Notes on An Introduction to Statistical Learning

Jonathan Chen

May 13, 2025

Contents

1 Statistical learning

- input variable X
- output variable Y
- Y can be modeled using this equation:

$$Y = f(X) + \epsilon \tag{1}$$

- equation 1: f is the fixed but unknown function of X; ϵ is the random error term that is independent of X and has mean 0
- f is unknown, the goal estimate f based on observed points using statistical learning – a set of approaches for estimating f

1.1 Why Estimate f?

$$\hat{Y} = \hat{f}(X) \tag{2}$$

- equation 2: $\hat{Y}$ is the predicted Y; $\hat{f}$ is the estimated f
- accuracy of $\hat{Y}$ depends on reducible error and irreducible error.
  - reducible error is error that can be reduced using statistical learning techniques
  - irreducible error (ϵ) is error that cannot be reduced because there maybe unmeasured variables or variations

$$E(Y - \hat{Y})^2 = [f(X) - \hat{f}(X)]^2 + Var(\epsilon) \tag{3}$$

$$E(Y - \hat{Y}) = \text{reducible} + \text{irreducible} \tag{4}$$

- equation 3-4: average expected value of the squared distance between the ground truth and the predictions is equal to reducible error + irreducible error
- in the real world, irreducible error is almost always unknown
- two reasons to estimate f
  - prediction
    - $\hat{f}$ can be entirely black-box. We don't care about the exact form of f, we just care about getting good predictions $\hat{f}$
  - inference
    - the goal of inference is to understand the association between Y and X
    - $\hat{f}$ is not black box, we estimate f but the goal is not to make the most accurate predictions for Y

    * inference is used for some main questions:
        · which predictors are associated with the response?
        · what is the relationship between response and each predictor?
        · can relationship between Y and each predictor be summarized using linear equation?

1.2 How to Estimate f?

- we have n training data
- x_{ij}
  - i is the observation number and j is the predictor/variable number
- y_i: i-th training sample
- statistical learning techniques can be split into parametric and non-parametric methods
  - parametric
    * make assumption about the functional form of f (simplify the problem)
    * advantage: less number of observations needed to accurately fit a model if the shape chosen is correct
    * disadvantage: if the shape chosen is incorrect, estimate can be poor
  - non-parametric
    * no assumption about the shape
    * advantage: can accurately fit wider range of possible shapes for f
    * disadvantage: larger number of observations needed

1.3 Prediction / Interpretability Tradeoff

- model flexibility: how many shapes a statistical learning technique can use to estimate f
  - inflexible: linear regression
  - flexible: neural network
- inflexible methods are much more interpretable while flexible methods can fit more shapes
  - for inference: use an inflexible / interpretable method
  - for prediction: use a flexible / uninterpretable method
- NOTE: you cannot just use the most flexible model possible for prediction because often times it will overfit and be less accurate than some of the less flexible models

1.4 Supervised v. Unsupervised

- supervised: for each x there exists a y
- unsupervised: have x, but no y
- semi-supervised: have x, but only have some of the ys

1.5 Regression v. Classification

- quantitative variables: numbers; qualitative variables: categories
- quantitative response: regression; qualitative response: classification
- distinctions can be blurred
  - least squares linear regression is used for regression
  - logistic regression is used for classification
  - KNN, boosting can be used for both regression and classification
- whether response is quantitative or qualitative matters more than predictors because qualitative predictors can be coded to be quantitative

1.6 Assessing Model Accuracy

- there are no free lunches in statistics; no one method dominates over all possible datasets
- model accuracy is used to select the best model for a dataset

1.7 Measuring Quality of Fit

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2 \tag{5}$$

- equation 5: mean squared error formula that is very popular for assessing regression accuracy. Basically take the square difference between the truth and the prediction for each observation / prediction. Then, add them all up and divide by number of samples
- when assessing accuracy of a model, we are interested in accuracy on UNSEEN data. Choose the method that gives the lowest TEST error, not TRAINING error
- there is NO guarantee that the method with the lowest training error will also have the lowest test error due to overfitting
- you increase the degrees of freedom of a model (flexibility of a model), test MSE will start declining then start increasing while training MSE continues decreasing

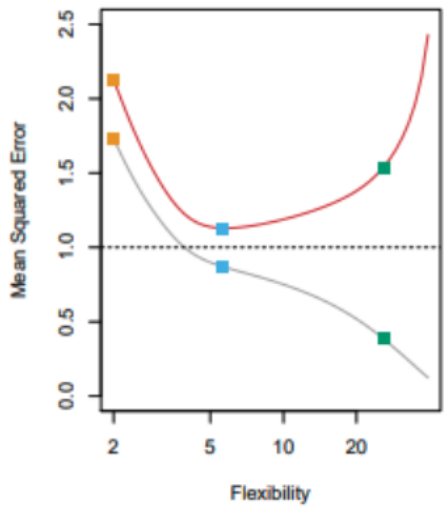

Figure 1: Red line is test MSE. Gray line is train MSE. Train MSE will decrease continuously as flexibility of a model increases. Test MSE will decrease at first but then increase as model starts overfitting on the training data.

- when you overfit, you see low training MSE and high testing MSE
  - the model picks up randomness in the data instead of the important patterns
- however, in all models, it can be expected that the training MSE ¡ test MSE because the model aims to reduce the training MSE
  - hence, overfitting refers specifically to when a less flexible model would've yielded a smaller test MSE

1.8 Bias / Variance tradeoff

$$E(y_0 - \hat{f}(x_0))^2 = Var(\hat{f}(x_0) + [Bias(\hat{f}(x_0)]^2 + Var(\epsilon) \tag{6}$$

- equation 6: the average test MSE if we were to repeatedly estimate f using different training sets is equal to variance + bias squared + irreducible error
  - variance is the amount by which $\hat{f}$ would change if we estimated it using a different training dataset
    - flexible methods have higher variance
  - bias is the error that comes from approximating a real-life problem which is most likely more complex than our model + data
    - flexible methods have lower bias

- as you increase flexibility, variance increases and bias decreases. Bias initially decreases faster than variance but at some point variance will continue increasing while bias stops decreasing
- getting good test set performance requires low variance as well as bias. Easy to have one but not the other

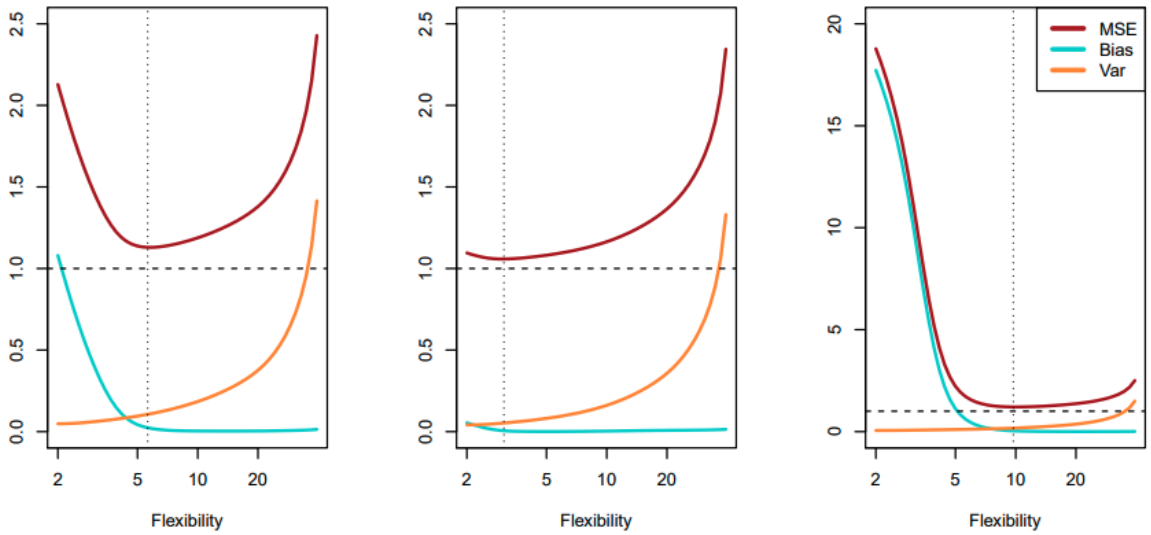

Figure 2: Squared bias (blue curve), variance (orange curve), Var(ϵ) (dashed line), and test MSE (red curve) for three data sets. The lowest test MSE at the point where the squared bias and the variance is low

1.9 Classification Setting

$$\text{error rate} = \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i) \tag{7}$$

- equation 7: the error rate is the most common error for classification. It is the proportion of mistakes the model makes. I is an identity function that returns 1 if the condition is true and 0 if the condition is false
- same train / test split applies to classification. You want to make decisions based on the smallest test error instead of training error
- bayes classifier
  - the test error rate is minimized, on average, by a very simple classifier that assigns each observation to the most likely class

$$Pr(Y = j|X = x_0) \tag{8}$$

- equation 8: j is the class. Assign class j to observation x such that the probability of j being the label for x is largest out of all possible classes
- this classifier is called the Bayes Classifier

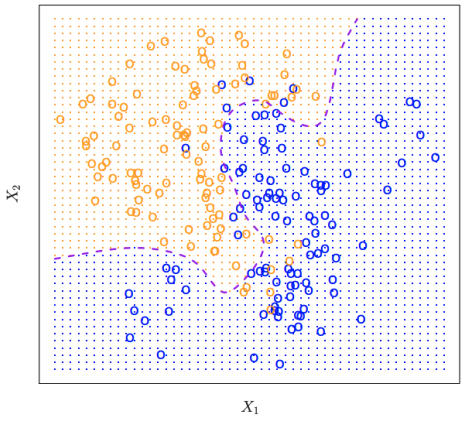

Figure 3: The Bayes classifier will form a Bayes Decision Boundary. The Bayes classifier's prediction depends on the section that the observation falls on

- The Bayes error rate is the test error rate produced by Bayes classifier. It should be the lowest possible test error rate

$$1 - max_j Pr(Y = j|X = x_0) \tag{9}$$
$$1 - E(max_j Pr(Y = j|X) \tag{10}$$

- equation 9 is the Bayes error rate for a single observation. Equation 10 is the Bayes error rate for a given dataset
- since Bayes error rate is the best possible test error rate, it is thus equivalent to the irreducible error
  - in real data, we do not know the conditional distribution of Y given X, so computing the Bayes classifier is impossible – it is just a gold standard to compare other methods
  - many approaches try to estimate conditional distribution of Y given X and then classify given observation to the class with the highest probability

- K-Nearest Neighbors is an approach that that tries to estimate the conditional distribution of Y given X
  - given an int k and x_0 in the test set, KNN identifies the k points in the training data closest to x_0 and then estimates conditional probability for class j as a fraction of points in N_0 that is j. It classifies the observation as the class that has the highest conditional probability

    - basically looks at the surrounding K points. Whatever class is the highest in those K points is the predicted class
- KNN is simple but can produce results surprisingly close to Bayes classifier
- low K makes the model flexible and prone to overfitting
- high K makes the model inflexible and prone to underfitting

2 Linear Regression

- purpose: useful for predicting quantitative response
- dull but effective, and many techniques build off of linear regression
- used to investigate several questions:
  - is there a relationship between variables / output?
  - how strong is the relationship?
  - which variables are connected to the output?
  - how accurate can we predict the target?
  - is there a synergy among the variables?

$$Y \approx \beta_0 + \beta_1 X \tag{11}$$

- equation 11: Y is the target; β_0 is the intercept; β_1 is the slope; X is the variables
- we want to estimate the βs (or coefficients) using training data so that we can predict the target for future data

2.1 Estimating the Coefficients

- β_0 and β_1 are unknown, so we need to estimate them: $\hat{\beta}_0$, $\hat{\beta}_1$
- we find prediction of coefficients using least squares

$$e_i = y_i - \hat{y}_i \tag{12}$$

- equation 12: ith residual is the difference between the ith observed value and ith actual value

$$RSS = e_1^2 + e_2^2 + \cdots + e_n^2 \tag{13}$$

- equation 13: residual sum of squares is the squared residuals for all observations summed together

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x - \bar{x})^2} \tag{14}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{15}$$

- equation 14: the $\bar{\beta}_1$ and $\bar{\beta}_0$ estimates that minimize the RSS can be solved using these formulas. $\bar{x}$ and $\bar{y}$ are the sample means of x and y respectively. The coefficients estimated this way are called least squares coefficient estimates

References

[1] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*. Springer, 2013.