STATISTICAL LEARNING

22/2/2021

- Linear Regression
- Linear Discriminant Analysis (LDA, Fisher, 1936) → QDA
- Logistic Regression (1940s)
- Generalized Linear Models (1970s, Nelder)
- Classification and Regression trees (Breiman and Freidman, 1980s)
  → Extended to random forests (1995)
- Machine Learning
  → Support vector machines, neural networks, clustering, principal component analysis

Individual methods, details, theory, ...
Model accuracy, model selection, ...

BOOK James et al (2013)

Hastie et al (2001) The elements of statistical learning

ASSESSMENT 3 Homeworks

Final exam

TUTORATO Monday at 14:30 (remotely)

Starting on 8th March

Ch 2 What is Statistical Learning?

Example Wage data: wage and other variables of 3000 workers in US

Objective Predict wage from a number of factors (age, education, ...)

WAGE DATASET : Statistical learning problem

We could formalise this problem:

Y : response / dependent variable / outcome

$X = (X_1, \dots, X_p)$: p predictors / features / covariates / independent variables

In general:

$$Y = \underbrace{f(X)}_{\substack{\text{fixed} \\ \text{but } \underline{\text{unknown}} \\ \text{function} \\ \text{(systematic / deterministic)}}} + \underbrace{\varepsilon}_{\substack{\text{error (random, stochastic)} \\ E[\varepsilon] = 0 \\ \varepsilon \perp X}}$$

Statistical Learning : a set of approaches to estimate f !

② Why estimating f?

Two main reasons
- PREDICTION
- INFERENCE

PREDICTION

Predict Y when you only observe X

Since ε has average zero,

$$\hat{Y} = \hat{f}(X)$$

with $\hat{f}$ our estimate of f

In this context, $\hat{f}$ can be a black-box.

The accuracy of $\hat{Y}$ as a prediction of Y can be evaluated by:

$$E\left[\left(Y - \hat{f}(X)\right)^2 \mid X = x\right] = E\left[\left(f(X) + \varepsilon - \hat{f}(X)\right)^2\right] =$$

($\hat{f}$ fixed)

$$= E\left[\left(f(X) + \epsilon - \hat{f}(X)\right)^2\right]$$

$$= E\left[\left(f(X) - \hat{f}(X)\right)^2\right] + \underbrace{E[\epsilon^2]}_{= E[\epsilon^2] - (E[\epsilon])^2} + 2E\left(\left(f(X) - \hat{f}(X)\right)\epsilon\right)$$

$$= \left(f(X) - \hat{f}(X)\right)^2 + \mathrm{Var}(\epsilon) + \underbrace{2\left(f(X) - \hat{f}(X)\right) \cdot \underbrace{E(\epsilon)}_{0}}_{= 0}$$

$$= \underbrace{\left(f(X) - \hat{f}(\downarrow X)\right)^2}_{\text{reducible error}} + \underbrace{\mathrm{Var}(\epsilon)}_{\text{irreducible error}}$$

INFERENCE

Understanding how Y changes in terms of X

1. Which predictors are most important?
2. What is the relationship between Y and X_j?
3. Is the relationship between Y and X linear or non-linear?

3. How to estimate f?

Given a training data $(\underline{x}_i, y_i)$, $i = 1, \ldots, n$ with

$\underline{x}_i = (x_{i1}, \ldots, x_{ip})^t$ vector of observations for unit i

y_i response

Different ways :

(1) Parametric methods

We make an assumption about the functional form of f :

$$f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Fit & train the model on training data

→ parameter estimation

$$\hat{\beta}_0, \ldots, \hat{\beta}_p \rightarrow \hat{f}(X)$$

Disadvantage : True f could be far from our parametric assumption

Remark :

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_1^3 + \beta_4 X^4$$

Polynomial Regression

Linearity in the parameters !

(2) Non-parametric Methods

- No explicit assumption about $f(x)$
- Estimating f by getting as close as possible to the data, but without being too rough or wiggly

Why would we choose the more restrictive models?

- Interpretability (inference)
- generalizability outside of training data (prediction)

④ Bias-Variance Trade off

Let (x_0, y_0) be a new observation from the same population as our training data

We are interested in a model $\hat{f}$ such that $\hat{f}(x_0)$ is close to y_0.

Too complex models tend to <u>overfit</u> the training data, not doing well outside of the training data.

<u>Mathematically</u>:

$$E_{x,y}\left[\left(y_0 - \hat{f}(x_0)\right)^2\right]$$

EXPECTED ERROR UNDER REPEATED TRAINING SETS

$$E\left[(Y_0 - \hat{f}(X_0))^2\right]$$

$$= E\left[\left(f(X_0) + \varepsilon - \hat{f}(X_0)\right)^2\right]$$

$$= E\left[\left(f(X_0) + \varepsilon - \hat{f}(X_0) + E(\hat{f}(X_0)) - E(\hat{f}(X_0))\right)^2\right]$$

$\pm E(\hat{f}(X_0))$ constant

$E[Z^2] - E[Z]^2$

$= E[(Z - E(Z))^2]$

$E[(\hat{f} - E(\hat{f}))^2]$

$$= E\left(\left(f(X_0) - E(\hat{f}(X_0))\right)^2\right) + E(\varepsilon^2) + E\left[\left(E(\hat{f}(X_0)) - \hat{f}(X_0)\right)^2\right]$$

$$+ 2\left[\left(f(X_0) - E(\hat{f}(X_0))\right)\varepsilon\right] \leftarrow \left(f(X_0) - E(\hat{f}(X_0))\right)\underbrace{E[\varepsilon]}_{0} = 0$$

$$+ 2\left[\left(E(\hat{f}(X_0)) - \hat{f}(X_0)\right)\varepsilon\right] \leftarrow \varepsilon \perp X \quad E\left(E(\hat{f}(X_0)) - \hat{f}(X_0)\right)\underbrace{E[\varepsilon]}_{0} = 0$$

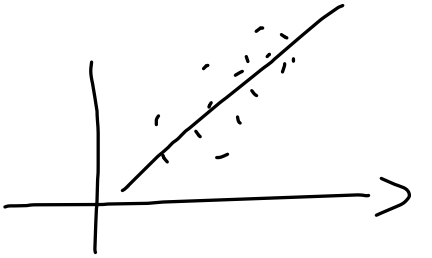
$$+ 2\left[\underbrace{\left(f(X_0) - E(\hat{f}(X_0))\right)}_{\text{constant}} \cdot \underbrace{\left(E(\hat{f}(X_0)) - \hat{f}(X_0)\right)}_{E[E(\hat{f}) - \hat{f}] = E(\hat{f}) - E(\hat{f}) = 0}\right]$$

$$= \underbrace{\left[f(X_0) - E(\hat{f}(X_0))\right]^2}_{\text{bias}} + \underbrace{\text{Var}(\varepsilon)}_{\text{irreducible error}} + \underbrace{\text{Var}(\hat{f}(X_0))}_{\text{variance}}$$

BIAS-VARIANCE TRADE OFF

Simple model — high bias
              — low variance

Complex model — low bias
               — high variance



Good model : good trade-off between bias and variance

→ Cross-validation of error

Supervised vs Unsupervised Learning

Supervised : (x, y) training data

Unsupervised : X observed

Supervised Methods :

① Regression methods $\Longleftrightarrow$ quantitative response (eg wage)

② Classification methods $\Longleftrightarrow$ qualitative response (categorical)

Objective : Given (x_i, y_i) training data, find a rule

$\hat{f}$ < regression function ①

$\hat{f}$ < classifier ②

that allows to predict y from X.