AI5213/EC4213: Machine Learning and Deep Learning **Student ID:**
Homework 1 (Statistical Learning) **Name:**

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:

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")

(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.).

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]

(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]

(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]

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]

(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]

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

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

(a) 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 nonparametric approach)? What are its disadvantages? [6 pts]

(b) 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? [6 pts]

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

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

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

(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]

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]

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

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

(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]

(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]

(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]

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]

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

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

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

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

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

(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]

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]

(b) If the black box returns the j nearest neighbors (and their corresponding class labels) instead of the single 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]