Chloe Quinto
CPE 695 WS
3/19/20
I pledge my honor that I have abided by the Stevens Honor System - Chloe Quinto

Question 1: Please answer the following questions related to Machine Learning Concepts

1) Explain what the bias-variance trade-off is? Describe a few techniques to reduce bias and variance respectively
    a) The Bias Problem: the hypothesis space made available by a particular classification method does not include sufficient hypotheses.
    b) The Variance Problem: the hypothesis space made available is too large for the training data, and the selected hypothesis may not be accurate to unseed data
    c) To reduce bias:
        i) Increase hypothesis space (i.e. model complexity)
    d) To Reduce Variance
        i) Resample
            (1) Bias of each tree is the same as full model but higher variance
            (2) Averaging many trees decreases variance without increasing bias
            (3) Theoretically the more trees, the less variance (if no computation limit)
2) What is K-fold cross-validation? Why do we need it?
    a) K-fold cross-validation randomly splits the training set into 10 distinct subsets called folds and then it trains and evaluates the decision tree model 10 times, picking a different fold for evaluation every time and training on the other 9 folds. We need it because it is a procedure used to estimate the skill of the model on new data. Specifically, Cross-Validation is to use a limited sample in order to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of a model.

Question 2: Assume the following confusion matrix of a classifier. Please compute its

1) Precision
2) Recall
3) F1 Score

		Predicted results	
Actual values		Class 1	Class 2
	Class 1	50	30
	Class 2	40	60

$$p = \frac{TP}{TP+FP} = \frac{50}{50+40} = \frac{5}{9} = 0.555$$
$$r = \frac{TP}{TP+FN} = \frac{50}{50+30} = \frac{5}{8} = 0.625$$

$$F_1 = \frac{2pr}{p+r} = \frac{2*0.555*0.625}{0.555+0.625} = 0.5879$$

Question 3:

1) Build a Decision Tree using the following training instances (using information gain approach)

Day	*Outlook*	*Temperature*	*Humidity*	*Wind*	*PlayTennis*
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes

Entropy

- Out of 10 instances, 6 have said yes

  $p_{yes} = -(6/10)log_2(6/10) = 0.442$

  $p_{no} = -(4/10)log_2(4/10) = 0.528$

  $H(s) = p_{yes} + p_{no} = 0.442 + 0.528 = 0.97$

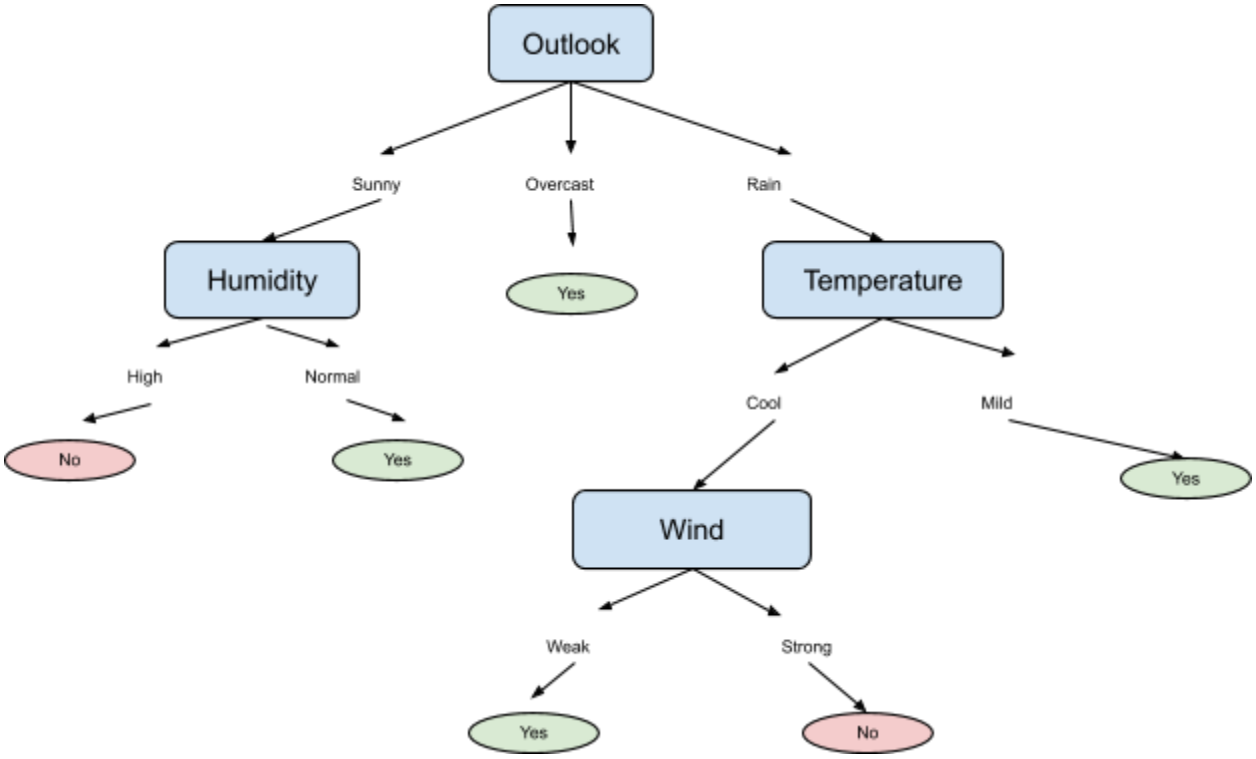
For every feature, calculate the entropy

- Outlook
  - $E(outlook = sunny) = -(1/4)log_2(1/4) = 0.5$
  - $E(outlook = overcast) = -(2/2)log_2(2/2) = 0$
  - $E(outlook = rain) = -(3/4)log_2(3/4) = 0.311$
  - $I(outlook) = \frac{4}{10}(0.5) + \frac{2}{10}(0) + \frac{4}{10}(0.311) = 0.3244$
  - $Gain(outlook) = 0.97 - 0.3244 = 0.6456$
- Temperature
  - $E(temperature = hot) = -(\frac{1}{3})log_2\frac{1}{3} = 0.528$
  - $E(temperature = mild) = -(\frac{2}{3})log_2(\frac{2}{3}) = 0.389975$
  - $E(temperature = cool) = -(\frac{3}{4})log_2(\frac{3}{4}) = 0.3112$
  - $I(temperature) = \frac{3}{10}(0.528) + \frac{3}{10}(0.389975) + \frac{4}{10}(0.3112) = 0.39987$

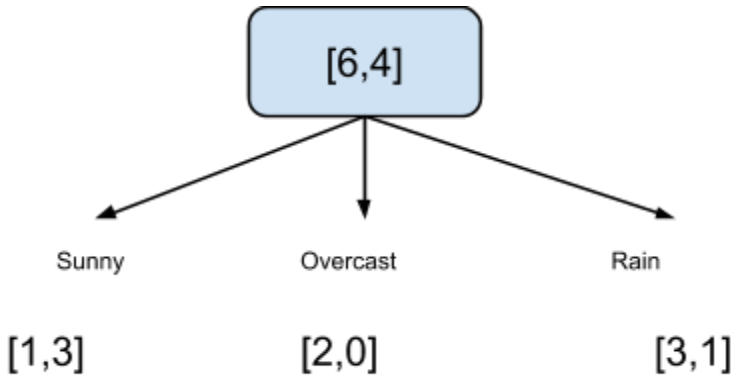
- $Gain(temperature) = 0.97 - 0.39987 = 0.57013$

- Humidity
  - $E(humidity = high) = -(\frac{2}{5})log_2\frac{2}{5} = 0.528$
  - $E(humidity = normal) = -(\frac{4}{5})log_2\frac{4}{5} = 0.257$
  - $I(humidity) = \frac{5}{10}(0.528) + \frac{5}{10}(0.257) = 0.3925$
  - $Gain(humidity) = 0.97 - 0.3925 = 0.5775$
- Wind
  - $E(wind = weak) = -(\frac{5}{7})log_2\frac{5}{7} = 0.3467334$
  - $E(wind = strong) = -(\frac{1}{3})log_2\frac{1}{3} = 0.5283$
  - $I(wind) = \frac{7}{10}(0.346) + \frac{3}{10}(0.5283) = 0.4012$
  - $Gain(wind) = 0.97 - 0.4012 = 0.5687$

Attribute	**Gain**
Outlook	0.6456
Temperature	0.571
Humidity	0.5775
Wind	0.57



2) Decide the p-value (i.e. pchance) of the root node using Chi-square test



$\widehat{p_1} = \frac{6}{6+4} * 4 = 2.4$
$\widehat{p_2} = \frac{6}{6+4} * 2 = 1.2$
$\widehat{p_3} = \frac{6}{6+4} * 4 = 2.4$

$\widehat{n_1} = \frac{4}{6+4} * 4 = 1.6$
$\widehat{n_2} = \frac{4}{6+4} * 2 = 0.8$
$\widehat{n_3} = \frac{4}{6+4} * 4 = 1.6$

$$Q = \sum_i \frac{(p_i - \widehat{p_i})^2}{\widehat{p_i}} + \frac{(n_i - \widehat{n_i})^2}{\widehat{n}} = \frac{(1-2.4)^2}{2.4} + \frac{(3-1.6)^2}{1.6} + \frac{(2-1.2)^2}{1.2} + \frac{(0-0.8)^2}{0.8} + \frac{(3-2.4)^2}{2.4} + \frac{(1-0.8)^2}{0.8} = 3.575$$

$$Degrees\ of\ Freedom = 3 - 1 = 2$$

$$p_{chance} = 0.1673$$

Question 4:

In ensemble learning, there are several popular fusion methods for class label type classifiers e.g. majority vote, weighted majority vote, and naive bayes methods. Assuming we have 3 classifiers, and their predicted results are given in the table 1. The confusion matrix of each classifier is given in table 2. Please give the final decision using Naive Bayes as the fusion method:

Table 1 Predicted results of each classifier

Sample x	Result
Classifier 1	Class 1
Classifier 2	Class 1
Classifier 3	Class 2

Table 2 Confusion matrix of each classifier

i) Classifier 1

	Class1	Class2
Class1	40	10
Class2	30	20

ii) Classifier 2

	Class1	Class2
Class1	20	30
Class2	20	30

iii) Classifier 3

	Class1	Class2
Class1	50	0
Class2	40	10

The outputs of individual classifiers are: 1 1 2

$\widehat{P}(\omega_1 \mid d_{1,1}(x) = 1) = \frac{40}{70}$ $\widehat{P}(\omega_1 \mid d_{2,1}(x) = 1) = \frac{20}{40}$ $\widehat{P}(\omega_1 \mid d_{3,2}(x) = 1) = \frac{0}{10}$ **Class 1: 0.2857**	$\widehat{P}(\omega_2 \mid d_{1,1}(x) = 1) = \frac{30}{70}$ $\widehat{P}(\omega_2 \mid d_{2,1}(x) = 1) = \frac{20}{40}$ $\widehat{P}(\omega_2 \mid d_{3,2}(x) = 1) = \frac{10}{10}$ Class 2: 0.214