STK-IN4300 Compendium

Gabriel Sigurd Cabrera

(Dated: Wednesday 27$^{\text{th}}$ November, 2019)

I. OVERVIEW OF TOPICS

A. Lecture 1

1. Basics

Typical Scenario:

An *outcome* Y (*dependent variable*, *response*) can be *categorical* or *qualitative*.

We want to predict this outcome based on a set of *features* $X_1, X_2, ..., X_p$ (*independent variables*, *predictors*).

In practice, we have a *training set* that is used to create a *learner* (or model/rule $f(X_i) \approx Y_i$.)

A *supervised learning problem* is when the outcome is measured in the training data, and can be used to construct a learner Y.

B. Least Squares Estimate

Given a training set $\{(x_{i1}, x_{i2}, ..., x_{ip}, y_i\}$ a *least-squares* model is given by:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i$$

With a least-square estimate:

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

Where $X^{\mathrm{T}} X$ is known as the *Gramian*.

C. Invertability

If $X^{\mathrm{T}} X$ is *not invertible*, then we can use *dimension reduction* or *shrinkage methods*.

Some dimension reduction methods are to:

- Remove variables with *low correlation* (forward selection/back substitution)
- More formal subset selection
- Selecting optimal linear combinations of variables (*principal component analysis*.)

Some shrinkage methods are:

- *Ridge regression*
- *LASSO*
- *Elastic net*

D. Conventions

Quantitative response: *Regression*

Qualitative response: *Classification*

E. Least-Squares

For *ordinary least-squares* (OLS), we estimate β by minimizing the *residual sum of squares* (RSS):

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

Where $X \in \mathbb{R}^{N \times p}$, $X \in \mathbb{R}^{N}$.

F. K-Nearest-Neighbors

The *k-nearest-neighbors* (KNN) of x is the mean:

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

G. Other Methods

OLS and KNN are the basis of most modern techniques; some of these are:

- *Kernel methods* that weigh data according to distance
- In higher dimensions, weighing variables based on correlation
- Local regression models
- Linear models of functions of X
- *Projection pursuit* and *neural network*

H. Statistical Decision Theory

Statistical decision theory gives a *mathematical framework* for finding the optimal learner.

Given $X \in \mathbb{R}^p$, $Y \in \mathbb{R}$ and a *joint distribution* $p(X,Y)$, our goal is to find a function $f(X)$ for predicting Y given X.

This requires a *loss function* $L(Y, f(X))$ for penalizing errors in $f(X)$ when the truth is Y.

An example is *squared error loss*:

$$L(Y, f(X) = (Y - f(X))^2$$

The *expected prediction error* of $f(X)$ is given by:

$$\mathrm{EPE}(f) = E_{X,Y}[L(Y, f(X)] = \int_{x,y} L(y, f(x))p(x,y)\ \mathrm{d}x\ \mathrm{d}y$$

Next, we must find the f that minimizes $\mathrm{EPE}(f)$.

For the *squared error loss* $L(Y, f(X) = (Y - f(X))^2$, we have:

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

It is sufficient to minimize $E_{Y|X}[(Y - f(X))^2|X]$:

$$f(x) = \mathrm{argmin}_c E_{Y|X}[(Y-c)^2|X = x] = E[Y|X = x]$$

This is known as the *conditional expectation*, or the *regression function.* This implies that the best prediction of Y at any point $X = x$ is the *conditional mean.*

I. Error Decomposition

$$E[(Y - \hat{f}(X))^2] = \underbrace{\sigma^2}_{\text{irreducible error}} + \underbrace{\underbrace{\mathrm{Var}(\hat{f}(X))}_{\text{variance}} + \underbrace{E[\hat{f}(X) - f(X)]^2}_{\text{bias}^2}}_{MSE}$$

J. Assumptions for OLS

- A function is linear in its arguments; $f(x) \approx x^{\mathrm{T}}\beta$.
- $\mathrm{argmin}_\beta E[(Y - X^{\mathrm{T}}\beta)^2|X = x] \rightarrow \beta = E[XX^{\mathrm{T}}]^{-1}E[XY]$.
- Replacing the expectations by averages over the training data leads to $\hat{\beta}$.

K. Assumptions for KNN

- Uses $f(x) = E[Y|X = x]$ directly.
- $\hat{f}(x_i) = \mathrm{mean}(y_i)$ for observed x_i.
- Normally, there is at most one observation for each point x_i.
- Uses points in the neighborhood:

$$\hat{f}(x) = \mathrm{mean}(y_i|x_i \in N_k(x))$$

- There are two approximations:

  - *Expectation* is approximated by averaging over sample data.

  - *Conditioning* on a point is related to conditioning on a neighborhood.

- $f(x)$ can be approximated by a *locally constant function.*
- For $N \to \infty$, all $x_i \in N_k(x) \approx x$.
- For $k \to \infty$, $\hat{f}(x)$ is getting more stable.
- Under mild regularity conditions on $p(X,Y)$:

  $\hat{f}(x) \to E[Y|X = x]$ for $N, k \to \infty$ s.t. $k/N \to 0$

- It is unnecessary to implement the *squared loss error* function (L_2 loss function.)
- A valid alternative is the L_1 loss function, whose solution is the conditional median:

$$\hat{f}(x) = \mathrm{median}(Y|X = x) \tag{1}$$

- More robust estimates than those obtained with conditional mean.
- The L_1 loss function has discontinuities in its derivatives which leads to numerical difficulties.

L. Conclusion

OLS: stable but biased

KNN: less biased and less stable

For higher dimensions, KNN suffers from the *curse of dimensionality*

II. LECTURE 2

A. Gauss-Markov Theorem

The least square estimator $\hat{\theta} = a^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$ is the:

B est smallest error (MSE)
L inear $\hat{\theta} = a^{\mathrm{T}}\beta$
U nbiased $E[\hat{\theta}] = \theta$
E stimator

Given the *error decomposition*, then any estimator $\tilde{\theta} = c^{\mathrm{T}}Y$ s.t. $E[c^{\mathrm{T}}Y] = a^{\mathrm{T}}\hat{\beta}$ has $\mathrm{Var}(c^{\mathrm{T}}Y) \geq \mathrm{Var}(a^{\mathrm{T}}\hat{\beta})$.

B. Hypothesis Testing

To test $H_0 : \beta_j = 0$ we use the *Z-score statistic*:

$$z_j = \frac{\hat{\beta}_j - 0}{sd(\beta_j)} = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{(X^{\mathrm{T}}X)^{-1}_{[j,j]}}} \tag{2}$$

When σ^2 is unknown, under H_0

$$z_j \sim t_{\mathrm{N}-p-1} \tag{3}$$

References