Exercises for ISLP*

Kevin Ma

May 2025

2 Statistical Learning

Conceptual

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) When the sample size n is extremely large and the number of predictors p is small, a flexible method is **better**, since it is better able to account for the patterns present in the more abundant training data, with less risk of overfitting, since outliers are balanced out.

   (b) When p is extremely large and n is small, a flexible method is **worse**, since it overfits to the hyper-specificity of the many parameters in the data, without a large sample size to support any inferred patterns. In these scenarios, flexible models are highly inconsistent and have high variance; that is, they are highly affected by small datasets that are subject to random variation.

   (c) When the relationship between the predictors and response is highly non-linear, a flexible method is **better**, since it is able to account for nuance, while inflexible methods remain rigid.

   (d) When the variance of the error terms, i.e. $\sigma^2 = \text{Var}(\epsilon)$, is extremely high, a flexible method is **worse**, since it will be overly influenced by the noise from the training data and prone to overfitting, while an inflexible method is more stable and better at generalizing to unseen data.

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.

   (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.

   > This is a **regression** problem dealing with a quantitative `salary`. n is 500 and p is 3.

An Introduction to Statistical Learning with Applications in Python: https://www.statlearning.com/

(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.

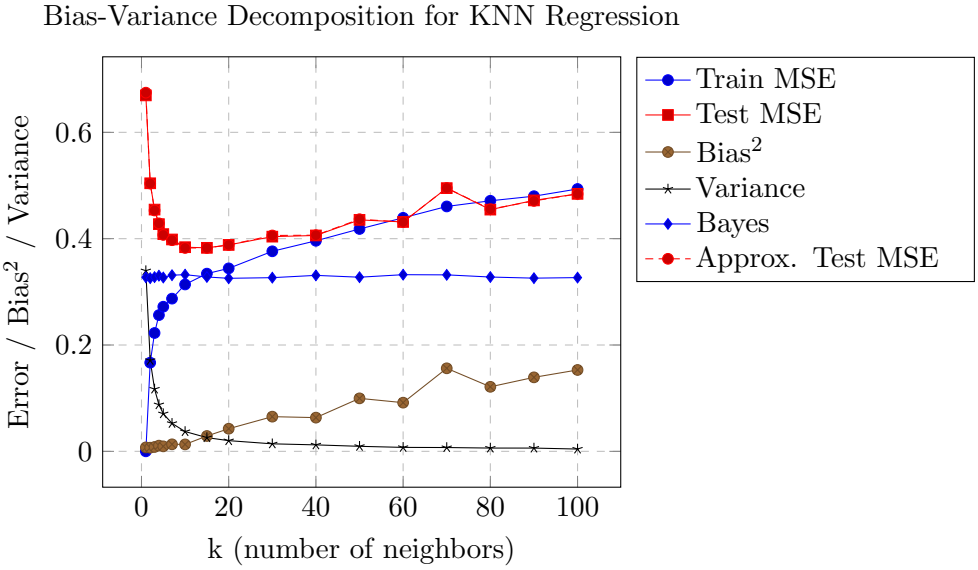
This is a **classification** problem with two categories as output, `success` and `failure`. n is 20 and p is 13.

(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.

This is a **regression** problem dealing with quantitative percent change. n is 52 (weeks in 2012), and p is 3.

3. We now revisit the bias-variance decomposition.

(a)



(b) Data generation process:
https://colab.research.google.com/drive/18Fiz4dJgnacXO8Ix8em3vGMLxiYEF9qE

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.

- Collecting students' academic data to improve effectiveness of teaching strategies. Responses could include indicators like grade point average or standardized test scores, and predictors could include teacher-to-student ratio, class

length, average minutes of each type of classroom activity (e.g. group work, lectures, labs) and amount of daily homework assigned. This is an **inference** problem, since it is important to interpret which of the predictors is most vital to student success.

(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.

- Predicting weather. Predictors could include temperature, pressure, wind, precipitation, and humidity for a specific point in time. Responses could include any of those same predictor variables. This is a **prediction** application, since it is more important to get an accurate forecast than to understand what variables cause the forecast and their relationships.

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

- Grouping professional sports players into archetypes by analyzing their in-game data. Useful for general managers who want to build more cohesive teams or fill a missing role.

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?

> The advantages of a very flexible approach for regression include being able to capture deeper relationships between predictors. However, when training data is not always consistent or does not always resemble the test data, flexible approaches may overfit to the training data, resulting in high variance. Less flexible approaches are less prone to being influenced by a "bad" dataset's noise or outliers, but may not be able to capture more detailed patterns.
>
> Flexible approaches are preferred when interpretability is not that important, and there is a large sample size of data, which minimizes variance. Less flexible approaches are preferred when there is a lot of noise in the dataset, the sample size is small, or interpretability is important.

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?

> A parametric approach assumes the distribution of the data and shape of f, the true function relating the predictors to the responses. A non-parametric approach does not make this assumption, but rather adapts the shape based on the data.
>
> Parametric approaches greatly simplify the fitting problem to estimating a fixed set of parameters, but come with the possibility that the true shape of f is different than assumed. Nonparametric approaches are much more adaptable to different shapes, but require a large sample size to catch on to the true shape.

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

Obs.	X_1	X_2	X_3	Y
1	0	3	0	Red
2	2	0	0	Red
3	0	1	3	Red
4	0	1	2	Green
5	−1	0	1	Green
6	1	1	1	Red

Suppose we wish to use this data set to make a prediction for Y when $X_1 = X_2 = X_3 = 0$ using K-nearest neighbors.

(a) Compute the Euclidean distance between each observation and the test point, $X_1 = X_2 = X_3 = 0$.

Obs.	Euclidean distance	Y
1	$\sqrt{(0-3)^2} = 3$	Red
2	$\sqrt{(0-2)^2} = 2$	Red
3	$\sqrt{(0-1)^2 + (0-3)^2} = \sqrt{10} \approx 3.162$	Red
4	$\sqrt{(0-1)^2 + (0-2)^2} = \sqrt{5} \approx 2.236$	Green
5	$\sqrt{(0-(-1))^2 + (0-1)^2} = \sqrt{2} \approx 1.414$	Green
6	$\sqrt{3(0-1)^2} = \sqrt{3} \approx 1.732$	Red

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

The 1 observation with the smallest Euclidean distance is Obs. 5, and our prediction is the most frequently occuring Y within this set of 1 observation, which is `Green`.

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

The 3 observations with the smallest Euclidean distance are Obs. 5, 6, and 2; and our prediction is the most frequently occuring Y within this set of 3 observations, which is `Red`, with $\Pr(Y = \text{Red} \mid X = (0, 0, 0)) = \frac{2}{3}$.

(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?

The best value for K is small, since a fine-grained decision boundary is necessary to handle the problem's high non-linearity.

Applied

See Google Colab[1].

[1] https://colab.research.google.com/drive/1gZ7HiWsGWmQqEyFaiqSYPD1kgsCf8bTk