# Solution to ISL Exercise 2

## Question 1

For each of parts (a) through (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 your answer.

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

If we have a sample size $n$ that's extremely large, we expect a flexible method to be able to capture the true relationship between the predictors and the response better than an inflexible method. The reason is that the model sees many more data points, so it learns the actual trend rather than the randomness or noise, which helps prevent overfitting.

Also, if the number of predictors $p$ is small, a flexible method would likely perform better too, since we're working in a low-dimensional space that's easier to handle.

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

For a small sample size, a flexible method is expected to perform worse because there might not be enough points to truly understand the relationship between the predictors and the response, which leads to overfitting.

Also, having a large number of predictors means we'll be working in a very high-dimensional space, which is not ideal.

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

If the relationship between the predictors and the response is highly non-linear, we expect a flexible method to perform better. Though this is more true if we have a large number of observations, which helps reduce overfitting. Otherwise, with a small dataset, it's possible an inflexible method might still perform better.

### (d) The variance of the error terms, i.e. $\sigma^2 = \textbf{Var}(\varepsilon)$, is extremely high.

If the variance of the irreducible error for each value from the true function is extremely high, then I think it's safe to say a flexible method is likely to perform worse than an inflexible method. Since it tries to fit all points, it might end up fitting the noise as well. I'm not sure I can say the same if the number of observations is large, though.

## Question 2

### (a)

In this case, we're interested in finding a relationship between predictor values and a quantitative response, which aligns with a regression problem for inference. Here, the number of observations is $n = 500$ and the number of predictors is $p = 3$.

### (b)

Since we're interested in determining whether a product will be a success or a failure, our response variable is qualitative. Therefore, this is a classification problem. The number of observations is $n = 20$ (20 past products) , and the number of predictors is $p = 13$ (price + marketing budget + competition price + 10 others).

### (c)

In this case, we're interested in the weekly percentage change, which is a quantitative variable. So, we're dealing with a regression problem. The number of observations is $n = 52$ (52 weeks in a year) and the number of predictors is $p = 3$ (US, British, and German stock market

# Question 3

## (a)

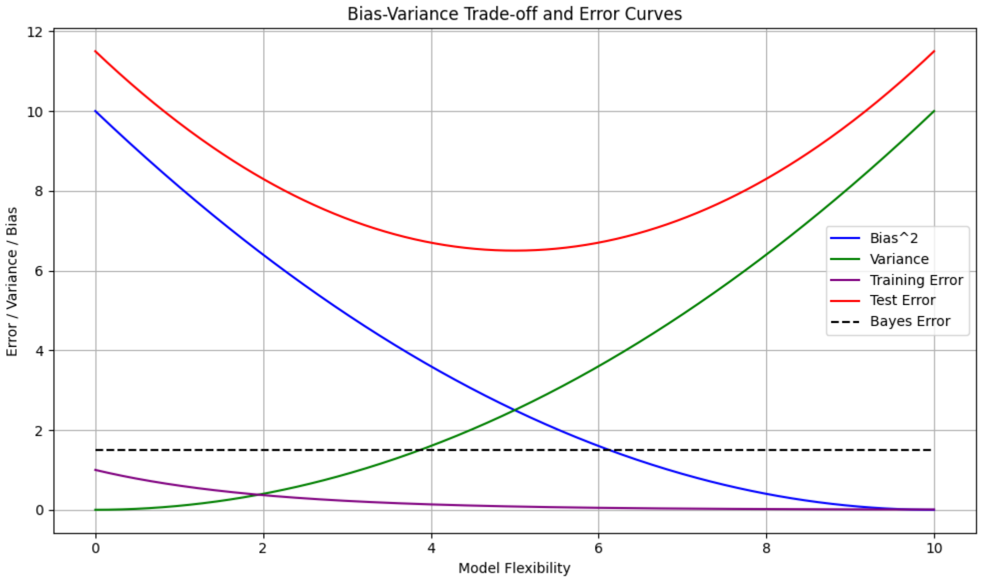

Figure 1: Bias$^2$, Variance, Training Error, Test Error, and Bayes Error as model flexibility increases.

**Note:** I used bias$^2$ to avoid negative values, since we're basically measuring the difference between our model's prediction and the actual value.

## (b)

- **Bias$^2$ (Blue Curve)**

  This decreases as model flexibility increases.

  Reason: Less flexible models cannot capture the underlying pattern well, so they have high bias. As flexibility increases, the model learns better, reducing bias.

- **Variance (Green Curve)**

  This increases as model flexibility increases.

  Reason: Flexible models adapt too much to training data, even to noise. This makes them unstable and increases variance.

- **Training Error (Purple Curve)**

  This goes down quickly as model flexibility increases.

Reason: Flexible models fit the training data more accurately, even perfectly if overfitting occurs.

- **Test Error (Red Curve)**
  This forms a U-shape.
  Reason: At first, as flexibility increases, both bias and test error drop. But after some point, variance increases too much, causing test error to rise again.

- **Bayes Error (Black Dashed Line)**
  This stays constant.
  Reason: This is the irreducible error. It comes from randomness or missing features, and no model can eliminate it.

# Question 4

## (a)

1. **Medical Test for a Disease**
   **Response:** Whether the patient has the disease (Yes/No) — binary classification.
   **Predictors:** Age, blood pressure, test results, symptoms, lifestyle factors, etc.
   **Goal:** *Prediction.* We are using data to predict if someone has the disease. We're more concerned about accuracy than understanding which predictor causes the disease.

2. **Handling Missing Values in a Dataset**
   **Response:** Whether a row with a missing value should be discarded or not (Keep/Discard).
   **Predictors:** Proportion of missing data, which column is missing, data type, overall row quality, etc.
   **Goal:** *Prediction.* The focus is on deciding the correct action for handling the missing value, based on observed patterns.

3. **Weather Forecasting (Rain or No Rain)**
   **Response:** Whether it will rain or not tomorrow (Rain/No Rain).
   **Predictors:** Temperature, humidity, wind speed, cloud cover, pressure readings, etc.
   **Goal:** *Prediction.* We are using current atmospheric data to predict the weather outcome. Again, accuracy matters more than understanding the precise influence of each variable.

## (b)

1. **Understanding the Factors that Increase Student Performance**
   **Response:** Student performance (e.g., exam score or GPA).
   **Predictors:** Study hours, attendance, parental education, school facilities, etc.
   **Goal:** *Inference.* The aim is to understand which factors influence student performance the most. That is we want insight, not just accurate predictions.

2. **Predicting Profit in a Business**
   **Response:** Profit amount (in naira).
   **Predictors:** Advertising budget, number of units sold, production cost, market trends, etc.
   **Goal:** *Prediction.* The objective is to accurately forecast profit based on input variables, not necessarily to explain relationships.

3. **Understanding What Increases Sales**
   **Response:** Sales amount (e.g., number of units sold or revenue).
   **Predictors:** Price, discounts, ad campaigns, product placement, seasonality, customer reviews, etc.
   **Goal:** *Inference.* We're trying to understand which factors significantly impact sales not just to predict, but to draw conclusions from the data.

## (c)

1. **Customer Segmentation in Marketing**
   **Description:** Businesses can group customers into clusters based on their purchasing behavior, age, income, location, or browsing patterns.
   **Goal:** To identify distinct customer types and tailor marketing strategies to each group. E.g: sending different ads to high-spending vs. price-sensitive customers.

2. **Grouping Similar News Articles**
   **Description:** News websites can group articles that talk about similar topics.
   **Goal:** To organize large collections of text and help users discover related content automatically.

3. **Identifying Patterns in Medical Data**
   **Description:** Doctors or researchers can use cluster analysis to find groups of patients with similar symptoms, lab results, or disease progression.
   **Goal:** To improve diagnosis, personalize treatment plans.

# Question 5

**Advantages of a Very Flexible Approach:**

1. It generally leads to a lower bias, as it can capture complex relationships in the data.
2. It performs well when the number of observations is large relative to the number of predictors, achieving both low bias and low variance.

**Disadvantages of a Very Flexible Approach:**

1. It tends to have high variance when the dataset is small, which can lead to overfitting.
2. It is more likely to capture random noise in the data, mistaking it for actual patterns.

**When to Prefer a More Flexible Approach:**

- When we have a large number of observations and relatively few predictors.
- When the goal is prediction rather than inference.

# Question 6

**6. Parametric vs Non-Parametric Approaches**

Parametric: assumes a specific form (like linear regression), estimates a set of parameters.
Non-parametric: does not assume a fixed form, more flexible (e.g. KNN, decision trees).
**Advantages of parametric:**

- Simple to interpret
- Needs less data
- Faster computation

**Disadvantages:**

- Can miss patterns (high bias)
- Less flexible

# Question 7

**(a)** Distances to test point $(0, 0, 0)$:

- Obs 1: $\sqrt{9} = 3$ (Red)
- Obs 2: $\sqrt{4} = 2$ (Red)
- Obs 3: $\sqrt{10} \approx 3.16$ (Red)
- Obs 4: $\sqrt{5} \approx 2.24$ (Green)
- Obs 5: $\sqrt{2} \approx 1.41$ (Green)
- Obs 6: $\sqrt{3} \approx 1.73$ (Red)

**(b)** $K = 1$: Closest is Obs 5 → **Green**

**(c)** $K = 3$: Nearest are Obs 5 (Green), 6 (Red), 2 (Red) → Majority is **Red**

**(d)** If Bayes boundary is highly nonlinear, we want small $K$ (more flexible). So, **small** $K$ **preferred**.