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

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:

Solution:

Calculate the average of x and y.

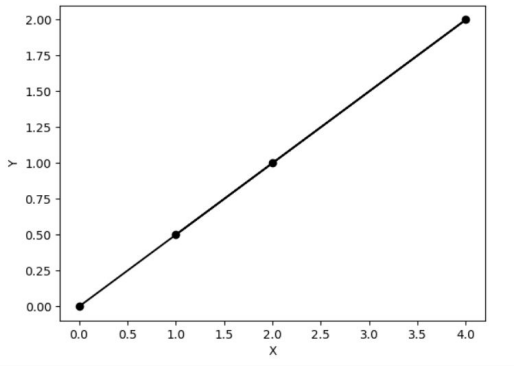
$\bar{y} = 0.875 \quad \bar{x} = 1.75$

Calculate the slope: $m = \frac{\sum_i (x_i - \bar{x}) * (y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = 0.5$

Calculate the intercept: $b = \bar{y} - m * \bar{x} = 0$

So our linear regression model is **y = 0.5*x + 0**

We can check the quality of the model fit with the R-squared value (calculated below in (b)) and also by looking at a plot of the data and our model:



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

Solution:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \qquad SS_{res} = \sum_i (y_i - f_i)^2 \qquad SS_{tot} = \sum_i (y_i - \bar{y})^2$$

$$SS_{res} = (0.5 - 0.5)^2 + (1 - 1)^2 + (2 - 2)^2 + (0 - 0)^2 = 0$$

$$SS_{tot} = (0.5 - 0.875)^2 + (1 - 0.875)^2 + (2 - 0.875)^2 + (0 - 0.875)^2 = 2.1875$$

$$R^2 = 1 - \frac{0}{2.1875} = 1$$

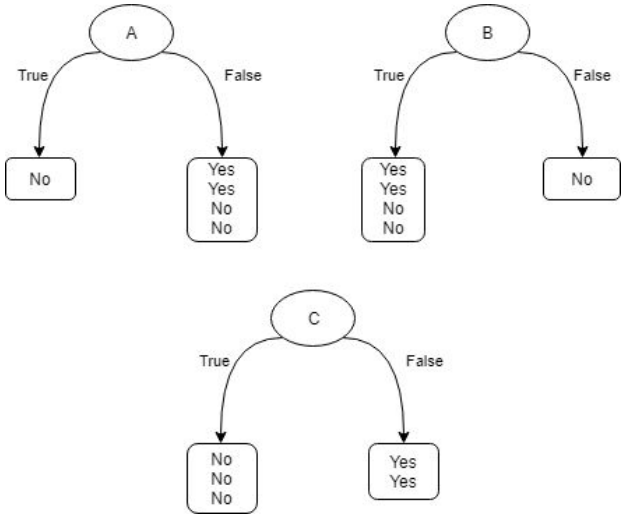
This makes sense from looking at the plot of the data and fit line since the model perfectly fits the data the error residual should be zero.

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?

Solution:
The three possible splits are shown below:



Initial tree entropy:
- Yes,Yes,No,No,No
- H = -⅖*log(⅖,2)-⅗*log(⅗,2) = 0.97 bits

Information gain from A:
- True
  - No
  - Information = 0 bits
- False
  - Yes, Yes, No, No
  - Information = 1 bits
- Average information (bits x occurrence) = 0*⅕ + 1*⅘ = 0.80
- Information gain = tree entropy - average information gain = 0.97 - 0.8 = 0.17 bits gained

Information gain from B:

- True
    - Yes, Yes, No, No
    - Information = 1 bits
- False
    - No
    - Information = 0 bits
- Average information (bits x occurrence) = 1*4/5 + 0*⅕ = 0.80
- Information gain = tree entropy - average information gain = 0.97 - 0.8 = 0.17 bits gained

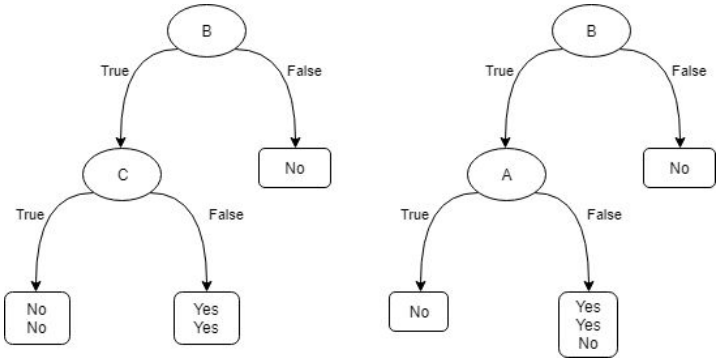
Information gain from C:
- True
    - No, No, No
    - Information = 0 bits
- False
    - Yes, Yes
    - Information = 0 bits
- Average information (bits x occurrence) = 0*3/5 + 0*2/5 = 0.0
- Information gain = tree entropy - average information gain = 0.97 - 0.0 = 0.97 bits gained

Summary:
Information gain from splitting on A: 0.17 bits
Information gain from splitting on B: 0.17 bits
Information gain from splitting on C: 0.97 bits

The root node should be C since it yields the greatest information gain when split.

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?

Solution:
The two possible splits are shown below.



Initial tree entropy:
- Yes,Yes,No,No
- H = -2/4*log(2/4,2)-2/4*log(2/4,2) = 1.0 bits

Information gain from A:
- True

        - No
        - Information = 0 bits
    - False
        - Yes, Yes, No
        - Information = 0.92 bits
    - Average information (bits x occurrence) = 0*1/4 + 0.92*3/4 = 0.69
    - Information gain = tree entropy - average information gain = 1.0 - 0.69 = 0.31 bits gained

Information gain from C:
- True
    - No, No
    - Information = 0 bits
- False
    - Yes, Yes
    - Information = 0 bits
- Average information (bits x occurrence) = 0*2/4 + 0*2/4 = 0.0
- Information gain = tree entropy - average information gain =1.0 - 0.0 = 1.0 bits gained

Summary:
Information gain from splitting on A: 0.31 bits
Information gain from splitting on C: 1.0 bits

The root node should be C since it yields the greatest information gain when 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

Solution:
P(Yes) = 2/5
P(No) = 3/5
P(P=F | Yes) = 2/2
P(P=F | No) = 2/3
P(Q=T | Yes) = 2/2
P(Q=T | No) = 2/3
P(R=F | Yes) = 1/2
P(R=F | No) = 1/3

P(X | Yes) = P(P=F | Yes) * P(Q=T | Yes) * P(R=F | Yes) = 2/2 * 2/2 * 1/2 = 0.5
P(X | No) = P(P=F | No) * (P(Q=T | No) * P(R=F | No) = 2/3 * 2/3 * 1/3 = 0.15
P(X | Yes) * P(Yes) / (P(X | Yes) * P(Yes)+P(X | No) * P(No)) = 0.5 * ⅖ /(0.5*⅖ + 0.15*⅗) = 69%
P(X | No) * P(No) / (P(X | Yes) * P(Yes)+P(X | No) * P(No)) = 0.15 * ⅗ /(0.5*⅖ + 0.15*⅗) = 31%

The Naïve Bayes classifier would classify X as a Yes since 69% > 31%.

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 | Yes)*? What about *p (S = 4.5 | 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

Solution:
Assuming a normal distribution, the probability of any value can be computed from the probability density function below.

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

S mean | No = 8
S stdev | No = 2.94
S mean | Yes = 5
S stdev | Yes = 2

$$p(S = 4.5 \mid No) = \frac{1}{\sqrt{2\pi} * 2.94} e^{-\frac{(4.5-8)^2}{2*2.94}} = 0.01$$

$$p(S = 7 \mid Yes) = \frac{1}{\sqrt{2\pi} * 2} e^{-\frac{(7-5)^2}{2*2}} = 0.03$$

Normalized:
p(S = 4.5 | No) = 0.01/0.04 = 25%
p(S = 7 | Yes) = 0.03/0.04 = 75%

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:

    Solution:

	Confusion Matrix	
	Actual Spam	Actual Not-Spam
Predicted Spam	10 (*TP*)	0 *(FP)*
Predicted Not-Spam	5 (*FN)*	85 *(TN)*

    b) What is the precision and recall of this classifier?
    Solution:
    Precision = TP / (TP + FP) = 10/10 = 1

    Recall = TP/ (TP + FN) = 10/15 = ⅔ = 0.67

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

    Solution: Accuracy: (TP + TN) / (TP + FP + TN + FN) = 95/100 = 0.95. Accuracy is not a good measure of evaluating the classifier as there is a large imbalance in the dataset. The classifier may not generalize well to spam emails and may result in a lot of false negatives as the recall is low. Using accuracy alone will not give a complete picture of how the classifier is doing.

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.

Solution

Randomly divide the data into three folds but use stratification such that there is an equal number of juniors and seniors per fold.

Fold 1: 20 juniors; 20 seniors
Fold 2: 20 juniors; 20 seniors
Fold 3: 20 juniors; 20 seniors

Training set A: Fold 1, Fold 2
Testing set A: Fold 3

Training set B: Fold 1, Fold 3
Testing set B: Fold 2

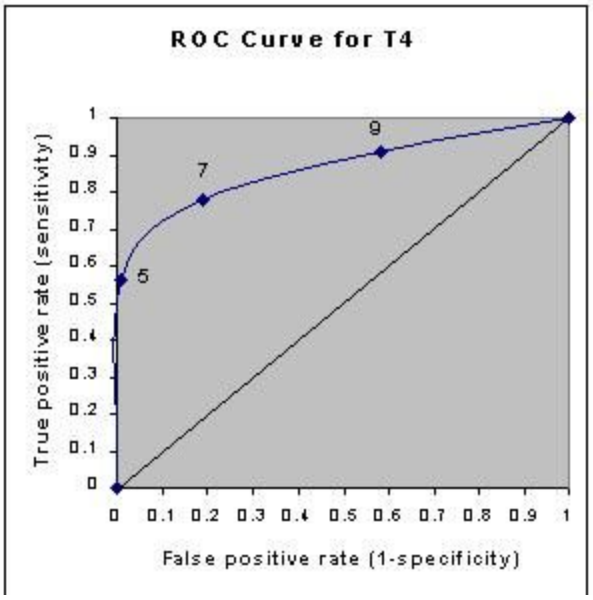
Training set C: Fold 2, Fold 3
Testing set C: Fold 1

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.

Solution



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

Solution
(c) - considerably above chance (chance AUROC = 0.5) - we estimate it to be around 0.8 (rough estimate)