Due: September 27, 2021, 12:00 PM

NOTE: You are allowed to have at most one page of answer for each problem - total of 5 pages (1 page for problems 1 to 5). This restriction is due to limited grader time. If you have more than 5, we will only grade the first 5 pages.

Problem 1 - PCA (Moved from HW1)

- Compute the covariance matrix for the following matrix: $\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$
- After performing eigenvalue decomposition on covariance matrix C, when is the smallest eigenvalue equal to zero?
- Assume after performing eigenvalue decomposition on covariance matrix C, the first two eigenvalues are equal. Explain when this could happen.

Problem 2 - Decision Trees

- In tree induction, can greedy splitting algorithm (based on impurity measures, assuming all attributes are not numerical) always reach the purest split at the end? If yes, explain why. If no, provide a counterexample.
- What is the maximum value for the entropy of a random variable that can take n values? Justify your answer.
- Why does $\text{GAIN}_{\text{split}}$ splits the tree into many small partitions? How does $\text{GainRATIO}_{\text{split}}$ address this issue? Explain using mathematical formulations.
- How does Pre-Pruning and Post-Pruning address over-fitting? What is the difference between them?

Problem 3 - Evaluations

- Calculate precision, recall and F-measure based on the following confusion matrix:

		Actual Value	
		Positive	Negative
Predict Value	Positive	5	15
	Negative	20	10

- Why do ROC curves always go from (0, 0) to (1, 1) regardless of the classifiers? What is the meaning of the straight line in the middle?

Problem 4 - k Nearest Neighbor

- How can we select the best k for kNN classifiers? What if k is too small? What if k is too large?
- What is the time complexity of kNN? Justify your answer.

Problem 5 - Naive Bayes

- A new medical test has been designed to detect the presence of the mysterious Brainlesserian disease. Among those who have the disease, the probability that the disease will be detected by the new test is 0.9. However, the probability that the test will erroneously indicate the presence of the disease in those who do not actually have it is 0.02. It is estimated that 11% of the population who take this test have the disease.

  If the test administered to an individual is positive, what is the probability that the person actually has the disease?

- In Naive Bayes, how can we estimate the conditional probabilities of continuous attributes?
- Please illustrate the independence assumption of Naive Bayes using an example.
- The independence assumption does not hold in most real-world data. However, Naive Bayes still works well in most scenarios. Why does this happen? What if the dataset has duplicate attributes?

Problem 6 - (Programming) - Naive Bayes Classifier

Let's build a Naive Bayes classifier from scratch (not using prebuilt packages). Download adult.csv from

https://datahub.io/machine-learning/adult

You can check the distributions of dataset attributes by visiting 'input' in

https://www.kaggle.com/prashant111/naive-bayes-classifier-in-python/data

1. In the dataset, the to-be-predicted label is the last attribute: 'class'. Perform the following preprocessing steps:
   - Feature selection: remove features 'capitalgain', 'capitalloss' and 'native-country.'
   - Remove instances with at least one missing value '?'.
   - Randomly split the dataset into 90% as train and 10% as test. Remember to utilize a random seed for sampling train and test instances.
2. Build the classifier:
   - Estimate the probabilities for continuous attributes ('age', 'fnlwgt', 'education-num' and 'hoursperweek') by fitting a Normal distribution.
   - Calculate the probabilities of discrete attributes (remaining attributes) using Laplace smoothing.
3. Perform the classification:
   - Train and test with the corresponding train and test datasets;
   - Report the testing accuracy; and
   - Only submit code and performance results (accuracy).