Statistical Learning I

1. Organization
2. Statistical Learning Intro:
   a) supervised learning: classification, regression (overfitting 1)
   b) simplistic: linear regression for classification (overfitting 2)
   c) nearest neighbour (overfitting 3)
3. Statistical decision theory:
   EPE (overfitting 4), Bayes optimal decision
5. Bias-variance (overfitting 5)
4. Curse of dimensionality — niet aan toegekomen
6. History, applications, beyond classification/regression

1. Organization

Tim van Erven
TA: Kevin Duisters

www.timvanerven.nl/teaching/statlearn2015

(deelnemerslijst uitdelen, boek laten zien)

- 2 large homework assignments > 5.5
- open book exam: Jan 6 > 5.5

HW1: Nov 12 - Dec 1
HW2: Dec 1 - end of Dec.

(first two weeks: 3 hours of lectures
then: 2 to 3 hours + work on homework)

Enroll in Blackboard + Usis

2a) Supervised Learning

teacher shows us what to do)

Training data $T = \begin{pmatrix} y_1 \\ x_1 \end{pmatrix}, \begin{pmatrix} y_2 \\ x_2 \end{pmatrix}, \dots, \begin{pmatrix} y_N \\ x_N \end{pmatrix}$

x_i : feature vectors
y_i : class / response variable

Goal: learn function $\hat{f}: X \to Y$
from model $\mathcal{F}$.

Evaluate $\hat{f}$ on test data:
new x from same source,
predict corresponding Y by $\hat{y} = \hat{f}(x)$.

Assume: $\begin{pmatrix} y_i \\ x_i \end{pmatrix}$ independent from same probability distribution P^*

Avoid further assumptions on P^*!
(P^* can be very complicated.)

Classification:

$\begin{pmatrix} G_1 \\ x_1 \end{pmatrix}, \dots, \begin{pmatrix} G_N \\ x_N \end{pmatrix}$

G_i is categorical variable (e.g. yes/no or apple/orange/pear)

$G_i \to Y$

yes → 1
no → 0

yes → +1
no → −1

apple → 1
orange → 2
pear → 3

apple → $\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$
orange → $\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$
pear → $\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$

"dummy coding"

(see slides)

Linear Regression:

Y_i is real-valued (X_i in reals or use dummy coding)
model $\mathcal{F}$ can be represented as linear functions:

$$f_\beta(x_i) = \hat{Y}_i = \beta_0 + \sum_{j=1}^{p} x_{i,j}\,\beta_j$$

$$= x_i^T \beta \qquad (x_{i,0} = \text{always } 1)$$

Least Squares: choose $\hat{\beta}$ to minimize

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

NB X_i random design, not set by experimenter

(slides)

Important trick: make more features

* Given $X_i \in \mathbb{R}$
* Make $\tilde{X}_{i,0} = 1$
  $\tilde{X}_{i,1} = X_i$
  $\tilde{X}_{i,2} = X_i^2$
  $\tilde{X}_{i,3} = X_i^3$
  etc.

Model linear in new features, but not in old features

(polynomial slides)

OVERFITTING is central concern

2b. Linear Regression for Classification

NB. Logistic regression is better!

$G_i \in \{\text{orange, blue}\} \rightarrow y_i \in \{1, 0\}$

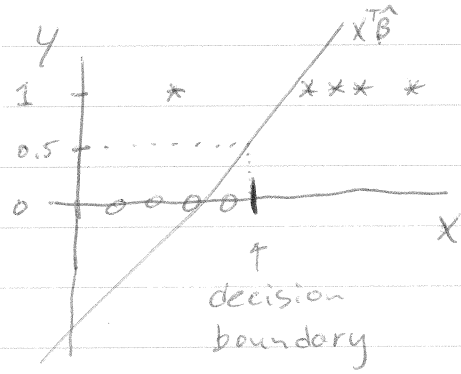
$x_i \in \mathbb{R}^2$

Simple idea:

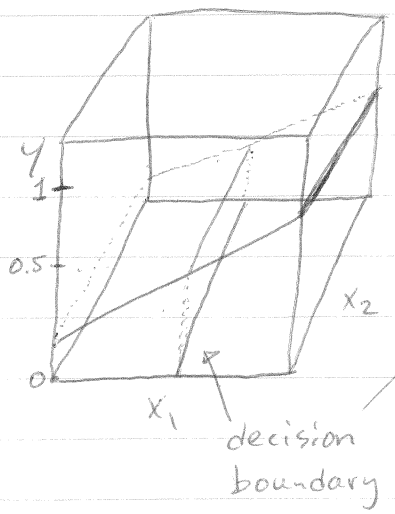
- use linear regression on this data
- convert back to binary predictions:

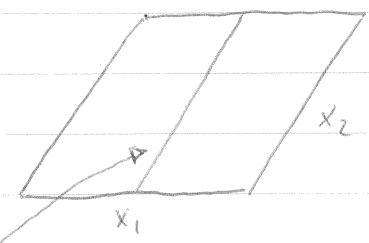
$x^T\hat{\beta} > 0.5 \Rightarrow \hat{f}(x) = 1$

$x^T\hat{\beta} \leq 0.5 \Rightarrow \hat{f}(x) = 0$



(1 feature)





(2 features)

(show first figure from book)

- Many errors on training set: avoidable?
- Maybe if we construct more features:
  $X_1, X_2, X_1^2, X_2^2, X_1 \cdot X_2, X_1^2 \cdot X_2, \ldots$
  Too few: underfitting
  Too many: overfitting

2C. K-Nearest Neighbour Classification

To classify new X:
- Ask k points in T closest to X
- Choose most common class

For two classes: $y_i \in \{0, 1\}$

$$\hat{y} = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i$$

⇐ estimates $Pr(Y=1/X)$ by average in region around X.

$\hat{y} > 0.5 \Rightarrow$ predict 1
$\hat{y} \leq 0.5 \Rightarrow$ predict 0

k small: very flexible decision boundary, but very unstable estimates based on few neighbours (risks overfitting)

k large: stable estimates, but inflexible decision boundary (risks underfitting)

Model $\hat{f}$ is implicit!

If neighbourhoods do not overlap: $\frac{N}{k}$ neighbourhoods
⇒ effective nr. of parameters, because estimate $\hat{y}$ for each neighbourhood.

(Show pictures book,
what is best k?
Need statistical decision theory!)

3 Statistical Decision Theory

Learn $\hat{f}$ from training data $\binom{y_1}{x_1}, \ldots, \binom{y_N}{x_N} \overset{i.i.d.}{\sim} p^*$
Evaluate on new $(x, y) \sim p^*$

Loss function $L(y, \hat{y})$ measures how bad it is if truth is y, but we predict $\hat{y} = \hat{f}(x)$
(smaller is better)

Regression: $L(y, \hat{y}) = (\hat{y} - y)^2$ "squared error"
Classification: $L(y, \hat{y}) = \begin{cases} 0 & \text{if } \hat{y} = y \\ 1 & \text{if } \hat{y} \neq y \end{cases}$ "0/1-loss"

Other choices possible?
(depends on what is important in your application)

$(x, y) \sim p^*$ random, so want small **expected prediction error**:

$$EPE(f) = \mathbb{E}_{x,y \sim p^*}\left[L(y, f(x)) \right]$$

often called "risk"

Classification:

$$EPE(f) = Pr(y \neq f(x)) \cdot 1 + Pr(y = f(x)) \cdot 0$$
$$= Pr(y \neq f(x))$$

Bayes Optimal Decisions

Bayes optimal f minimizes $EPE(f)$ over all possible functions:

$$f_B = \underset{f}{\operatorname{argmin}} \; \mathbb{E}_{x,y}\left[L(y, f(x)) \right]$$

$$f_B(x) = \underset{\hat{y}}{\operatorname{argmin}} \; \mathbb{E}_{P^*(y|x)}[L(y, \hat{y})]$$

- unknown, because depends on P^*
- No learning alg can do better (by definition)
- Nothing to do with "Bayesian statistics"

Classification with 0/1-loss:

$$f_B(x) = \underset{\hat{y}}{\operatorname{argmin}} \; P^*_{\xi}(y \neq \hat{y} | x)$$

$$= \underset{\hat{y}}{\operatorname{argmax}} \; P^*_{\xi}(y = \hat{y} | x)$$

$$= \begin{cases} 1 & \text{if } P^*_{\xi}(y=1|x) > 0.5 \\ 0 & \text{if } P^*(y=1|x) \leq 0.5 \end{cases} \quad \text{(for two classes)}$$

Regression with Squared Error:

$$f_B(x) = \underset{\hat{y}}{\operatorname{argmin}} \; \mathbb{E}_{P^*(y|x)}[(y - \hat{y})^2]$$

$$= \mathbb{E}_{P^*(y|x)}[y] = \mathbb{E}[y|x]$$

If $y = f^*(x) + \text{noise}$ (noise: mean 0)

Then $f_B(x) = \mathbb{E}[y|x] = f^*(x) + \mathbb{E}[\text{noise}|x]$

$$= f^*(x)$$

Empirical Risk Minimization

- Approximate P^* by empirical distribution on T and choose $f \in \mathcal{F}$ that minimizes

$$\frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i))$$

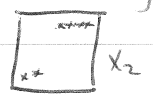
Is this always best? No

1. Overfitting (see k-NN picture train and test error as function of k)
2. Restricted to model $\mathcal{F}$

4. Curse of Dimensionality

1 feature	2 features	...	p features
N samples	N^2 samples for same coverage		N^p samples for same coverage

Conclusion: In high dimensions you only see samples from very small part of domain.



("A high-dimensional space is a lonely place." Bernard Schölkopf)

5. Bias-variance decomposition for Squared Error

Understand balance between under and overfitting

Thm: Let $\bar{f}(x) = \mathbb{E}_T[\hat{f}(x)]$. Then

$$\mathbb{E}_T \mathrm{EPE}(\hat{f}) = \mathbb{E}_X \mathrm{Var}(Y|X) \quad \leftarrow \text{Bayes error}$$

$$+ \mathbb{E}_X (f_B(x) - \bar{f}(x))^2 \quad \leftarrow \text{bias (typically smaller with larger } \mathcal{F})$$

$$+ \mathbb{E}_T \mathbb{E}_{X,Y} (\hat{f}(x) - \bar{f}(x))^2 \quad \leftarrow \text{variance (typically larger with larger } \mathcal{F})$$

Proof:

$$\mathop{\mathbb{E}}_{T,X,Y}\{Y-\hat{f}(x)\}^2 = \mathbb{E}\left(Y-f_B(x)+f_B(x)-\hat{f}(x)\right)^2$$

(subscripts belangrijk)

$$= \mathop{\mathbb{E}}_{X,Y}\left(Y-f_B(x)\right)^2$$

$$+ \mathop{\mathbb{E}}_{T,X,Y}\left(Y-f_B(x)\right)\left(f_B(x)-\hat{f}(x)\right)$$

$$+ \mathop{\mathbb{E}}_{T,X}\left(f_B(x)-\hat{f}(x)\right)^2$$

$$\mathop{\mathbb{E}}_{X,Y}\left(Y-f_B(x)\right)^2 = \mathop{\mathbb{E}}_{X}\left(Y-\mathbb{E}[Y|X]\right)^2 = \mathop{\mathbb{E}}_{X}\mathrm{Var}(Y|X).$$

$$\mathop{\mathbb{E}}_{T,X,Y}\left(Y-f_B(x)\right)\left(f_B(x)-\hat{f}(x)\right)$$

$$= \mathop{\mathbb{E}}_{T,X}\left(\mathbb{E}[Y|x]-f_B(x)\right)\left(f_B(x)-\hat{f}(x)\right)$$

$$= 0.$$

$$\mathop{\mathbb{E}}_{T,X}\left(f_B(x)-\hat{f}(x)\right)^2 = \mathop{\mathbb{E}}_{T}\left(f_B(x)-\bar{f}(x)+\bar{f}(x)-\hat{f}(x)\right)^2$$

$$= \mathop{\mathbb{E}}_{X}\left(f_B(x)-\bar{f}(x)\right)^2 \quad \Leftarrow \text{bias}$$

$$+ \mathop{\mathbb{E}}_{T,X}\left(f_B(x)-\bar{f}(x)\right)\left(\bar{f}(x)-\hat{f}(x)\right)$$

$$+ \mathop{\mathbb{E}}_{T,X}\left(\bar{f}(x)-\hat{f}(x)\right)^2 \Leftarrow \text{variance}$$

$$\mathop{\mathbb{E}}_{T,X}\left(f_B(x)-\bar{f}(x)\right)\left(\bar{f}(x)-\hat{f}(x)\right)$$

$$= \mathop{\mathbb{E}}_{X}\left(f_B(x)-\bar{f}(x)\right)\left(\bar{f}(x)-\mathop{\mathbb{E}}_{T}[\hat{f}(x)]\right)$$

$$= 0. \qquad \square$$