldots, X_p$.
- Y est à valeurs dans un ensemble discret fini de modalités qui peuvent être numérotées par des les indices $\{1, 2, \ldots, K\}$ et les variables $X_1, \ldots, X_p$ peuvent être qualitatives et/ou quantitatives.
- Néanmoins, pour présenter les méthodes, on se restreint au cas où Y est à 2 modalités (0 et 1).

Table of Contents