CSE 590 **Introduction to Machine Learning** Spring 2022

# Mid-Term Exam 2

**4/29 10:00am – 4/30 11:59pm**

**NO CODING for any of the problems**

**Problem #1 (15 points)**

**FOR SECTIONS 590-12 and 590-53 (undergraduate) ONLY**

1. Explain the motivation of using an ensemble of classifiers. What are the advantages/disadvantages of using this strategy?
2. Identify and explain cases where an ensemble approach can result in a performance that is:
3. Better than the best individual classifier
4. Worse than the best individual classifier
5. Comparable to the best performing individual classifier

**Problem #1 (15 points)**

**FOR SECTIONS 590-07 and 590-56 (graduate) ONLY**

For a given application (with training/validation data sets), you are provided with 10 classifiers (already trained and optimized). For each classifier, you are provided with a description of the algorithm and its output (probability of each sample in each class). You are tasked with training and optimizing an ensemble method that selects and combines ONLY 5 of the 10 classifiers. Identify 4 criteria that you can use in selecting these classifiers. You are allowed to train/validate only once after selecting the 5 classifiers.

**Problem #2 (15 points)**

Create a 2-dimensional data set with 20 samples that has the following properties:

1. Samples should belong to 2 clusters (10 samples per class)
2. Data cannot be clustered correctly using the K-Means algorithm
3. Data can be clustered correctly using the Hierarchical Agglomerative clustering

Explain why the K-Means cannot generate the correct clusters.

What kind of linkage is needed for the Agglomerative algorithm to cluster the data correctly?

**Problem #3 (15 points)**

Create a 2-dimensional data set with 22 samples that has the following properties:

1. 20 of the samples should belong to 2 clusters (10 samples per class)
2. The remaining 2 samples are noise
3. Data cannot be clustered correctly using the K-Means algorithm (with K=2)
4. Data cannot be clustered correctly using the Hierarchical Agglomerative clustering (with K=2)
5. Data can be clustered correctly using DBSCAN

Explain why the K-Means and the Agglomerative algorithms cannot generate the correct clusters.

Explain why the DBSCAN is the appropriate algorithm for this dataset.

**Problem #4 (15 points)**

List three different reasons for trying to reduce the number of features prior to applying a machine learning algorithm. Justify and explain each reason.

**Problem #5 (15 points)**

Identify two cases where accuracy may be an inadequate measure to evaluate the performance of a classification algorithm. Explain the reasons.

For each case, provide an alternative scoring measure and explain why it is more reliable than the accuracy.

**Problem #6 (10 points)**

Given the following pseudo-code that is supposed to train and test an SVM classifier on the Iris data.

* *Load Iris data*
* *Normalize data to have zero mean and unit variance*
* *X\_train, X\_test, y\_train, y\_test = train\_test\_split (iris.data, iris.target)*
* *svm.fit (X\_train, y\_train)*
* *score = svm.score (X\_test, y\_test)*

Is the above algorithm logically correct? If not, identify the problem and correct it.

**Problem #7 (15 points)**

Suppose that we have 3 classification algorithms. Each algorithm has two parameters: P1 and P2.

After performing a grid search for each algorithm (using {0.01, 0.1, 1, 10} for each parameter), we obtain the following accuracy results:

1. Algorithm 1

***P2***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **10** | 0.71 | 0.80 | 0.93 | 0.93 |
| **1** | 0.70 | 0.70 | 0.92 | 0.91 |
| **0.1** | 0.65 | 0.71 | 0.90 | 0.89 |
| **0.01** | 0.63 | 0.65 | 0.88 | 0.87 |
|  | **0.01** | **0.1** | **1** | **10** |

***P1***

Did we use the correct range of values for each parameter? Justify your answer.

If the answer is no, then what other values for P1 and P2 do you recommend exploring?

1. Algorithm 2

***P2***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **10** | 0.90 | 0.95 | 0.91 | 0.89 |
| **1** | 0.88 | 0.93 | 0.87 | 0.82 |
| **0.1** | 0.75 | 0.88 | 0.84 | 0.70 |
| **0.01** | 0.63 | 0.79 | 0.73 | 0.67 |
|  | **0.01** | **0.1** | **1** | **10** |

***P1***

Did we use the correct range of values for each parameter? Justify your answer.

If the answer is no, then what other values for P1 and P2 do you recommend exploring?

1. Algorithm 3

***P2***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **10** | 0.85 | 0.90 | 0.88 | 0.82 |
| **1** | 0.83 | 0.93 | 0.91 | 0.85 |
| **0.1** | 0.75 | 0.89 | 0.84 | 0.70 |
| **0.01** | 0.63 | 0.79 | 0.73 | 0.67 |
|  | **0.01** | **0.1** | **1** | **10** |

***P1***

Did we use the correct range of values for each parameter? Justify your answer.

If the answer is no, then what other values for P1 and P2 do you recommend exploring?