Name: Qian Cai
NUID: 001389278

Data Structures and Algorithms
INFO 6205, Wed
Homework 9
Due: November 23, 2019

Put all your java, compiled class files and documentation files into a zip file named Homework9.zip and submit it via the dropbox on the blackboard before the END of
due date. Put your name on all .java files. There will be a short Quiz on this homework.

1. Describe these concepts:

Predictive accuracy?

Answer: Number of correct classification/ Total number of test cases

What is the key in building a decision tree?

Answer: The key to building a decision tree - which attribute to choose in order to branch.

What is evaluation method?

Answer: Evaluation methods are the criteria for evaluating the success of a program or project.

Describe ALL evaluation classification methods and their differences?

Answer:

1.holdout method: The holdout method is the simplest kind of cross validation. The data set is separated into two sets, called the training set and the testing set. The function approximator fits a function using the training set only. Then the function approximator is asked to predict the output values for the data in the testing set (it has never seen these output values before). The errors it makes are accumulated as before to give the mean absolute test set error, which is used to evaluate the model.
2.n-fold cross-validation: The available data is partitioned into n equal-size disjoint subsets.
Use each subset as the test set and combine the rest n-1 subsets as the training set to learn a classifier.
3.Leave-one-out cross-validation: The available data is partitioned into n equal-size disjoint subsets.
Each fold of the cross validation has only a single test example and all the rest of the data is used in training.
4.Validation set: the available data is divided into three subsets, a training set, a validation set and a test set. A validation set is used frequently for estimating parameters in learning algorithms.
Differences:
"holdout method" is usually preferable to the residual method and takes no longer to compute.
"n-fold cross-validation": This method is used when the available data is not large. The advantage of this method is that it matters less how the data gets divided.
"Leave-one-out cross-validation" method is used when the data set is very small. Each fold of the cross validation has only a single test example and all the rest of the data is used in training.

Scoring and Ranking Method, How does it work?

Answer:

Answer: Scoring is related to classification. We are interested in a single class. Instead of assigning each test instance a definite class, scoring assigns a probability estimate (PE) to indicate the likelihood that the example belongs to the positive class. After each example is given a PE score, we can rank all examples according to their PEs.

Lift Analysis Curve

Answer:

Answer: A lift curve can be drawn according how many positive examples are in each bin. This is called lift analysis.

How does Naive Bayes is different from other Evaluation Classification methods?

Answer:

Advantages: Easy to implement. Very efficient. Good results obtained in many applications

Disadvantages:
Assumption: class conditional independence, therefore loss of accuracy when the assumption is seriously violated (those highly correlated data sets)

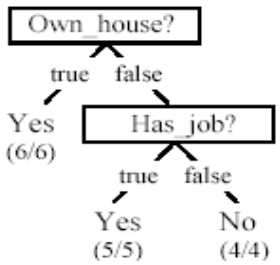
2. Consider the sequences x = GCCCTAGCG and y = GCGCAATG. Assume that the match score is +1, and the mismatch is -1, and gap penalties is -2.

A) Fill out the dynamic programming table for a global alignment between x and y.
B) Draw arrows in the cells to store traceback information.
C) What is the score of the optimal global alignment and what alignment(s) achieves this score?

A)

		G	C	G	C	A	A	T	G
	0	-2	-4	-6	-8	-10	-12	-14	-16
G	-2	1	-1	-3	-5	-7	-9	-11	-13
C	-4	-1	2	0	-2	-4	-6	-8	-10
C	-6	-3	0	1	1	-1	-3	-5	-7
C	-8	-5	-2	-1	2	0	-2	-4	-6
T	-10	-7	-4	-3	0	1	-1	-1	-3
A	-12	-9	-6	-5	-2	1	2	0	-2
G	-14	-11	-8	-5	-4	-1	0	1	1
C	-16	-13	-10	-7	-4	-3	-2	-1	0
G	-18	-15	-12	-9	-6	-5	-4	-3	0

B)
G->CG->GCG-> GCGA-> GCGAT-> GCGATC-> GCGATCC-> GCGATCCC-> GCGATCCCG
G->GT-> GT- -> GT-A-> GT-AA-> GT-AAC-> GT-AACG-> GT-AACGC-> GT-AACGC G
C)
GCCCTAGCG
GCGCAA-TG
Score:5*1+(-2)*1+(-1)*3=0



$$Accuracy = \frac{\text{Number of correct classifications}}{\text{Total number of test cases}}$$

3. Consider this decision tree. Is this an optimal decision tree? What is the key to decision Tree learning?

Data: Loan application data

Task: Predict whether a loan should be approved or not.

Performance measure: accuracy

Answer: Yes. This is an optimal decision tree
The key to building a decision tree - which attribute to choose in order to branch.

4. Consider n-fold cross-validation method:

a) How does algorithm work for Training and Test

https://towardsdatascience.com/why-and-how-to-cross-validate-a-model-d6424b45261f

b) Explain this code. Compile and Run if you can, analyze the results.

https://github.com/haifengl/smile/blob/master/core/src/main/java/smile/validation/CrossValidation.java

a)Split the entire data randomly into k folds (value of k shouldn't be too small or too high, ideally we choose 5 to 10 depending on the data size). The higher value of K leads to less biased model (but large variance might lead to overfit), where as the lower value of K is similar to the train-test split approach we saw before.

Then fit the model using the K — 1 (K minus 1) folds and validate the model using the remaining Kth fold. Note down the scores/errors.

Repeat this process until every K-fold serve as the test set. Then take the average of your recorded scores. That will be the performance metric for the model.

b) split the data into k folds

n=0, then repeat n+1

Use kth fold as test data, and use using the remaining fold as train data.
While n<k output the result

5. Naive Bayes Classification,

https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c

Naive Bayes three types of classifiers. What are they, explain.

How does it work? Provide an example showing results

1.Multinomial Naive Bayes:

This is mostly used for document classification problem, i.e whether a document belongs to the category of sports, politics, technology etc. The features/predictors used by the classifier are the frequency of the words present in the document.

2.Bernoulli Naive Bayes:

This is similar to the multinomial naive bayes but the predictors are boolean variables. The parameters that we use to predict the class variable take up only values yes or no, for example if a word occurs in the text or not.

3.Gaussian Naive Bayes:

When the predictors take up a continuous value and are not discrete, we assume that these values are sampled from a gaussian distribution.

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)}$$

The variable y is the class variable(play golf), which represents if it is suitable to play golf or not given the conditions. Variable X represent the parameters/features.

$$X = (x_1, x_2, x_3, ..., x_n)$$

X is given as,

Here x_1,x_2….x_n represent the features, i.e they can be mapped to outlook, temperature, humidity and windy. By substituting for X and expanding using the chain rule we get,

$$P(y|x_1, ..., x_n) = \frac{P(x_1|y)P(x_2|y)...P(x_n|y)P(y)}{P(x_1)P(x_2)...P(x_n)}$$

Now, you can obtain the values for each by looking at the dataset and substitute them into the equation. For all entries in the dataset, the denominator does not change, it remain static. Therefore, the denominator can be removed and a proportionality can be introduced.

$$P(y|x_1, ..., x_n) \propto P(y) \prod_{i=1}^{n} P(x_i|y)$$

In our case, the class variable(y) has only two outcomes, yes or no. There could be cases where the classification could be multivariate. Therefore, we need to find the class y with maximum probability.

$$y = argmax_y P(y) \prod_{i=1}^{n} P(x_i|y)$$

Using the above function, we can obtain the class, given the predictors.

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Frequency Table		
Weather	No	Yes
Overcast		4
Rainy	3	2
Sunny	2	3
Grand Total	5	9

Likelihood table				
Weather	No	Yes		
Overcast		4	=4/14	0.29
Rainy	3	2	=5/14	0.36
Sunny	2	3	=5/14	0.36
All	5	9		
	=5/14	=9/14		
	0.36	0.64		

Will he play when the weather is sunny?
P(Yes | Sunny) = P(Sunny | Yes) * P(Yes) / P (Sunny)
P (Sunny |Yes) = 3/9 = 0.33, P(Sunny) = 5/14 = 0.36, P(Yes)= 9/14 = 0.64
P (Yes | Sunny) = 0.33 * 0.64 / 0.36 = 0.60
0.6>0.5 it is true

6. Consider the following example using Naive Bayes classifier:
https://www.codingame.com/playgrounds/6734/machine-learning-with-java---part-5-naive-bayes
a) Describe Example
b) Run program, describe outputs
c) Take the Java code and build it in your environment

a) To find whether an email is ham or spam, we have training samples listed below.

Text	Category
Congratulation you are selected	ham
Congrats you won lottery	spam
travel for free	spam
selected for credit cards	spam
very Good	ham
Good night	ham
lottery	spam

In order to determine which category the sentence "you won lottery" belongs to, we calculate the probability of the sentence "you won lottery" is ham and the probability that its spam. Then take the largest one.

P (A | B) = (P (B | A) * P(A))/ P(B)
P (ham| you won lottery) = (P (you won lottery | ham) * P(ham))/ P (you won lottery)
P (you won lottery | ham) = P (you | ham) * P (won | ham) * P (lottery | ham)
P (you won lottery | spam) = P (you | spam) * P (won | spam) * P (lottery | spam)
P (you | ham) = (1 + 1)/ (7 + 15)
P (you | spam) = (1 + 1)/ (10 + 15)
"You won Lottery" belongs to spam category
b)

Correctly Classified Instances	**7**	**100**	**%**
Incorrectly Classified Instances	**0**	**0**	**%**
Kappa statistic	**1**		
Mean absolute error	**0.1378**		
Root mean squared error	**0.1444**		
Relative absolute error	**28.0006 %**		
Root relative squared error	**29.1716 %**		
Total Number of Instances	**7**		

the expression for the input data as per alogorithm is The independent probability of a class

----

spam	**0.5555555555555556**
ham	**0.4444444444444444**

The probability of a word given the class

----

	spam	**ham**
Congrats	**0.07407407407407407**	**0.043478260869565216**
cards	**0.07407407407407407**	**0.043478260869565216**
credit	**0.07407407407407407**	**0.043478260869565216**
for	**0.11111111111111109**	**0.043478260869565216**
free	**0.07407407407407407**	**0.043478260869565216**
lottery	**0.11111111111111109**	**0.043478260869565216**
selected	**0.07407407407407407**	**0.08695652173913045**
travel	**0.07407407407407407**	**0.043478260869565216**
won	**0.07407407407407407**	**0.043478260869565216**
you	**0.07407407407407407**	**0.08695652173913045**
Congratulation	**0.037037037037037035**	**0.08695652173913045**
Good	**0.037037037037037035**	**0.13043478260869565**
are	**0.037037037037037035**	**0.08695652173913045**
night	**0.037037037037037035**	**0.08695652173913045**
very	**0.037037037037037035**	**0.08695652173913045**

c)
The code is in H9_6
Run_NaiveBayesDemoTest

Tests passed: 1 of 1 test – 435ms

```
/Library/Java/JavaVirtualMachines/jdk1.8.0_131.jdk/Contents/Home/bin/java ...

** Naive Bayes Evaluation with Datasets **

Correctly Classified Instances           7               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0.1378
Root mean squared error                  0.1444
Relative absolute error                 28.0006 %
Root relative squared error             29.1716 %
Total Number of Instances                7

 the expression for the input data as per alogorithm is The independent probability of a class
--------------------------------------
spam    0.5555555555555556
ham 0.4444444444444444

The probability of a word given the class
-----------------------------------------
    spam    ham
Congrats    0.07407407407407407 0.043478260869565216
cards   0.07407407407407407 0.043478260869565216
credit  0.07407407407407407 0.043478260869565216
for 0.11111111111111109 0.043478260869565216
free    0.07407407407407407 0.043478260869565216
lottery 0.11111111111111109 0.043478260869565216
selected    0.07407407407407407 0.08695652173913045
travel  0.07407407407407407 0.043478260869565216
won 0.07407407407407407 0.043478260869565216
you 0.07407407407407407 0.08695652173913045
Congratulation  0.037037037037037035    0.08695652173913045
Good    0.037037037037037035    0.13043478260869565
are 0.037037037037037035    0.08695652173913045
night   0.037037037037037035    0.08695652173913045
very    0.037037037037037035    0.08695652173913045

{0 ?,1 1,2 1,3 1}
spam
```