CMPE 462 Machine Learning
Spring 2020 Final Exam
Due: July 4 by 11.59pm

DIRECTIONS AND RULES:

- Late submission via email is NOT allowed. Your grade will be 0.
- This is an individual exam. Do NOT share your answers with your classmates. Do NOT ask your classmates for answers. Do NOT forget that you signed an academic integrity document in the beginning of the semester.
- If you get help from an outside source, you are obliged to cite your references.
- Use a clear and concise language. Proofread your answers. Please make sure that there are no typos or grammatical issues that may cause vague answers. If your answer is not clear, you may lose points.
- Please type your answers. Scanned handwritten submissions will NOT be accepted. Your grade will be 0.
- Submit a single pdf file named `surname.pdf` through the designated Turnitin assignment on Moodle.
- When it is required in a question, include your code snippets or screenshots in your pdf file. You will NOT upload a notebook.
- Please type your name and student ID on the first page of your pdf file.

TIPS

- All the questions can be answered based on course material, in-class discussions, and the projects you did. Do not waste your energy by googling answers.
- Do not wait till the last moment to type your answers. You have one and a half day to work on the questions and type.
- Q3 is very short. You will not calculate the entropy mathematically.
- In Q4, you may think of the answer intuitively. However, you need to verify your answer mathematically to get the full credit.
- Q5 is similar to the toy example in the slides and quiz. You do not need any software to find the optimal parameters. You have only two data points. For this reason, derivations will be short.
- In Q1 and Q2, you do not need to implement anything from scratch. You just need to decide how to approach the problem. You can use numpy and sklearn functions to implement your solution.

Good luck and congratulations to graduating students!

Questions

1. (30 pts) You are given a small dataset and you would like to classify the data points. Please find the train and test splits in `data` folder. You can load the dataset using the code below.

```
import numpy as np

train_data = np.load('train_data1.npy')
train_label = np.load('train_label1.npy')
test_data = np.load('test_data1.npy')
test_label = np.load('test_label1.npy')
```

   To classify the data points, you will train an SVM classifier using the training data, and evaluate the performance on the test data using the code below.

```
import numpy as np
from sklearn import svm

clf = svm.SVC(gamma=0.001, C=100.)
clf.fit(train_data, train_label)
y_pred = clf.predict(test_data)
correct_prediction = np.equal(y_pred, test_label)
accuracy = np.mean(correct_prediction.astype(np.float32))
```

   When you follow the directions above, you can achieve a 59% classification accuracy. The performance is poor. We would like to improve the performance. However, the classifier is fixed. We are NOT allowed to change the classifier, its kernel, and hyper-parameters.

   (a) (10 pts) If the classifier is fixed, what could be the reason of poor performance? What would be the possible steps to investigate the problem?

   (b) (10 pts) After your investigation, what is your conclusion? Please clearly state your solution and explain the reason behind it. Implement your solution and report the accuracy.

   (c) (10 pts) Please include your code as a snapshot.

2. (30 pts) In this question, you are again given a dataset, but this time without the labels in `data/data2.npy`. However, we still would like to classify the data points.

   (a) (20 pts) Please propose a solution to obtain labels. Elaborate the reason behind your solution. Include your code as a snapshot. You can again load the dataset using `numpy.load`.

(b) (10 pts) Please classify the data points using the labels you found. Report the classification accuracy. You can use any classifier of your choice, or the classifier in Q1 (it will be faster for you). Include the snapshot of your code.

3. (10 pts) Assume you would like to classify patients as health vs. diabetes using decision tree. You have the following information about your patients:

```
patient ID, glucose, insulin, age, gender
```

Suppose you try each attribute as the root node. Which attribute do you think gives the lowest entropy after splitting the data? Do you think all the attributes above are meaningful to learn a generalizable tree?

4. (15 pts) Consider a data set of 1,000 customers. You would like to apply principal component analysis to the customer data. Suppose one attribute of 200 customers is missing. All other attributes are assumed to be non-missing. There could be two ways to handle the missing values. You could either discard the customers with missing values, or you could impute the missing attribute with the average value of 800 customers. Are the principal components found using the two approaches different? To get the full point, you should show your steps clearly.

5. (15 pts) You are given two data points $x_1 = (1, 0)^T$, $y_1 = -1$, and $x_2 = (3, 0)^T$, $y_2 = 1$.

(a) Compute the optimal w and b in support vector machine by solving the primal formulation given below.

$$\begin{aligned} \min_{w,b} \quad & \tfrac{1}{2} w^T w \\ \text{subject to} \quad & y_i(w^T x_i + b) \geq 1, \forall i. \end{aligned}$$

Do NOT use any quadratic programming tool. Solve the optimization problem above by hand. Clearly show your steps. Only the numerical results without the steps will NOT be graded.

(b) Compute the optimal α in the dual formulation of support vector machine given below.

$$\begin{aligned} \max_{\alpha \in \mathbb{R}^N} \quad & \textstyle\sum_{n=1}^N \alpha_n - \frac{1}{2} \sum_{n=1}^N \sum_{m=1}^N y_n y_m \alpha_n \alpha_m x_n^T x_m \\ \text{subject to} \quad & \textstyle\sum_{n=1}^N y_n \alpha_n = 0 \\ & \alpha_n \geq 0, n = 1, \cdots, N. \end{aligned}$$

Do NOT use any quadratic programming tool. Solve the optimization problem above by hand. Clearly show your steps. Only the numerical results without the steps will NOT be graded.

(c) Compute the optimal w based on the optimal α obtained from the dual formulation of support vector machine. Compare with the results in (a). You should be able to obtain the same result.