Chapter 2 – Statistical Learning

Zhe Rao

1. What is Statistical Learning?

- We assume that there is some relationship between Y and X

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

  - f represents the *systematic* information that X provides about Y
    - However, the functional form is unknown
    - We are interested in estimating it
  - In essence, **statistical learning** refers to a set of approaches for estimating f

1.1. Why estimating f?

Prediction

- In many situations, a set of inputs X are readily available, but the output Y cannot be easily obtained
  - Can treat the model as *black-box*

- The accuracy of $\hat{Y}$ depends on two quantities
  - *Reducible error*: that can be potentially improved by using the most appropriate statistical learning technique
  - *Irreducible error*: ϵ
    - It may contain unmeasured variables that are relevant in predicting Y
    - Or unmeasurable variables

Inference

- Understanding the association between Y and X
  - Models should not be a *black-box*

- Depending on whether ultimate goal is prediction, inference, or combination of two, different methods may be appropriate

- ➢ Linear models → simple and interpretable inference, may not yield as accurate predictions
- ➢ Highly non-linear approaches → quite accurate predictions for Y, but less interpretable for inference

1.2. *How to estimate f?*

Parametric Methods

- Involves two-step model-based approach
  - ➢ First, assumption about the functional form
    - ✧ E.g. f is linear in X
  - ➢ Second, a procedure that uses the training data to fit or train the model
    - ✧ The most common approach is *(ordinary) least squares*
    - ✧ There are many possible ways

- The potential disadvantage of a parametric approach is model will not match the true unknown form of f
  - ➢ Address this problem by choosing *flexible* models that can fit many possible functional forms for f
    - ✧ Requires estimating a greater number of parameters → overfitting

Non-Parametric Methods

- Do not make explicit assumptions about the functional form of f
  - ➢ Major disadvantage: requires a very large number of observations in order to obtain an accurate estimate for f
  - ➢ E.g. *thin-plate spline*.
  - ➢ Problem: overfitting

1.3. *Trade-off Between Prediction Accuracy and Model Interpretability*

- We might prefer a more restrictive model because they are much more interpretable
  - ➢ Highly non-linear models could lead to complicated estimates of f that it is difficult to understand how nay individual predictor is associated with the response

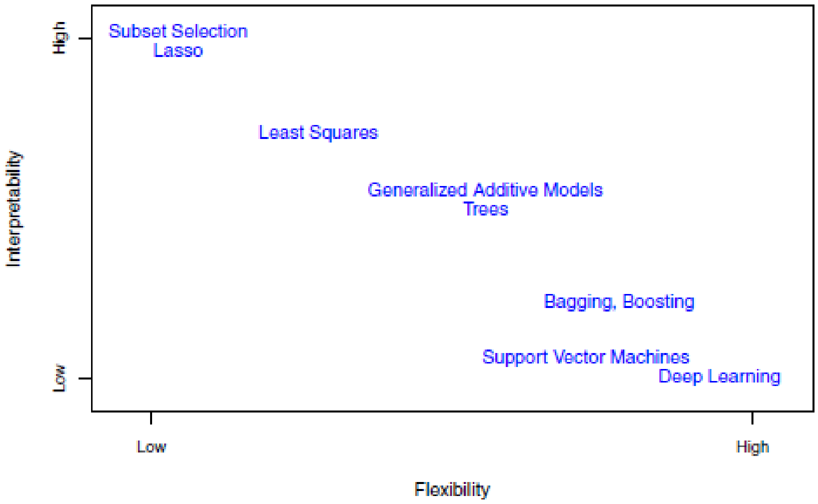


FIGURE 2.7. *A representation of the tradeoff between flexibility and interpretability, using different statistical learning methods. In general, as the flexibility of a method increases, its interpretability decreases.*

1.4. Supervised VS. Unsupervised Learning

- Unsupervised learning describes the situation where for every observation, we observe a vector of measurement x_i but not associated response y_i
  - ➢ What sort of statistical analysis is possible? We can seek to understand the relationship between the variables or between the observations
  - ➢ One statistical learning tool that we may use in this setting is *cluster analysis*, or clustering
    - ✧ The goal of clustering is to ascertain whether the observations fall into relatively distinct groups

- In reality, suppose that we have a set of n observations, m of which have both predictors and responses, whereas the rest have only predictors
  - ➢ It may be the case that predictors can be measured relatively cheaply but responses are much more expensive to collect
  - ➢ This setting is referred to as *semi-supervised learning*

1.5. Regression VS. Classification

- Variables can be characterized as quantitative or qualitative

- Quantitative variables take on numerical values
    - Usually refer to problems with a quantitative response as *regression* problems
- Qualitative variables take on values in one of K different classes, or categories
    - Usually refer to problems with qualitative responses as *classification* problems

- We tend to select statistical learning methods on the basis of whether the response is quantitative or qualitative
    - But this is generally considered less important
    - Most of the statistical learning methods can be applied regardless of the predictor variable type

2. *Assessing Model Accuracy*

- There is no one method dominates all others over all possible dataset
    - On a particular data set, one specific method may work best, but some other method may work better on a similar but different data set
    - Selecting the best approach can be one of the most challenging parts of performing statistical learning in practice

2.1. *Measuring the Quality of Fit*

- We need to measure how well the models' predictions match the truth
    - In the regression setting, the most commonly-used measure is the MSE

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

- It is important to note that we are interested in the accuracy of test predictions rather than training predictions
    - We want to choose the method that gives the lowest test MSE

- When a given method yields a small training MSE but a large test MSE, it's overfitting the data
    - This happens because our statistical learning procedure is working too hard to find patterns in the training data, and may be picking up some patterns that are just caused by random chance rather than by true properties of the unknown function f

2.2. The Bias- Variance Trade-Off

$$E\left(y_0-\hat{f}(x_0)\right)^2 = Var\left(\hat{f}(x_0)\right)+\left[Bias\left(\hat{f}(x_0)\right)\right]^2+Var(\epsilon)\geq Var(\epsilon)$$

- ➢ $E\left(y_0-\hat{f}(x_0)\right)^2$ defines the expected test MSE at x_0,
  - ✧ If we repeatedly estimate f using a large number of training sets, this is the average test MSE
  - ✧ The overall expected test MSE can be computed by averaging $E\left(y_0-\hat{f}(x_0)\right)^2$ over all possible values of x_0 in the test set

- Variance refers to the amount by which $\hat{f}$ would change if we estimated it using a different training data set
  - ➢ Ideally the estimate for f should not vary too much between training sets
  - ➢ If a method has high variance then small changes in the training data can result in large changes in $\hat{f}$
  - ➢ More flexible statistical methods have higher variance

- Bias refers to the error that is introduced by approximating a real-life problem
  - ➢ For example, if assumed linear regression, the relationship between Y and X is assumed to be linear, but in real life, no relationship is hardly linear, so there is the bias in estimation of f

- As a general rule, as we use more flexible methods, the variance will increase and the bias will decrease
  - ➢ As we increase the flexibility of a class of methods, the bias tends to initially decrease faster than the variance increases → test MSE decreases
  - ➢ At some point increasing flexibility has little impact on the bias but starts to significantly increase the variance → test MSE increases

2.3. The Classification Setting

- The most common approach for quantifying the accuracy of our estimate $\hat{f}$ is the training *error rate*:

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

The Bayes Classifier

- The test error rate is minimized, on average, by assign each observation to the most likely class, given its predictor values – *Bayes Classifier*

$$P(Y = j|X = x_0)$$

  - Bayes classifier corresponds on predicting class one if $P(Y = 1|X = x_0) > 0.5$, and class two otherwise
  - There the probability is exactly 50% is called the *Bayes decision boundary*.
  - Bayes error rate

$$1 - E\left(\max_j P(Y = j|X)\right)$$

    - The Bayes error rate is analogous to the irreducible error.

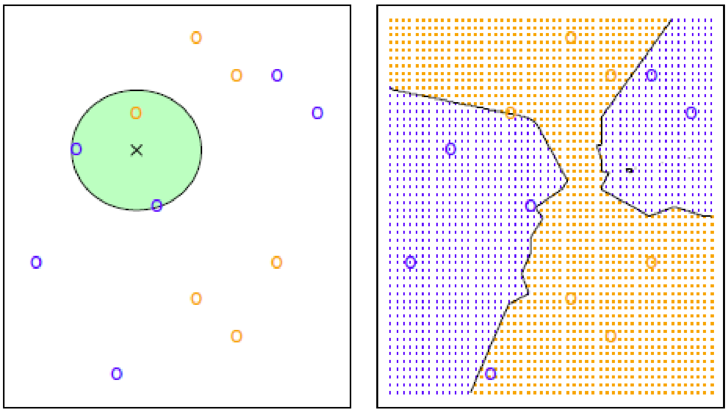
K – Nearest Neighbors

- For real data, we do not know the conditional distribution of Y given X, so computing the Bayes classifier is impossible
  - Many approaches attempt to estimate the conditional distribution of Y given X, and then classify a given observation to the class with highest *estimated* probability

- K-nearest neighbors (KNN)
  - Given a positive integer K and a test observation x_0, the KNN classifier first identifies the K points in the training data that are closest to x_0, represented by N_0.
  - It then estimates the conditional probability of class j as the fraction of points in N_0 whose response values equal j

$$P(Y = j|X = x_0) = \frac{1}{K}\sum_{i \in N_0} I(y_i = j)$$

    - KNN classified the test observation x_0 to the class with the largest probability form



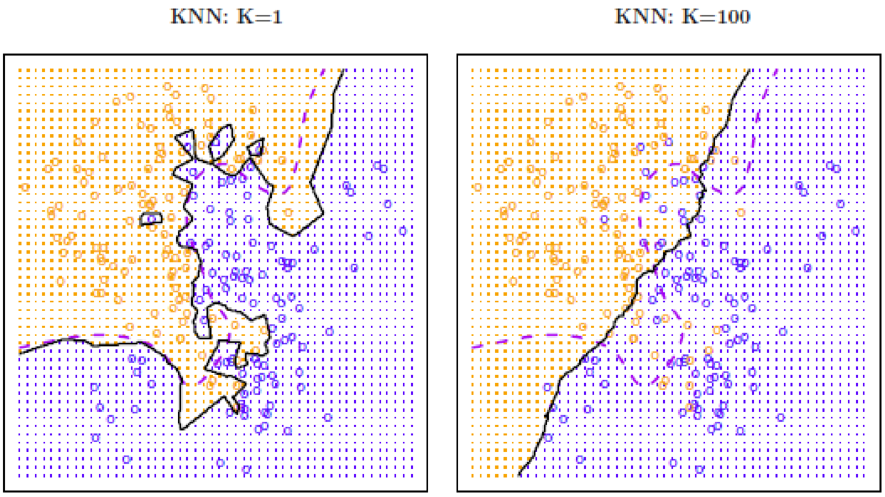
  - On the left-hand panel

- ✧ We want to estimate the class of the black cross
- ✧ We choose K = 3, the estimated probability is 2/3 for blue class and 1/3 for orange class
- ✧ Hence, KNN will predict that the black cross belongs to the blue class

➢ On the right-hand panel, K = 3 KNN is being applied to all of the possible values for X_1 & X_2

- ✧ The corresponding KNN decision boundary is drawn

- Despite the fact that it is a very simple approach, KNN can often produce classifiers that are surprisingly close to the optimal Bayes classifier.



➢ The choice of K has a drastic effect on the KNN classifier obtained

- ✧ When K = 1, the decision boundary is overly flexible and finds patterns in the data that don't correspond to the Bayes decision boundary – low bias and high variance
- ✧ When K = 100, the method becomes less flexible and produces a decision boundary close to linear – low variance and high bias