AI5213/EC4213: Machine Learning and Deep Learning
Homework 1 (Statistical Learning)

Student ID: 20195024
Name: 김민준 (Kim, Minjun)

Due: Sunday, Sep 10, 11:59PM

This homework comprises a set of conceptual problems and one coding exercise. Some problems are trivial, while others will require a lot of thought. Start this homework early!

> Guideline for those new to data analysis using Python:
> We recommend you to review the Lab section within each chapter (e.g., p. 40 of Ch. 2 or https://islp.readthedocs.io/en/latest/labs/Ch02-statlearn-lab.html) prior to tackling the programming tasks. Additionally, find datasets and Jupyter notebooks at https://github.com/intro-stat-learning/ISLP_labs/ and https://islp.readthedocs.io/en/latest/.
>
> Visit https://www.statlearning.com/forum — a dedicated forum created by and for the ISL community. Whether you have a question or encounter issues with ISLP labs, this platform is your go-to resource for assistance and collaborative discussions.

Deliverables:

1. Submit a PDF of your homework, with an appendix listing all your code, to the Gradescope assignment entitled "HW1 Write-Up". You may typeset your homework in LaTeX or Word or submit neatly handwritten and scanned solutions. Please start each question on a new page. If there are graphs, include those graphs in the correct sections. Do not put them in an appendix. We need each solution to be self-contained on pages of its own.

   • On the first page of your write-up, please sign your signature next to the following statement. (Mac Preview, PDF Expert, and Foxit PDF Reader, among others, have tools to let you sign a PDF file.) We want to make extra clear the consequences of cheating.

   "I certify that all solutions are entirely in my own words and that I have not looked at another student's solutions. I have given credit to all external sources I consulted."

   • On the first page of your write-up, please list students who helped you or whom you helped on the homework. (Note that sending each other code is not allowed.)

2. Submit all the code needed to reproduce your results to the Gradescope assignment entitled "HW1 Code". You must submit your code twice: once in your PDF write-up (above) so the readers can easily read it, and again in compilable/interpretable form so the readers can easily run it. Do NOT include any data files we provided. Please include a short file named README listing your name, student ID, and instructions on how to reproduce your results. Please take care that your code doesn't take up inordinate amounts of time or memory.

For staff use only

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Total
/ 12	/ 12	/ 12	/ 12	/ 24	/ 14	/ 14	/ 100

Honor Code

Declare and sign the following statement:

"I certify that all solutions in this document are entirely my own and that I have not looked at anyone else's solution. I have given credit to all external sources I consulted."

Signature: I certify that all solutions in this document are entirely my own and that I have not looked at anyone else's solution. I have given credit to all external sources I consulted

We welcome group discussions, but the work you submit should be entirely your own. If you use any information or pictures not from our lectures or readings, make sure to say where they came from. Please note that breaking academic rules can lead to severe penalties.

(a) Did you receive any help whatsoever from anyone in solving this assignment? If your answer is 'yes', give full details (e.g. "Junho explained to me what is asked in Question 2a")

(b) Did you give any help whatsoever to anyone in solving this assignment? If your answer is 'yes', give full details (e.g. "I pointed Josh to Ch. 2.3 since he didn't know how to proceed with Question 2")

최유현 and 김아현 and I discussed about whether black box returns neighbors in order or randomly in Q7(b). We concluded that black box will not give neighbors in order of distance.

(c) Did you find or come across code that implements any part of this assignment? If your answer is 'yes', give full details (book & page, URL & location within the page, etc.).

NO.

Q1. Solve ISLP Ch.2, Exercise #2 [12 pts]

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 U.S. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary. [4 pts]

① In this senario, we are not focus on classifying CEO salay into specific categories, and salary varies continously depending on given features (profit, # of emplyes. industry).

② We want to focus how each features affect the CEO salary, not predicting.

③ Data samples are from 500 firms in the US. Features are profit, # of employes and industry.

∴ regression problem, interference, $n=500$ and $p=3$

(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. [4 pts]

① The goal of the senario is to classify new product into two categories, a success and a failure.

② Unlike senario (a), we want to predict whether new product will success or not. It's not relevant to understanding whether which features are associated with the response.

③ Data samples are 20 similar products which were launch before. Features are price charged, marketing budget, competition price, and ten variables which are not shown in this problem.

∴ classification problem, prediction, $n=20$ and $p=13$.

(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. [4 pts]

① This is not classifying the response into classes, but predicting the continous values of the % change in the USD/Euro exchange rate.

② As I mentioned in ①, we want to predict and estimate the response rather than getting interested in the relationships between the response and features.

③ Data samples are 52 weekly data in 2012. Features are the % change in the US market, British market, and German market.

∴ regression problem, prediction, $n=52$ and $p=3$

Q2. Solve ISLP Ch.2, Exercise #4 [12 pts]

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. [4 pts]

i) Predicting lung cancer would be the first example. The response is whether patients will have cancer in the future or not. There are many predictors containing age, tumor size, blood test values, blood type, whether there are any cancer patients in the family, and such things. This is a prediction problem.

ii) Determining whether a email is spam or not is a popular classification example. Response is spam or normal. Predictors can be email address, subject, # of repeated words, country of senders, and such things. This a prediction problem.

iii) Classifying images into certain two groups like dogs and cats is an example. Responses are dogs and cats. Predictors are pixel values. This is a prediction problem.

(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. [4 pts]

i) Realizing which attributes are related to the income of person. Predictors are ages, year, education level. The response is income. This problem is focusing on relationships between X an Y, so it's a interference problem.

ii) Predicting the number of moviegoers before the release of the movie. The responses are the number of moviegoers. And predictors are genre of movie, country, movie release date, and the age of target audience. The purpose is prediction.

iii) Predicting house price is regression problem. The response is the price of the house. Responses are the size of the house, distance from subway stations, surrounding school districts and such things. This is prediction problem since we don't care about how each factors affect the price of houses.

(c) Describe three real-life applications in which cluster analysis might be useful. [4 pts]

i) Marketing companies divide customers into several groups by cluster analysis. They classify customers based on for example, types of items purchased, number of purchases number of visiting the website and such things. So, company can make different events for different groups of customers.

ii) Youtube utilize cluster analysis by dividing users into several groups. They determine user's pattern by user's viewing history, like/dislike history, and subcription information. Then Youtube recommend videos depending on the user group.

iii) SNS platforms analyze data such as posts, hashtags, and places to group posts with similar topics. They use this cluster analysis for recommendation or advertisement.

Q3. Solve ISLP Ch.2, Exercise #5–6 [12 pts]

(a) Describe the ① differences between a parametric and a non-parametric statistical learning approach. What are the ② advantages of a parametric approach to regression or classification (as opposed to a nonparametric approach)? What are its ③ disadvantages? [6 pts]

① There's big differences between parametric methods and non-parametric methods in that the former build the functional shape of model f in advance and the latter does not. For example, in parametric methods, we assume that f is linear and satisfies $f(X)=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_pX_p$ where p is the number of features. And then, we fit the model $f(X)$ by using training datasets while finding appropriate parameters β. In non-parametric methods, we don't assume model f. Instead, we make smooth estimate of f which is very close to data points, requiring large amount of samples.

② Parametric methods have powerful advantages in that we just need to find simple parameters instead of determining estimate of f which is hard. In addition, we need smaller amount of samples(observations) than non-parametric methods if there are a few parameters needed for making a model.

③ The disadvantages of a parametric approach is that the model f doesn't follow true model, which means the performance of f may not be good.

(b) What are the ① advantages and ② disadvantages of a very flexible (versus a less flexible) approach for regression or classification? Under ③ what circumstances might a more flexible approach be preferred to a less flexible approach? ④ When might a less flexible approach be preferred? [6 pts]

① If the real model is non-linear, a linear model (=less flexible model) cannot fit to the real values well. However, flexible model are likely to capture those of non-linear data.

② In a case of a very flexible model, there are some possbities of overfitting, which show bad performance on new observation data sets. In addition to this, a flexible model usually demand much more training data sets than a less flexible model, because the flexible model pursue smooth and accurate fitting on real values of training data sets.

③ As I said before, a more flexible approch is preferable when there are non-linear relationships between observations and real values of Y. And also, when there are so many training data sets or when we want to get accurate prediction, a more flexible model is preferred.

④ A less flexible approach is benificial to realizing relationships between variables and interpret the tendency of model. Moreover, if we only have small amount of training data sets, we can get stable model through a less flexible approach.

Q4. Solve ISLP Ch.2, Exercise #7 [12 pts]

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	$d_{i,test}$
1	0	3	0	Red	3.00
2	2	0	0	Red	2.00
3	0	1	3	Red	3.16
4	0	1	2	Green	2.24
5	-1	0	1	Green	1.41
6	1	1	1	Red	1.73

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, $X_1 = X_2 = X_3 = 0$. [4 pts]

Assume that $d_{i,test}$:= Euclidean distance between ith observation and the test point.

$d_{1,test} = \sqrt{3^2} = 3.00$ $\quad$ $d_{4,test} = \sqrt{1^2+2^2} \doteq 2.24$

$d_{2,test} = \sqrt{2^2} = 2.00$ $\quad$ $d_{5,test} = \sqrt{(-1)^2+1^2} \doteq 1.41$

$d_{3,test} = \sqrt{1^2+3^2} \doteq 3.16$ $\quad$ $d_{6,test} = \sqrt{1^2+1^2+1^2} \doteq 1.73$

(b) What is our prediction with $K = 1$? Why? [2 pts]

Green. 'K=1' indicates nearest observation which is fifth one.

(c) What is our prediction with $K = 3$? Why? [2 pts]

$d_{5,test} < d_{6,test} < d_{2,test} < d_{4,test} < d_{1,test} < d_{3,test}$

Red is our prediction with k=3. This is because three closest observations are fifth, sixth, and second one, which are green, red, red respectively. Therefore test point is more likely to be red.

(d) If the Bayes decision boundary in this problem is highly nonlinear, then would we expect the best value for K to be large or small? Why? [4 pts]

The best value for K should be small. This is largely due to the fact that the boundary becomes linear as K becomes larger. When K is large, we consider more neighbor observations, so this leads to make boundary non-flexible.

Q5. Solve ISLP Ch.2, Exercise #9 [24 pts]

This exercise involves the AUTO data set studied in the lab. Make sure that the missing values have been removed from the data.

[Note] Your code for all of the programming exercises including this one should be submitted to the corresponding Programming submission slot on Gradescope.

(a) Which of the predictors are quantitative, and which are qualitative? [4 pts]

Quantitative predictors are mpg, cylinders, displacement, horsepower, weight, acceleration, and year.

Qualitative predictors are origin and name.

(b) What is the range of each quantitative predictor? You can answer this using the min() and max() methods in numpy. [4 pts]

range of mpg: 9.0 ~ 46.6
range of cylinders: 3 ~ 8
range of displacement: 68.0 ~ 455.0
range of horsepower: 46.0 ~ 230.0
range of weight: 1613 ~ 5140
range of acceleration: 8.0 ~ 24.8
range of year: 70 ~ 82

(c) What is the mean and standard deviation of each quantitative predictor? [4 pts]

	mean	standard deviation
mpg	23.4459	7.8050
cylinders	5.4719	1.7058
displacement	194.4120	104.6440
horsepower	104.4694	38.4912
weight	2977.5842	849.4026
acceleration	15.5413	2.7589
year	75.9796	3.6837

(d) Now remove the 10th through 85th observations. What is the range, mean, and standard deviation of each predictor in the subset of the data that remains? [4 pts]

	range	mean	standard deviation
mpg	11.0 ~ 46.6	24.3685	7.8809
cylinders	3 ~ 8	5.3817	1.6581
displacment	68.0 ~ 455.0	187.7539	99.9395
horsepower	46.0 ~ 230.0	100.9558	35.8956
weight	1649 ~ 4997	2939.6435	812.6496
acceleration	8.5 ~ 24.8	15.7183	2.6938
year	70 ~ 82	77.1325	3.1100

(e) Using the full data set, investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings. [4 pts]

Through "scatter_matrix" function from module "pandas.plotting", we can easily create plots demonstrating the relationships among quantitative predictors.

In addition, we can import "seaborn" to get clear plots.

By analyzing graphs, I could find #. of cylinders tends to rise as displacement rises. And also, displacement rises as weight rises. These tendencies are related to correlation coefficient.

(f) Suppose that we wish to predict gas mileage (mpg) on the basis of the other variables. Do your plots suggest that any of the other variables might be useful in predicting mpg? Justify your answer. [4 pts]

As we can see correlation between each quantitative predictors, mpg has strong correlation with 'weight', 'displacement', 'horsepower', and 'cylinders'. Otherwise, mpg has some correlation with 'year' and 'acceleration', but they are not strong enough.

'weight', 'displacement', and 'horsepower' might be useful for predicting mpg.

Q6. From UW-CSE446-20: Conceptual questions with T/F or short answers. [14 pts]

(a) In your own words, describe what bias and variance are? What is bias-variance tradeoff? [2 pts]

① Bias indicates how far the prediction is from the true value on average.
② Variance is how much the predicted values change for different training data sets.
③ Finally, bias-variance tradeoff is finding appropriate model complexity since bias and variance change on the contrary as the complexity increases. We can also minimize MSE if we deal with bias-variance tradeoff.

(b) What happens to bias and variance when the model complexity increases/decreases? [2 pts]

When the model complexity increases, bias decreases and variance increases.

In the contrast, when the model complexity decreases, bias increases and variance decreases.

(c) True or False: The bias of a model increases as the amount of training data available increases. [2 pts]

False

(d) True or False: The variance of a model decreases as the amount of training data available increases. [2 pts]

True

(e) True or False: A learning algorithm will always generalize better if we use less features to represent our data [2 pts]

False

(f) To get better generalization, should we use the train set or the test set to tune our hyperparameters? [2 pts]

We should not use the test set to tune our hyperparameters. This is largely due to the fact that the test set is just for the final evaluation of the performance of model. Instead, we should devide the train set into several training sets and a validation set, which is for modulating the model before using the test set.

(g) True or False: The training error of a function on the training set provides an overestimate of the true error of that function. [2 pts]

False

Q7. From CMU-10701-20: k-NN Black Box [14 pts]

(a) In a k-NN classification problem, assume that the distance measure is not explicitly specified to you. Instead, you are given a "black box" where you input a set of instances $P_1, P_2, ..P_n$ and a new example Q, and the black box outputs the nearest neighbor of Q, say P_i and its corresponding class label C_i. Is it possible to construct a k-NN classification algorithm (w.r.t the unknown distance metrics) based on this black box alone? If so, how and if not, why not? [7 pts]

It is possible to construct a k-NN classification algorithm based on this black box alone.

We want to know kth nearest neighbor of Q.

For example when k=2, find nearest neighbor of Q, say $P_{i,1}$ and its label $C_{i,1}$. Then, drop $P_{i,1}$ from the observation sets $P_1 \sim P_n$. Only (n-1) instances will remain in the observation sets. We can find nearest neighbor of Q from new sets, say $P_{i,2}$ and its label $C_{i,2}$. $P_{i,2}$ was actually second nearest neighbor before dropping $P_{i,1}$. Therefore we can determine the label of Q from $C_{i,1}$ and $C_{i,2}$.

Likewise, for unkrown K, we can get $C_{i,1}, C_{i,2}, \cdots C_{i,K}$ while dropping P for (k-1)times. Finally, we can determine the label of Q, through comparing the number of different C.

(b) If the black box returns the j nearest neighbors (and their corresponding class labels) instead of the single most nearest neighbor (assume $j \neq k$), is it possible to construct a k-NN classification algorithm based on the black box? If so how, and if not why not? [7 pts]

It is impossible to construct a k-NN classification algorithm based on this black box.

Imagine that k is not a multiple of j. The black boxes don't return the neighbors in order that's close to Q. Therefore, it's impossible to find the k^{th} nearest neighbor, if k is not a multiple of j.

For example, let j=4, k=7. Firstly we get the four nearest neighbors and its labels. Like a method in (a), drop those four neighbors from the set. Then we need the three neighbors in new set, but we can not determine since j=4.

Therefore, it's impossible to construct a k-NN classification algorithm.