EEL 4930 Machine Learning **Name:** SOLUTIONS
Spring 2020
Midterm Exam
March 13, 2020
Time Limit: 120 Minutes (7:20 PM - 9:20 PM)

Grade Table (for teacher use only)

Question:	1	2	3	4	5	6	7	8	9	10	11	Total
Points:	5	10	10	10	10	10	10	20	5	10	0	100
Score:												

Answer the questions in the spaces provided on the question sheets. If you run out of room for an answer, continue on the back of the page.

Write your name in the formula sheet.

- One-page formula sheet (front and back, written with pen or printed), calculator, no computer, closed book, closed notes
- Write legibly
- There are 10 questions for a total of 100 points
    - Question 11 is a **bonus question** and is worth up to 5 points

1. (5 points) What is the difference between regression and classification?

   Regression is a method to determine a manifold that passes through the data (line, hyperplane or curved surface). The output of regression is a continuous value number. Regression requires supervision.

   Classification is a method that attributes class labels (integer-valued labels) to data samples. Classification methods can typical be discriminative (where feature space in partitioned into class regions) or generative (where each class' samples are modeled with a probability distribution).

2. (10 points) This question has three parts.

   (a) (3 points) What is overfitting in machine learning? Be precise.

   Overfitting refers to the phenomenon of a model/algorithm memorize training data such that generalization to a test set is poor.

   Generally results from having too many parameters in the model.

   (b) (3 points) In practice, how can you determine whether you have overfitted your machine learning system?

   Cross-validation provides an indication as well as using a hold-out test set.

   If the results are very good on training/validation data and poor on test data, that generally indicates overfitting.

(c) (4 points) What strategies can you apply in order to avoid overfitting?

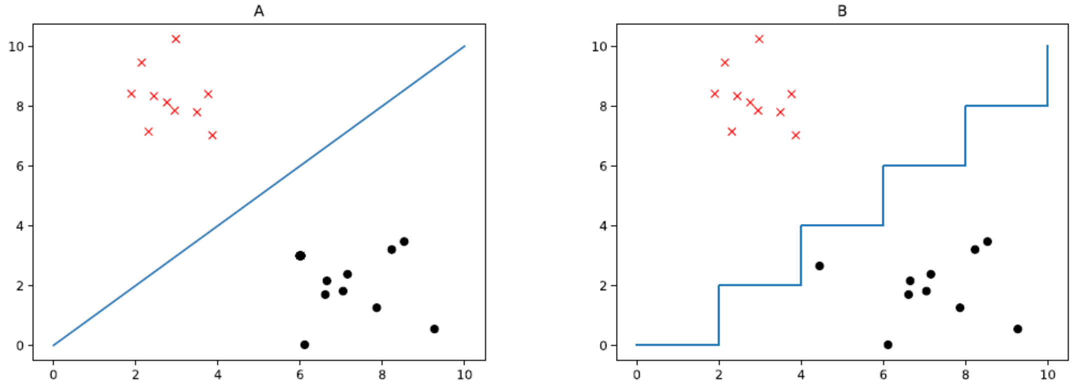
There are a few strategies to use in order to avoid overfitting. They include:

- Add more data. More data prevents the model from memorizing the training data.
- Regularization. Including a regularization penalty on the model's parameters. It has been observed that the model parameters become too large when the model is overfitting. So we can include a regularization penalty that enforces the parameters to be small.
- Cross-validation. Perform cross-validation to make sure the model is generalizes and learns from all sets of samples.
- Decrease model complexity.

3. (10 points) Suppose you are designing a k-Nearest Neighbors (k-NN) classifier for a 10-dimensional data set where each dimension (or feature) is equally informative for classification but each dimension has a different scale and range. For example, suppose the first feature has a mean of 0.1 with a variance of 0.001 whereas the second feature has a mean of 100 with a variance of 35. Would normalization of the data prior to application of k-NN be helpful? Clearly and completely describe why or why not. What sort of distance measure would you use for this data given your response?

   Yes, if you are using Euclidean distance then the second feature would outweigh the first feature in the distance computation. This is because Euclidean distance does not account for relative variance and scale during computation.

4. (10 points) Consider the following two decision boundaries estimated using the same training data but with two different learning algorithms, the support vector machine and a single decision tree that branches based on a single feature at each node. Which method estimated each decision boundary? Clearly provide the reasons for your selection.



Decision boundary A was generated with a SVM. A SVM creates a decision boundary that maximizes the margin that separates the two classes.

Decision boundary B was generated using a single decision tree with a single feature split. This is visible as each (single feature) split in the decision tree corresponds to a straight line (or hyperplane) in the feature space. This is why the decision boundary looks like a staircase.

5. (10 points) Recall our discussion of the *volume of the crust*, i.e., the case of the a inner sphere S_2 of radius $r - \epsilon$ inscribed within an outer sphere S_1 of radius r and the relative volume of the crust and the outer sphere S_1 as we increase dimensionality:

$$\frac{V_{crust}}{V_{S_1}} = \frac{V_{S_1} - V_{S_2}}{V_{S_1}} = 1 - \frac{V_{S_2}}{V_{S_1}} = 1 - \frac{\frac{(r-\epsilon)^D \pi^{\frac{D}{2}}}{\Gamma\left(\frac{D}{2}+1\right)}}{\frac{r^D \pi^{\frac{D}{2}}}{\Gamma\left(\frac{D}{2}+1\right)}} = 1 - \left(1 - \frac{\epsilon}{r}\right)^D$$

Describe how this concept relates to the number of data points needed during classification as we increase the number of features we use for classification. In your description, be sure to answer: (1) What is the volume of the crust going to converge to as the number of dimensions increases? (2) What concept is this example illustrating? (3) Why is this an important issue in Machine Learning in general. (4) How strategies can we apply when features are uncorrelated and when features are strongly correlated.

The volume of the crust will convergence to 1. That means that in high dimensions, the samples will be sparse and in the corners of the feature space. This concept illustrates the Curse of Dimensionality.

The Curse of Dimensionality is an important concept in ML as it relates the number of samples needed as the number of dimensions increases. The number of samples needed is exponentially larger than the number of dimensions.

If the features are correlated, then we can perform feature selection (e.g., L1-norm or Sequential Backward/Forward Selection) techniques to reduce the feature space dimensionality without loss of significant information.

If the features are uncorrelated, we should collect a lot more data to explain/populate the entire feature space. If no more data is available, then we can perform feature extraction (e.g., PCA or LDA).

6. (10 points) In order to decide whether or not a group of friends should go out to play soccer, let's build a decision tree algorithm. To train the model, consider the data set of features about each time the team has decided to plat, including the outlook conditions (1 means sunny, 2 means overcast and 3 means rainy), temperature (1 means hot, 2 means mild and 3 means cool), humidity (0 for False and 1 for True), and windy (0 for False and 1 for True), along with the final decision about whether or not to go out to play soccer ($y = 0$ for "do not play", $y = 1$ for "play").

outlook	temperature	humidity	windy	play
1	1	1	0	0
1	1	1	1	0
2	1	1	0	1
3	2	1	0	1
3	3	0	0	1
3	3	0	1	0
2	3	0	1	1
1	2	1	0	0
1	3	0	0	1
3	2	0	0	1
1	2	0	1	1
2	2	1	1	1
2	1	0	0	1
3	2	1	1	0

In the case of any ties, we will prefer to predict class $y = 1$ ("play"). For the next steps define $0 \log_2(0) = 0$.

Compute the information gain (entropy function) for each feature. Which feature should be the root node? Show your work.

Entropy for feature "oulook":

$$-\left(\frac{5}{14}\left(\frac{3}{5}\log_2\frac{3}{5}+\frac{2}{9}\log_2\frac{2}{9}\right)+\frac{4}{14}\left(0\log_2 0+\frac{4}{9}\log_2\frac{4}{9}\right)+\frac{5}{14}\left(\frac{2}{5}\log_2\frac{2}{5}+\frac{3}{9}\log_2\frac{3}{9}\right)\right)\approx 0.865$$

Entropy for feature "temperature":

$$-\left(\frac{4}{14}\left(\frac{2}{5}\log_2\frac{2}{5}+\frac{2}{9}\log_2\frac{2}{9}\right)+\frac{6}{14}\left(\frac{2}{5}\log_2\frac{2}{5}+\frac{4}{9}\log_2\frac{4}{9}\right)+\frac{4}{14}\left(\frac{1}{5}\log_2\frac{1}{5}+\frac{3}{9}\log_2\frac{3}{9}\right)\right)\approx 1.022$$

Entropy for feature "humidity":

$$-\left(\frac{7}{14}\left(\frac{1}{5}\log_2\frac{1}{5}+\frac{6}{9}\log_2\frac{6}{9}\right)+\frac{7}{14}\left(\frac{3}{5}\log_2\frac{4}{5}+\frac{3}{9}\log_2\frac{3}{9}\right)\right)\approx 0.820$$

Entropy for feature "windy":

$$-\left(\frac{8}{14}\left(\frac{2}{5}\log_2\frac{2}{5}+\frac{6}{9}\log_2\frac{6}{9}\right)+\frac{6}{14}\left(\frac{3}{5}\log_2\frac{3}{5}+\frac{3}{9}\log_2\frac{3}{9}\right)\right)\approx 0.941$$

For this training data, we should use the feature ”humidity” as it is the one that minimizes the entropy gain.

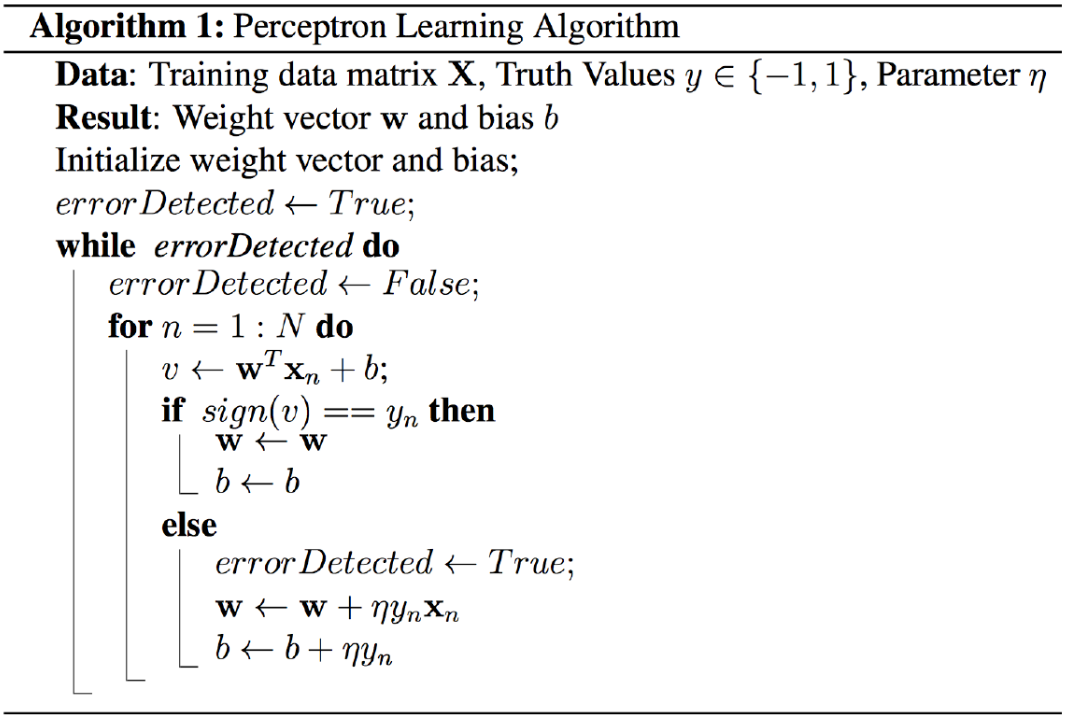
7. (10 points) What are the common issues of decision trees and how can they be solved? Be specific.

   Decision trees are prone to overfitting. There are techniques to help prune the decision tree, which include: define a maximum number of misclassified samples in order to allow to tree to create a new split and control the depth of the tree.

   Another strategy to combat overfitting of a single tree model is to use a collection of Bootstrap decision trees, or a Random Forest. A Random Forest will combine the decision from multiple overfitted learners and retrieve a model with a generalization ability that is better than any individual overfitted tree.

8. (20 points) The following three questions are about the perceptron learning algorithm.

Suppose you have a linearly separable classification problem with training data, $\mathbf{X}$, that has been normalized to have zero mean and unit variance. You would like to use the Perceptron Learning Algorithm to train a classifier. The perceptron learning algorithm requires the learning rate parameter, η, to be set.



Algorithm 1: Perceptron Learning Algorithm

```
Data: Training data matrix $\mathbf{X}$, Truth Values $y \in \{-1, 1\}$, Parameter $\eta$
Result: Weight vector $\mathbf{w}$ and bias $b$
Initialize weight vector and bias;
$errorDetected \leftarrow True$;
while errorDetected do
    $errorDetected \leftarrow False$;
    for $n = 1 : N$ do
        $v \leftarrow \mathbf{w}^T \mathbf{x}_n + b$;
        if $sign(v) == y_n$ then
            $\mathbf{w} \leftarrow \mathbf{w}$
            $b \leftarrow b$
        else
            $errorDetected \leftarrow True$;
            $\mathbf{w} \leftarrow \mathbf{w} + \eta y_n \mathbf{x}_n$
            $b \leftarrow b + \eta y_n$
```

(a) (5 points) How will the perceptron learning algorithm perform if η is set to be ”too large?” (e.g., $\eta = 1000$). Will it converge to the correct answer?

No, it won’t. If the learning rate, η, is too large, the algorithm may diverge.

(b) (5 points) How will the perceptron learning algorithm perform if η is set to be "too small?" (e.g., $0 < \eta < 0.01$)? Will it converge to the correct answer?

It will. But it will take a long time to reach a minima.

(c) (5 points) How will the perceptron learning algorithm perform if η is set to be $\eta < 0$? Will it converge to the correct answer?

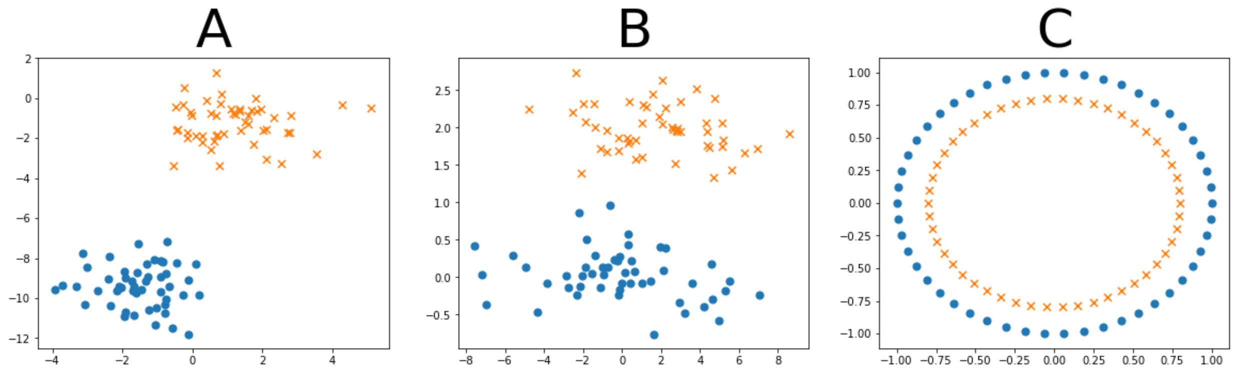
No, it will diverge. Instead of descend to minimize the cost function, it will move in the direction of the gradient, which is pointing towards the largest increase in the cost function.

9. (5 points) Suppose you were segmenting a data set into three classes and wanted to evaluate your results. Would using a ROC curve be an appropriate method for evaluation? Why or why not? What about the overall accuracy measure?

   No, a ROC curve is primarily for a 2-class classification problem. However, we can combine multiple ROC curves for a classification problem with more than 2 classes. ROC curves are a better measure than accuracy, even for a multi-class problem as ROC curves are not sensitive to imbalance classes.

   A more typical measure to summarize results for a multi-class classification is to compute a confusion matrix.

10. (10 points) Consider the following three two-dimensional data sets each containing two clusters of data points (shown with ”circles” and ”crosses”). Suppose you would like to apply Principal Components Analysis (PCA) to reduce the dimensionality of each of these data sets from 2-D to 1-D where the two clusters remain separated in the reduced dimensional data set. For each data set, address each of the following two questions:



(a) (5 points) Will PCA be effective at keeping the two clusters separated in the reduced dimensionality data? Why or why not? If yes, state what characteristics of the data set allow PCA to be effective. If no, state what characteristics of the data set cause PCA to fail.

PCA will be effective at keeping the two clusters separated in the 1-D projection, as the direction of maximum variance is parallel to the y-axis, and in that direction, the clusters will be linearly separable.

PCA will not be effective at keeping the two clusters separated in the 1-D projection, as the direction of maximum variance is parallel to the x-axis, and in that direction, the clusters will overlap.

PCA will not be effective at keeping the two clusters separated in the 1-D projection, as the relationship between features is non-linear. As so PCA is not able to preserve non-linear representations.

(b) (5 points) Can you think of another dimensionality reduction technique that would be successful at reducing the dimensionality of this data set while maintaining (or increase) separation between the two clusters? State the other method and describe why it would be successful.

For data set B, LDA will be a good dimensionality reduction to apply as it will choose the direction of projection that maximize class separability.

Because LDA is also only able to represent linear representations, it will not be useful to data set C. However, we can represent the feature space in a higher dimensional space using an appropriate kernel function (for example, radial basis function or RBF) and then perform PCA or LDA in that transformed space.

11. **Extra credit (5 points)**

    Suppose that you want to learn the the regions of the feature space that belong to different classes (decision surface) using the k-NN classifier from a set of training set with imbalanced classes, in particular: 200 points from class 1 and 10 points from class 2. Further assume that you are using Euclidean distance as your metric and you will tally the votes based on majority vote.

    How would the decision surface look like as you increase the parameter k? How would you solve this issue?

    As the value of k starts to increase, in particular, starting at $k = 21$, we will have always more classes from class 1 than class 2, and with majority vote we will always choose class 1. Even if points from class 1 are much closer in the Euclidean-distance sense. That means that the decision surface will become smoother and in majority belonging to class 1 as k increases.

    An alternative to using *regular* Euclidean distance, we can use the weighted Euclidean distance, where training samples that are closer to the test point will have a larger weight towards test point's assignment.

HONOR STATEMENT

I understand that I am bound to uphold the honor code of the University of Florida. I have neither given nor received assistance on this examination. In addition, I did not use any outside materials on this exam other than the one page of formulas that was allowed.

Sign Your Name: ____

Write the Date: ____

Print Your Name: ____

Turn in your formula sheet with your exam!!!