SHEET - An Introduction to Statistical Learning
Chapter 2 - Statistical Learning

PLAYE Nicolas

1 December 2023

1 Statistical Learning

1.1 Courses' Demonstrations

We suppose that we observe a quantitative response Y and p different predictors, $X_1, X_2, \ldots, X_p$. We assume that there is some relationship between Y and $X = (X_1, X_2, \ldots, X_p)$, which can be written in the very general form

$$Y = f(X) + \varepsilon.$$

Here f is some fixed but unknown function of $X_1, \ldots, X_p$, and ε is a random error term, which is independent of X and has mean zero. In this formulation, f represents the systematic information that X provides about Y.

Consider a given estimate $\hat{f}$ and a set of predictors X, which yields the prediction $\hat{Y} = \hat{f}(X)$. Assume for a moment that both $\hat{f}$ and X are fixed. Then, we show that

$$\begin{align*}
E(Y - \hat{Y})^2 &= E[f(X) + \varepsilon - \hat{f}(X)]^2 \\
&= [f(X) - \hat{f}(X)]^2 + \mathrm{Var}(\varepsilon) \\
&= [f(X) - \hat{f}(X)]^2 + \mathrm{Var}(\varepsilon) \tag{2.3} \\
&= \mathrm{Var}(\hat{f}) + \mathrm{Bias}(\hat{f})^2 + \mathrm{Var}(\varepsilon) \tag{2.7}
\end{align*}$$

Demonstration:

$$\begin{aligned}
\mathrm{Var}[X] &= E[X^2] - E[X]^2 \\
E[X^2] &= \mathrm{Var}[X] + E[X]^2 \\
E[f] &= f \\
y &= f + \varepsilon \\
E[\varepsilon] &= 0 \\
E[y] &= E[f + \varepsilon] = E[f] = f \\
\mathrm{Var}[y] &= E[(y - E[y])^2] = E[(y - f)^2] \\
E[(y - \hat{f})^2] &= E[y^2 + \hat{f}^2 - 2y\hat{f}] \\
&= E[y^2] + E[\hat{f}^2] - E[2y\hat{f}] \\
&= \mathrm{Var}[y] + E[y]^2 + \mathrm{Var}[\hat{f}] + E[\hat{f}]^2 - 2fE[\hat{f}] \\
&= \mathrm{Var}[y] + \mathrm{Var}[\hat{f}] + (f - E[\hat{f}])^2 \\
&= \mathrm{Var}[y] + \mathrm{Var}[\hat{f}] + E[(f - \hat{f})]^2 \\
&= \mathrm{Var}[\varepsilon] + \mathrm{Var}[\hat{f}] + \mathrm{Bias}[\hat{f}]^2
\end{aligned}$$

1.2 Answers of Exercises

Question 1: For each of parts (a)-(d), indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method. Justify the answer.

Definition:

Flexible models and **Inflexible models** generally refer to how adaptable or rigid a model is in learning from data:
- **Flexible Model**: has Adaptability (capable of adapting to a wide range of data inputs and variations in data patterns), Complexity (have a higher level of complexity), Over-fitting Risk (more prone to over-fitting).
- **Inflexible Model**: has Limited Adaptability (more rigid structure and are less adaptive to a wide range of data inputs, Simplicity (generally simpler, with fewer parameters or less complexity), Generalizable (often better at generalizing from the training data to new, unseen data, as they are less likely to overfit.

(a) The sample size n is extremely large, and the number of predictors p is small

- Flexible methods are great at capturing complex relationships in the data. With a large sample size, they have enough data to ”learn” these complexities without over-fitting, allowing them to potentially perform better by capturing more nuances in the data.
- Inflexible methods are less likely to over-fit since they don't model complex relationships as aggressively. If the true relationship between predictors and response is simple, these methods might perform better

In general, opting for flexible models in the case of a sample size n extremely large, and a number of predictors p small is a preferred option.

(b) The number of predictors p is extremely large, and the number of observations n is small

When p is large and n is small, the choice generally leans towards inflexible methods. These methods are less likely to over-fit and are more capable of providing a reasonable generalization from the limited data available.

(c) The relationship between the predictors and response is highly non-linear

Flexible methods excel in modeling complex, non-linear relationships. They can adapt their parameters to fit the intricate patterns in the data that linear or less flexible methods might miss. Unlike Flexible models, Flexible models are typically designed to model linear relationships. When faced with non-linear data, these methods struggle because they cannot adapt their structure to fit the non-linear nature of the data effectively.

Strictly most of the time, opting for flexible models in the case highly non-linear relationship between the predictors and response is a better choice.

(d) The variance of the error terms, i.e., $\sigma^2 = \mathbf{Var}(\varepsilon)$, is extremely high.
The goal in such a scenario is to find a balance between bias and variance. Inflexible models, by not chasing complex patterns that might be noise, can offer better predictive performance and reliability under these conditions. In the presence of high variance in error terms, an inflexible model is generally a better choice. Its simplicity and robustness against over-fitting make it more suitable for handling the uncertainty and noise associated with high error variance.

Question 2: Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide n and p.

Definitions:

- **Classification**: If the target is categorical, it's a classification problem.
- **Regression**: If the target is a continuous value, it's a regression problem.
- **Inference**: Inference is when we are more interested in understanding the relationships between variables.
- **Prediction**: Prediction is about forecasting an outcome without necessarily understanding the exact relationships between variables.
- **n** is the total number of data points/observations you have.
- **p** is the number of variables/predictors we're using to predict the outcome.

(a) We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.
- Type of Problem: Regression (CEO salary is a continuous variable).
- Interest: Inference (understand relationship between variables and CEO salary).
- n (number of observations): 500 (one for each firm).
- p (number of predictors): 3 (profit, number of employees, industry).

(b) We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.
- Type of Problem: Classification (categorical outcome success/failure).
- Interest: Prediction (forecasting the success or failure of a new product).
- n (number of observations): 20 (for each previously launched similar product).
- p (number of predictors): 13 (price charged, marketing budget, competition price, and ten other variables).

(c) We are interested in predicting the % change in the USD/Euro exchange rate in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market.

-Type of Problem: Regression (predicting a continuous outcome - the % change in the USD/Euro exchange rate).

-Interest: Prediction (forecasting the future value of the exchange rate based on changes in stock markets).

- n (number of observations): The number of weeks in 2012. Since 2012 was a leap year, there would be 52 weeks.
- p (number of predictors): 3 (% change in the US market, % change in the British market, and % change in the German market).

Question 3: We now revisit the bias-variance decomposition.

Definitions:

In statistical learning, **bias-variance decomposition** is a fundamental concept used to understand and improve the performance of machine learning models. It involves breaking down the error of a model into two main components: bias and variance, along with an irreducible error term.

- **Bias**: This refers to the error due to overly simplistic assumptions in the learning algorithm. High bias can cause the model to miss relevant relations between features and target outputs (underfitting). It's essentially the difference between the average prediction of our model and the correct value which we are trying to predict.
- **Variance**: This is the error due to too much complexity in the learning algorithm. High variance can cause the model to model the random noise in the training data, rather than the intended outputs (overfitting). It's the variability of model prediction for a given data point.
- **Irreducible Error**: This is the error inherent in the problem itself, often due to randomness or noise in the data. It can't be reduced by any model.

As previously mentioned and demonstrated, we have:

$$E[(y-\hat{f})^2] = \mathrm{Var}(\hat{f}) + \mathrm{Bias}(\hat{f})^2 + \mathrm{Var}(\varepsilon)$$

(a) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, and the y-axis should represent the values for each curve. There should be five curves. Make sure to label each one.

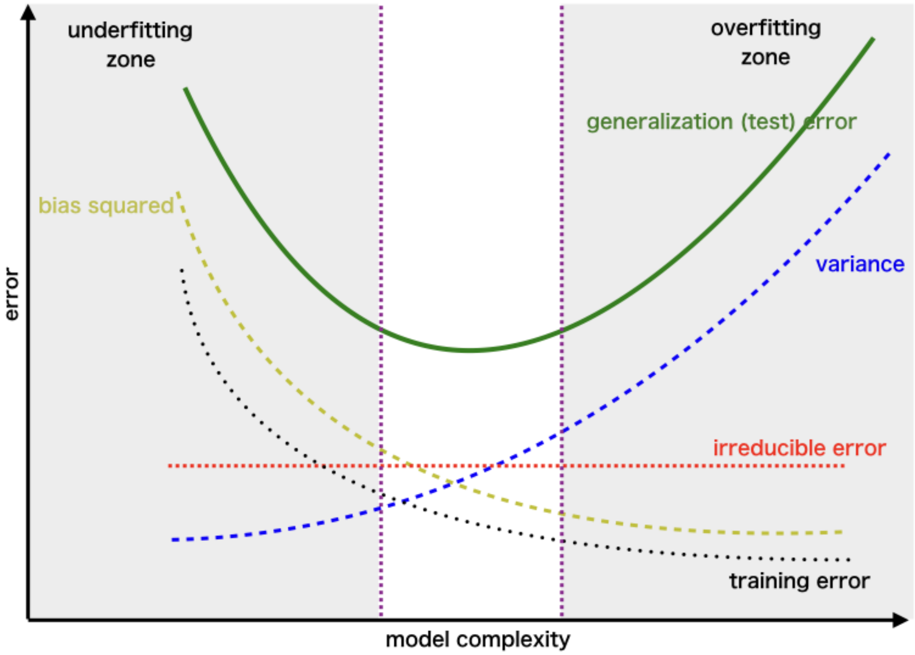

Figure 1: Biais-Variance Tradeoff

(b) Explain why each of the five curves has the shape displayed in part (a).

Bias Squared (yellow curve): As model complexity increases, the model has more capacity to fit the data, and thus bias decreases. This is why the curve starts high and slopes downwards.

Variance (blue curve): High variance can cause over-fitting. As model complexity increases, a model's sensitivity to data fluctuations increases, leading to higher variance. Hence, the curve starts low and slopes upwards.

Training Error (yellow line): This is the error on the training set, which typically decreases as the model becomes more complex, because a more complex model can fit the training data better. Therefore, the curve trends downward. It goes down the irreducible error because it learns the noise at a certain level of complexity.

Generalization (Test) Error (green curve): This is the total error of the model on unseen data. It's the sum of bias squared, variance, and irreducible error. Initially, as complexity increases, the model learns better, and the error decreases. However, after a certain point, increasing complexity only fits to noise, increasing the variance without reducing bias, causing the error to rise again. This creates a U-shaped curve.

Irreducible Error (black dotted line): This is the error that cannot be reduced regardless of how good the model is, often due to noise in the data itself. It's constant, hence the flat line; it doesn't change with model complexity.

Question 4: You will now think of some real-life applications for statistical learning.

(a) Describe three real-life applications in which classification might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer

- **Medical Diagnosis**
  - Response: Diagnosis (e.g., presence or absence of a disease).
  - Predictors: Patient symptoms, medical test results, age, gender, etc.
  - Goal: prediction – accurately diagnose the presence or absence of a disease in new patients. However, inference can also be important, for understanding which factors are most indicative of certain diseases, which can guide treatment and prevention strategies.
- **Credit Scoring in Finance**
  - Response: Creditworthiness (e.g., high risk or low risk).
  - Predictors: Credit history, debts, income, employment status, age.
  - Goal: prediction - determine the likelihood that a person will repay their debts. This helps financial institutions decide whether to grant a loan. Inference can be secondary, useful for understanding which factors most strongly predict default risk.
- **Customer Churn Prediction in Telecommunications**
  - Response: Churn (whether a customer will leave or quite the service).
  - Predictors: patterns, customer interactions, billing history, plans.
  - Goal: prediction – anticipate which customers are at risk of leaving so that the company can take proactive steps to retain them. While prediction is the primary goal, inference can help understand the reasons behind churn, aiding in developing better customer retention strategies.