DS5220: Supervised Machine Learning and Learning Theory Fall 2022

Problem Set 1

Instructor: Hongyang Ryan Zhang *Due:* ***September 30, 2022, 11:59pm***

Instructions:

- You are expected to write up the solution on your own. Discussions and collaborations are encouraged; remember to mention any fellow students you discussed with when you turn in the solution.
- There are up to three late days for all the problem sets and project submissions. Use them wisely. After that, the grade depreciates by 20% for every extra day. Late submissions are considered case by case. Please reach out to the instructor if you cannot meet the deadline.
- Submit your written solutions to Gradescope and upload your code to Canvas. You are recommended to write up the solution in LaTeX.

Problem 1 (20 points)

(a) (1 point) Calculate $Var(X)$ when X represents the outcome when a fair coin flip (i.e., $E[X] = 1/2$).

(b) (1 point) Find the expected value of the sum obtained when n fair coin flips are rolled independently.

(c) (2 point) For three events A, B, and C, we know that: A and C are independent, B and C are independent, A and B are disjoint, $P(A \cup C) = 2/3$, $P(B \cup C) = 3/4$, $P(A \cup B \cup C) = 11/12$. Find $P(A)$, $P(B)$ and $P(C)$.

(d) (2 points) Consider a test to detect a disease (e.g., COVID-19), assuming that 0.6% of the population has it. The test is 97% effective in detecting an infected person. However, the test gives a false-positive result in 1% of cases (meaning that it shows a positive result if the person is not infected). What is the probability that a person gets a negative test result?

(e) (2 points) If a person tests positive for the disease, what is the probability that they actually have COVID?

(f) (2 points) If a person tests negative for the disease, what is the probability that they are infected with COVID?

Along with the tests, data regarding the number of symptoms shown by the patients was also recorded and is given below. The data was collected from 2 different sources.

No. of Symptoms	Patients
1	20
2	20
3	20
4	20

No. of Symptoms	Patients
1	70
2	15
3	10
4	5

(g) (2 points) Suppose you pick one patient from each of the above 2 sources independently. What would be the expected number of symptoms detected in each of them?

(h) (2 points) Prove that $Var(X) = \mathbb{E}[X]^2 - (\mathbb{E}[X])^2$. Explain the interpretation of this derivation.

(i) (2 points) Let Y_1 and Y_2 denote the number of symptoms detected in each of the above two patients respectively, where $Y_1, Y_2 \in [1, 2, 3, 4]$. Then calculate the following probabilities: (i) $E[Y_1 Y_2]$; (ii) $Var[Y_1 - Y_2]$.

(j) (2 points) Among a population of n people, let X be the number of people that test positive. What is the expectation of X, $E[X]$? What is the variance of X, $Var[X]$? Make sure to include all the steps in the calculation.

(k) (2 points) Define bias error and variance error. What do you understand by Bias-Variance trade-off?

Problem 2 (20 points)

(a) (2 points) Show that for any arbitrary matrix $X \in \mathbb{R}_{m \times n}$, the matrix $XX^\top$ is always positive semi-definite.

(b) Recall that the SVD of a rank-r matrix M has the form

$$M = \sum_{i=1}^{r} \sigma_i u_i v_i^T,$$

where $\{u_i\}_{i=1}^r$ denote the left singular vectors, $\{v_i\}_{i=1}^r$ denote the right singular vectors, and $\{\sigma_i\}_{i=1}^r$ denote the singular values.

i) (2 points) Let

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

Calculate the left and right singular vectors $\{u_i\}_{i=1}^r$ and $\{v_i\}_{i=1}^r$ of A. Then show that $\{u_i\}_{i=1}^r$ and $\{v_i\}_{i=1}^r$ are the eigenvectors of $AA^\top$ and $A^\top A$.

ii) (5 points) Let $M \in \mathbb{R}^{m\times n}$ be an arbitrary real-valued rank-r matrix, show that the eigenvectors of MM^T and M^TM are $\{u_i\}_{i=1}^r$ and $\{v_i\}_{i=1}^r$ respectively.

(c) Recall that the best rank-k approximation of M in Frobenius norm is attained by

$$B = \sum_{i=1}^{k} \sigma_i u_i v_i^T.$$

i) (2 points) For the matrix A defined above, calculate the best rank-1 approximation of A in Frobenius norm. Then find out the approximation error $||M - B||_F$.

ii) (5 points) Let $M \in \mathbb{R}^{m\times n}$ be an arbitrary real-valued rank-r matrix. Show that

$$||M - B||_F = \sqrt{\sum_{i=k+1}^{r} \sigma_i^2}.$$

(d) (4 points) Write a Python file to verify your calculation in (b-i) and (c-i). You may find the library `numpy.linalg.svd` and `numpy.linalg.eig` useful.

Problem 3 (15 points)

(a) (6 points) For vectors $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{a} \in \mathbb{R}^n$ and matrices $\mathbf{X} \in \mathbb{R}^{n\times n}$, $\mathbf{A} \in \mathbb{R}^{n\times n}$, show the following:

(i) $\dfrac{\partial \boldsymbol{a}^T\boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{a}$.

(ii) $\dfrac{\partial \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = (\boldsymbol{A} + \boldsymbol{A^T})\boldsymbol{x}$.

(iii) $\dfrac{\partial ||\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}||_2^2}{\partial \boldsymbol{x}} = 2\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y})$.

(b) (4 points) You are given a training set $\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$, where $\boldsymbol{x_i} \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. Consider the regression problem

$$\min_{\theta\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{\theta}^T\boldsymbol{x}_i))^2.$$

What is the minimizer of the above regression problem? Provide all steps of your derivation. Feel free to assume that the rank of $\{\boldsymbol{x}_i\}_{i=1}^n$ is equal to d.]

(c) (5 points) Let the cost function to minimize is:

$$J(w) = \sum_{i=1}^{n} (y_i - \boldsymbol{\theta}^T \boldsymbol{x}_i))^2 + \lambda \sum_{j=0}^{d} {\theta_j}^2$$

Prove that the vector $w^\star$ that minimizes $J(w)$ is:

$$w^\star = (X^\top X + \lambda I)^{-1} X^\top y,$$

where X is the n by d design matrix, whose i-th row is x_i, and $y = (y_1, ..., y_n)^\top$.

Problem 4 (15 points)

We consider a regression problem for predicting the demand of bike-sharing services in Washington D.C.[1] The prediction task is to predict the demand for the bikes (column cnt) given the other features: ignore the columns instant and dteday. Use the day.csv file from the data folder.

(a) (4 points) Write a Python file to load day.csv.[2] Compute the correlation coefficient of each feature with the response (i.e., cnt). Include a table with the correlation coefficient of each feature with the response. Which features are positively correlated (i.e., have positive correlation coefficient) with the response? Which feature has the highest positive correlation with the response?

(b) (2 points) Were you able to find any features with a negative correlation coefficient with the response? If not, can you think of a feature that is not provided in the dataset but may have a negative correlation coefficient with the response?

(c) (5 points) Now, divide the data into training and test sets with the training set having about 70 percent of the data. Import `train_test_split` from `sklearn` to perform this operation. Use an existing package to train a multiple linear regression model on the training set using all the features (except the ones excluded above). Report the coefficients of the linear regression models and the following metrics on the training data: (1) RMSE metric; (2) R^2 metric.

[Hint: You may find the libraries `sklearn.linear_model.LinearRegression` useful.]

(d) (2 points) Next, use the test set that was generated in the earlier step. Evaluate the trained model in step (c) on the testing set. Report the RMSE and R^2 metrics on the testing set.

(e) (2 points) Interpret the results in your own words. Which features contribute mostly to the linear regression model? Is the model fitting the data well? How large is the model error?

[1] `https://www.kaggle.com/datasets/marklvl/bike-sharing-dataset?search=bike+demand+Washington&select=Readme.txt`. You can also find a Readme.txt file that explains all the features in the dataset.

[2] Refer to `https://docs.python.org/3/library/csv.html` on how to load a csv file in Python.

Problem 5 (10 points)

This question should be answered using the `Diabetes` data set that is readily available in the `Scikit-learn` library.[3] This data set has information about 442 patients and whether they have suffered from diabetes or not.

(a) (2 points) Fit a multiple regression model to predict `Diabetes` using `Age`, `Sex`, `BMI`, and `BP`.

(b) (4 points) Provide an interpretation of each coefficient in the model. Be careful—some of the variables in the model are qualitative!

(c) (2 points) Write out the model in equation form, being careful to handle the qualitative variables properly.

(d) (2 points) Using the model from (c), obtain 95% confidence intervals for the coefficient(s).

Problem 6 (20 points)

We will now perform cross-validation on a simulated data set.

(a) (2 points) Generate a simulated data set as follows:

```
numpy.random.seed(12345)
x = numpy.random.normal(0, 1, (200))
y = x + 2 * x**2 - 2 * x**3 + numpy.random.normal(0, 1, (200))
```

In this data set, what is n and what is p? Write out the model used to generate the data in equation form.

(b) (2 points) Create a scatterplot of X against Y. Comment on what you find. (Hint: You may find `matplotlib.pyplot.plot()` helpful)

(c) (9 points) Set a random seed `123`, and then compute the leave-one-out cross validation errors that result from fitting the following five models using least squares:

(i) $Y = \beta_0 + \beta_1 X + \varepsilon$

(ii) $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$

(iii) $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$

(iv) $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \varepsilon$

[3] `https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_diabetes.html`. You can find the description of this data set at `https://scikit-learn.org/stable/datasets/toy_dataset.html`.

(v) $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \beta_5 X^5 + \varepsilon$

[Hint: You may find `LeaveOneOut()` and `cross_val_score()` in `sklearn.model_selection` helpful.]

(d) (2 points) Repeat (c) using another random seed `12345`, and report your results. Are your results the same as what you got in (c)? Why?

(e) (5 points) Which of the models in (c) had the smallest leave-one-out cross validation error? Is this what you expected? Explain your answer.