CH2_Notes

August 8, 2019

1 Statistical Learning

1.1 What is Statistical Learning?

Given paired data (X, Y), assume a relationship between X and Y modeled by

$$Y = f(X) + \epsilon$$

where $f : \mathbb{R}^p \to \mathbb{R}$ is a function and ϵ is a random error term with $\mathbb{E}(\epsilon) = 0$.
Statistical learning is a set of approaches for estimating $f0$

1.1.1 Why Estimate f?

Prediction

- We may want to ***predict*** the output Y from an estimate $\hat{f}$ of f. The predicted value for a given Y is then
$$\hat{Y} = \hat{f}(X)$$
. In prediction, we often treat f as a ***black-box***
- The mean squared-error2 $\mathbf{mse}(\hat{Y}) = \mathbb{E}(Y - \hat{Y})^2$ is a good measure of the accuracy of $\hat{Y}$ as a predictor for Y.
- One can write

$$\mathbf{mse}(\hat{Y}) = \left(f(X) - \hat{f}(X)\right)^2 + \mathbb{V}(\epsilon)$$

These two terms are known as the ***reducible error*** and ***irreducible error***, respectively3

Inference

- Instead of predicting Y from X, we may be more interested how Y changes as a function of X. In inference, we usually do not treat f as a black box.

Examples of important inference questions:

- *Which predictors have the largest influence on the response?*
- *What is the relationship between the response and each predictor?*
- *Is f linear or non-linear?

1.1.2 How to Estimate f?

Parametric methods Steps for parametric method:

1. Assume a parametric model for f, that is assume a specific functional form4

$$f = f(X, \boldsymbol{\beta})$$

for some vector of ***parameters*** $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$

2. Use the training data to ***fit*** or ***train*** the model, that is to choose β_i such that

$$Y \approx f(X, \boldsymbol{\beta})$$

Non-parametric methods These methods make no assumptions about the functional form of f.

1.1.3 Accuracy vs. Interpretability

- In inference, generally speaking the more flexible the method, the less interpretable.
- In prediction, generally speaking the more flexible the method, the less accurate

1.1.4 Supervised vs. Unsupervised Learning

- In ***supervised learning***, training data consists of pairs (X, Y) where X is a vector of predictors and Y a response. Prediction and inference are supervised learning problems, and the response variable (or the relationship between the response and the predictors) *supervises* the analysis of model
- In ***unsupervised learning***, training data lacks a response variable.

1.1.5 Regression vs. Classification

- Problems with a quantitative response ($Y \in S \subseteq \mathbb{R}$) tend to be called ***regression*** problems
- Problems with a qualitative, or categorical response ($Y \in \{y_1, \dots, y_n\}$) tend to be called ***classification*** problems

1.2 Assessing Model Accuracy

There is no free lunch in statistics

1.2.1 Measuring Quality of Fit

- To evaluate the performance of a method on a data set, we need measure model accuracy (how well predictions match observed data).
- In regression, the most common measure is the ***mean-squared error***

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

where y_i and $\hat{f}(x_i)$ are the i true and predicting responses, respectively.

- We are usually not interested in minimizing MSE with respect to training data but rather to test data.
- There is no guarantee low training MSE will translate to low test MSE.
- Having low training MSE but high test MSE is called ***overfitting***

1.2.2 The Bias-Variance Tradeoff

- For a given x_0, the expected 5 MSE can be written

$$\begin{aligned}\mathbb{E}\left[\left(y_0 - \hat{f}(x_0)\right)^2\right] &= \left(\mathbb{E}\left[\hat{f}(x)\right] - f(x)\right)^2 + \mathbb{E}\left[\left(\hat{f}(x_0) - \mathbb{E}\left[\hat{f}(x_0)\right]\right)^2\right] + \mathbb{E}\left[(\epsilon - \mathbb{E}[\epsilon])^2\right] \\ &= \mathbf{bias}^2\left(\hat{f}(x_0))\right) + \mathbb{V}\left(\hat{f}(x_0)\right) + \mathbb{V}(\epsilon)\end{aligned}$$

- A good method minimizes variance and bias simultaneously.
- As a general rule, these quantities are inversely proportional. More flexible methods have lower bias but higher variance, while less flexible methods have the opposite. This is the ***bias-variance tradeoff***
- In practice the mse, variance and bias cannot be calculated exactly but one must keep the bias-variance tradeoff in mind.

1.2.3 The Classification Setting

- In the classification setting, the most common measure of model accuracy is the ***error rate*** 6

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$

- As with the regression, we are interested in minimizing the test error rate, not the training error rate.

The Bayes Classifier

- Given K classes, the ***Bayes Classifier*** predicts

$$\hat{y_0} = \operatorname*{argmax}_{1 \leqslant j \leqslant K} \mathbb{P}\left(Y = j \mid X = x_0\right)$$

- The set of points

$$\{x_0 \in \mathbb{R}^p \mid \mathbb{P}\left(Y = j \mid X = x_0\right) = \mathbb{P}\left(Y = k \mid X = x_0\right) \text{ for all } 1 \leqslant j, k \leqslant K\}$$

  is called the ***Bayes decision boundary***

- The test error rate of the Bayes classifier is the ***Bayes error rate***, which is minimal among classifiers. It is given by

$$1 - \mathbb{E}\left(\max_j \mathbb{P}\left(Y = j \mid X\right)\right)$$

- The Bayes classifier is optimal, but in practice we don't know $\mathbb{P}\left(Y \mid X\right)$.

K-Nearest Neighbors

- The ***K-nearest neighbors*** classifier works by estimating $\mathbb{P}\left(Y \mid X\right)$ as follows.

1. Given $K \geqslant 1$ and x_0, find the set of points

$$\mathcal{N}_0 = \{K \text{ nearest points to } x_0\} \subseteq \mathbb{R}^p$$

2. For each class j set

$$\mathbb{P}\left(Y = j \mid X\right) = \frac{1}{K} \sum_{x_i \in \mathcal{N}_0} I(y_i = j)$$

3. Predict

$$\hat{y_0} = \operatorname*{argmax}_{1 \leqslant j \leqslant K} \mathbb{P}\left(Y = j \mid X = x_0\right)$$

1.3 Footnotes

0. Reading the rest of the chapter, one realized this is the situation for *supervised* learning, which is the vast majority of this book is concerned with.

1. Here $X = (X_1, \dots, X_p)^T$ is a vector.

2. This is usual definition of the mean squared-error of $\hat{Y}$ as an estimator of the (non-parametric) quantity $Y = f(X)$.

3. We can in principle control the reducible error by improving the estimate $\hat{f}$, but we cannot control the irreducible error.

4. For example, a simple but popular assumption is that f is linear in both the parameters and the features, that is:

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

This is linear regression.

5. Here the random variable is $\hat{f}(x_0)$, so the average is taken over all data sets

6. This is just the proportion of misclassified observations.