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

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

Demonstration:

$$
\begin{aligned}
\textbf{We have } \mathrm{Var}[X] &= E[(X - E[X])^2] = E[X^2 - 2XE[X] + E[X]^2] \\
&= E[X^2] - 2E[XE[X]] + E[X]^2 = E[X^2] - 2E[X]E[E[X]] + E[X]^2 \\
\textbf{Because } E[X + Y] &= E[X] + E[Y] \\
\textbf{Then } Var[X] &= E[X^2] - E[X]^2 \\
\textbf{Then } E[X^2] &= \mathrm{Var}[X] + E[X]^2 \\
\textbf{We have } E[f] &= f \\
\textbf{And } y &= f + \varepsilon \\
\textbf{And } E[\varepsilon] &= 0 \\
\textbf{Then } E[y] &= E[f + \varepsilon] = E[f] = f \\
\textbf{Then } \mathrm{Var}[y] &= E[(y - E[y])^2] = E[(y - f)^2] \\
\mathbf{E}[(\mathbf{y} - \hat{\mathbf{f}})^{\mathbf{2}}] &= E[y^2 + \hat{f}^2 - 2y\hat{f}] \\
&= E[y^2] + E[\hat{f}^2] - E[2y\hat{f}] \\
\textbf{Because } E[y^2] &= Var(y) + E[y]^2 \\
\textbf{And } E[\hat{f}^2] &= Var(\hat{f}) + E[\hat{f}]^2 \\
\textbf{And } E[y] &= f \\
\textbf{Then } \mathbf{E}[(\mathbf{y} - \hat{\mathbf{f}})^{\mathbf{2}}] &= \mathrm{Var}[y] + E[y]^2 + \mathrm{Var}[\hat{f}] + E[\hat{f}]^2 - 2fE[\hat{f}] \\
&= \mathrm{Var}[y] + \mathrm{Var}[\hat{f}] + f^2 - 2fE[\hat{f}] + E[\hat{f}]^2 \\
\textbf{Because } f^2 - 2fE[\hat{f}] + E[\hat{f}]^2 &= (f^2 - E[\hat{f}])^2 \\
&= (E[f^2 - \hat{f}])^2 \\
\textbf{And } Var(y) = E[(y - E[y])^2] &= E[(f + \varepsilon - f)^2] \\
&= E[(\varepsilon - 0)^2] = E[(\varepsilon - E[\varepsilon])^2] \\
&= Var(\varepsilon) \\
\textbf{We have finally } \mathbf{E}[(\mathbf{y} - \hat{\mathbf{f}})^{\mathbf{2}}] &= \mathrm{Var}[y] + \mathrm{Var}[\hat{f}] + (f - E[\hat{f}])^2 \\
&= \mathrm{Var}[y] + \mathrm{Var}[\hat{f}] + E[(f - \hat{f})]^2 \\
&= \mathrm{Var}[\varepsilon] + \mathrm{Var}[\hat{f}] + \mathrm{Bias}[\hat{f}]^2 \\
\mathbf{E}[(\mathbf{y} - \hat{\mathbf{f}})^{\mathbf{2}}] &= \mathbf{Var}[\varepsilon] + \mathbf{Var}[\hat{\mathbf{f}}] + \mathbf{Bias}[\hat{\mathbf{f}}]^{\mathbf{2}}
\end{aligned}
\tag{1}
$$

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:

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

(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.

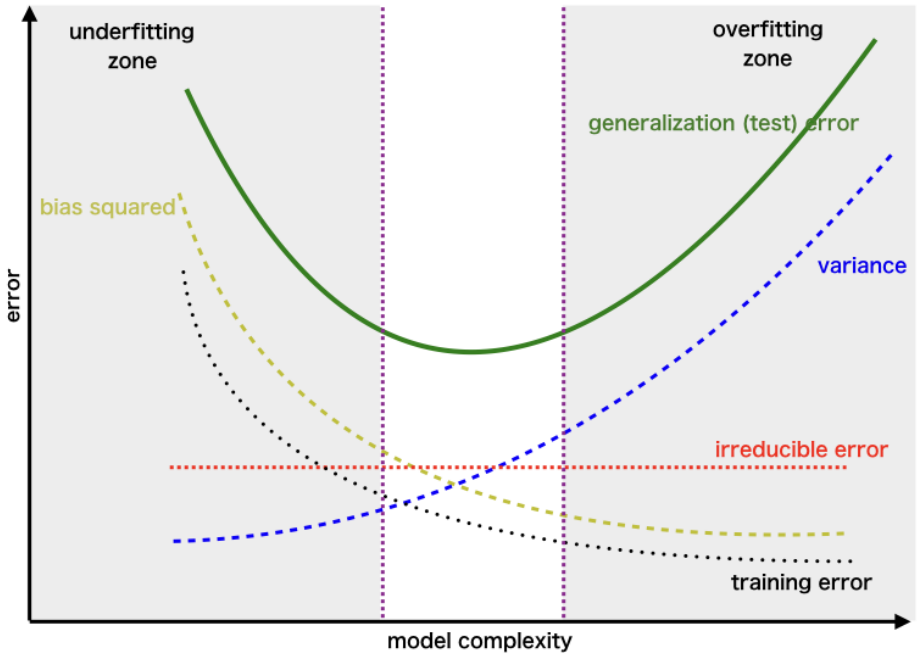

Figure 1: Biais-Variance Tradeoff

(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.

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

- **Real Estate Pricing:**
  - Response: The price of a house.
  - Predictors: Location, size (square footage), number of bedrooms, age of the house, proximity to amenities, etc.
  - Goal: prediction. Real estate agents and buyers use regression to predict the market value of a house based on its characteristics. While some inference about the importance of each predictor (like how much value a bedroom adds) might be made, the primary focus is on accurately predicting prices.
- **Healthcare - Patient Outcomes Prediction:**
  - Response: Patient recovery time or treatment effectiveness.
  - Predictors: Age, gender, pre-existing health conditions, type of treatment, lifestyle factors (such as smoking, diet, exercise), genetic information, etc.
  - Goal: This application serves both prediction and inference. For healthcare providers, predicting patient outcomes helps in tailoring treatment plans. Moreover, understanding how different predictors affect recovery (inference) can guide research and policy decisions in healthcare.
- **Marketing - Customer Lifetime Value (CLV):**
  - Response: The lifetime value of a customer.
  - Predictors: Purchase history, engagement with marketing campaigns, demographic information, browsing behavior on the company website, customer service interactions, etc.
  - Goal: Primarily prediction, as businesses use this information to predict how valuable a customer might be in the long term, aiding in resource allocation for marketing and customer retention strategies. In some cases, there might be an interest in inference as well, particularly in understanding which factors most strongly influence CLV.

(c) Describe three real-life applications in which cluster analysis might be useful.

- **Market Segmentation in Marketing:**
  - Application: Businesses often use cluster analysis to identify distinct groups within their customer base.
  - Data used for clustering: Purchasing habits, customer preferences, demographic data, engagement with marketing channels.
  - Purpose: By understanding different market segments, businesses can tailor their products, services, and marketing strategies to meet the specific needs and preferences of each group. This targeted approach can lead to more effective marketing and higher customer satisfaction.
- **Genomic Data Analysis in Biology:**
  - Application: In biology, particularly in genomics, cluster analysis is crucial for understanding genetic similarities and differences.
  - Data used for clustering: Genetic sequences, expression levels of genes, protein-protein interaction data.
  - Purpose: Scientists use clustering to group genes or organisms based on genetic features, helping in identifying functional relationships, evolutionary patterns, and classifying new species or gene functions. This can guide research into disease mechanisms, drug discovery, and understanding evolutionary biology.
- **Document Clustering in Information Retrieval:**
  - Application: In information retrieval, such as search engines or digital libraries, cluster analysis helps in organizing large volumes of text data.
  - Data used for clustering: Keywords, text content, metadata, user interaction data with documents.
  - Purpose: By grouping similar documents, users can navigate information more efficiently, improving the search experience. It also aids in automatic topic extraction and summarization, which is useful in fields like academic research and legal document analysis.

Question 5: What are the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression or classification? Under what circumstances might a more flexible approach be preferred to a less flexible approach? When might a less flexible approach be preferred

- **Very Flexible Approach**
  - *Advantages:* **Complex Pattern Recognition**: Better at detecting complex relationships in data. **High Predictive Accuracy**: More effective for datasets with high variance and intricate structures. **Adaptability**: Can adapt to a wide range of data shapes and structures.
  - *Disadvantages:* **Overfitting**: Prone to overfitting the training data. **Interpretability**: Often less interpretable. **Computational Intensity**: Requires more computational resources and time. **Data Requirements**: Needs more data to perform effectively.
- **Less Flexible Approach**
  - *Advantages:* **Interpretability**: Easier to understand and interpret. **Generalization**: Better at generalizing to unseen data. **Implicity**: Simpler to implement and requires less computational power. **Stability**: Less susceptible to noise in the training data.
  - *Disadvantages:* **Underfitting**: Can miss important complexities in data. **Predictive Accuracy**: May have lower predictive accuracy in complex data patterns. **Flexibility**: Not as adaptable to varying data structures and relationships.
- **When to Prefer Each Approach**
  - *Prefer More Flexible Approach:* When data has complex patterns that simpler models cannot capture. In scenarios where predictive accuracy is critical, especially in large datasets. When capturing as much information from the data is a priority.
  - *Prefer Less Flexible Approach:* When interpretability is crucial, such as in medicine or social sciences. In cases with smaller datasets to avoid overfitting. When computational resources are limited or simplicity and speed are desired. In situations with linear or nearly-linear data relationships or high noise level.

Question 6: Describe the differences between a parametric and a non-parametric statistical learning approach. What are the advantages of a parametric approach to regression or classification (as opposed to a non-parametric approach)? What are its disadvantages?

Definitions:

- **Parametric Approach**: Parametric methods involve making assumptions about the form of the function that relates the predictors to the response. (i.e.: Linear Regression,Logistic Regression).
- **Non-Parametric Approach**: Non-parametric methods do not make explicit assumptions about the functional form of the relationship between predictors and response. Instead, they seek to estimate the relationship between predictors and response directly from the data. (i.e.: K-Nearest Neighbors (K-NN), Decision Trees, etc.).

- **Advantages of a Parametric Approach**
  - **Simplicity**: By assuming a specific form, parametric models are generally simpler and easier to understand.
  - **Efficiency**: They require fewer data points to estimate the parameters effectively.
  - **Interpretability**: The relationship between variables is often easier to interpret in parametric models.
  - **Predictive Performance**: In cases where the parametric form aligns well with the underlying data structure, these models can perform exceptionally well.
- **Disadvantages of a Parametric Approach**
  - **Bias**: If the chosen model form does not align well with the true underlying data structure, this can introduce bias.
  - **Flexibility**: Parametric models are less flexible in adapting to complex or non-linear relationships unless the correct form is known and specified.
  - **Overfitting to Model Assumptions**: There is a risk of over-reliance on model assumptions, which might not hold in real-world scenarios.

Question 7: The table below provides a training data set containing six observations, three predictors, and one qualitative response variable

(a) Compute the Euclidean distance between each observation and the test point, X1 = X2 = X3 = 0

- Distance from Observation 1: 3.0
- Distance from Observation 2: 2.0
- Distance from Observation 3: 3.162
- Distance from Observation 4: 2.236
- Distance from Observation 5: 1.414
- Distance from Observation 6: 1.732

(b) What is our prediction with K = 1? Why?

With K=1 in K-nearest neighbors (K-NN), the prediction is made based on the single closest data point in the training set to the test point. The prediction is simply the response value of this nearest neighbor. From the previously calculated Euclidean distances, we see that the closest observation to the test point X1=X2=X3=0 is Observation 5, which has the smallest distance of 1.41. Observation 5 has the response value "Green". Therefore, with K=1, our prediction for the test point X1=X2=X3=0 using K-nearest neighbors is "Green".

(c) What is our prediction with K = 3? Why?

For K-nearest neighbors (K-NN) with $K = 3$, the prediction is based on the three closest data points in the training set to the test point. The prediction is typically the most common response value among these three nearest neighbors. From the Euclidean distances calculated earlier, the three closest observations to the test point $X1 = X2 = X3 = 0$ are:

- Observation 5 (Distance = 1.414) with response "Green"
- Observation 6 (Distance = 1.732) with response "Red"
- Observation 2 (Distance = 2.0) with response "Red"

Among these three observations, there are two "Red" responses and one "Green" response. Since "Red" is the most common response among the three nearest neighbors, the K-NN prediction with $K = 3$ for the test point $X1 = X2 = X3 = 0$ is "Red".

(d) If the Bayes decision boundary in this problem is highly non-linear, then would we expect the best value for K to be large or small? Why?

- **Small K Provides More Flexibility**: A smaller value of K makes the K-NN algorithm more flexible and capable of adapting to more complex, non-linear decision boundaries. It allows the model to more closely follow the intricate patterns in the data.
- **Large K Smoothens the Boundary**: In contrast, a larger value of K leads to a smoother, more linear decision boundary. This is because as K increases, the model starts to consider more neighbors, effectively averaging out the local variations and thus, ironing out the non-linearities.
- **Risk of Overfitting vs. Underfitting**: While a smaller K can capture complex decision boundaries, it also increases the risk of overfitting to the noise in the training data. On the other hand, a larger K might underfit the data, especially when the true decision boundary is complex.
- **Balancing Bias and Variance**: The choice of K involves balancing bias and variance. A small K leads to low bias but high variance, while a large K leads to high bias but low variance. In the case of a highly non-linear Bayes decision boundary, the risk of high bias (i.e., oversimplifying the model) is more concerning than high variance.

In summary, for a highly non-linear decision boundary, a smaller value of K is generally preferred as it allows the model to capture the complex patterns in the data more effectively. However, care must be taken to avoid too small a value that might lead to overfitting.