Take-Home Practice on Machine Learning (Non-Graded)

Quiz 3 (Machine learning)

- What to know conceptually
    - Basics of supervised classification, classification vs. regression, underfitting vs. overfitting, linear separability, etc

    - Logistic regression
    - Support vector machine
    - Random forest

    - Bias and variance
    - Bootstrap
    - Fairness in assessing machine learning methods

- What to know in detail (i.e., problem solving)
    - Linear regression
    - Naïve Bayes classifiers
    - Decision trees

    - Cross validation
    - Confusion matrix, precision, recall
    - RSq. (coefficient of determination)
    - ROC curves and AUC

- Materials for Quiz 3
    - Slides are your primary source of information
    - In-class activities are an excellent source for additional practice
    - Take-home practice and solutions to check your understanding

    - The posted reading is optional
    - There are excellent on-line resources. But these sometimes go beyond what we have covered, so only use them if you needed added explanations. Here are some examples:
        - Understanding AUC-ROC Curves
        - A Tutorial on Fairness in Machine Learning
        - Logistic regression
        - Bias and variance tradeoff
        - and so on

1. Given the following dataset:

X	Y
1	0.5
2	1
4	2
0	0

   a. Fit a linear regression model to the above dataset:

   b. Calculate the R-squared value for the model you computed in (a):

2. Assuming you're using a Decision Tree to classify data from the below dataset, where A, B, and C are features and the Class is a Yes or No.

A	**B**	**C**	**Class**
F	T	F	Yes
F	F	T	No
F	T	F	Yes
T	T	T	No
F	T	T	No

2a. What would the root node be?

2b. Assume you chose node B as the root (this may or may not be the correct answer), what would be the next node to split?

3. Given the below dataset, where P, Q, and R are features and the Class is a Yes or No. Use Naïve Bayes to classify a new datapoint X with features:
   < P = False, Q = True, R = False > as a Yes or No.

P	Q	R	Class
F	T	F	Yes
F	F	T	No
F	T	T	Yes
T	T	F	No
F	T	T	No

4. A fourth feature S was added to the dataset from problem 3, the new dataset is shown below. This feature is numeric. What is $p\ (S = 7 \mid Yes)$? What about $p\ (S = 4.5 \mid No)$?

P	Q	R	S	Class
F	T	F	7	Yes
F	F	T	4	No
F	T	T	3	Yes
T	T	F	9	No
F	T	T	11	No

5. Consider the following observations from an email classifier that marks emails as Spam or Non-Spam:
    - Altogether, the classifier predicted Spam or Non-Spam for 100 emails.
    - The model correctly classified 95 emails: 85 were correctly classified as Non-Spam, and 10 of them were correctly classified as Spam.
    - 5 emails, which were actually Spam, were predicted as Non-Spam.
    - 0 Non-Spam emails were predicted as Spam.

    Answer the following questions based on the above data:

    a) Construct a confusion matrix for this classifier:

    b) What is the precision and recall of this classifier?

    c) What is the accuracy of this classifier? Is using accuracy alone sufficient for determining if this is a good classifier? Justify your reasoning.

6. Cross-validation item

You have a data set of 120 people with 60 juniors and 60 senior students. You are asked to train a decision tree on the data using three-fold cross-validation. Sketch out the training and testing sets.

7. Consider the following values of specificity and sensitivity computed at three decision thresholds, 5, 7, and 9 for a classifier whose output ranges from 1 to 100%.

Threshold	**Sensitivity**	**Specificity**
5	.56	.99
7	.78	.81
9	.91	.42

Plot the ROC curve based on the data above.

Based on the ROC curve you generated, is the AUROC: (a) at chance; (b) slightly above chance; (c) considerably above chance; (d) almost perfect.