CSE 160: Introduction to Data Science
Solutions to Exam #2

1 December 2017

Your name: ____

As with all exams, I suggest reading through the entire exam, and then answering the questions that seem easiest first. Since no electronic devices are permitted, you may leave any arithmetic expressions in an unsimplified form.

1. True or False. **Circle** the best single answer for each problem.

   (a) 2 points TRUE / FALSE : The value of the AUC can be greater than 1. *FALSE*

   (b) 2 points TRUE / FALSE : A “false positive” is when a classifier correctly assigns a negative classification. *FALSE*

   (c) 2 points TRUE / FALSE : ROC graphs are advantageous because anyone can interpret them with ease. *FALSE*

   (d) 2 points TRUE / FALSE : While topic modeling can be informative, it can take a lot of time for a computer to perform. *TRUE*

   (e) 2 points TRUE / FALSE : Expected Value is a valuable tool when making business decisions because it can predict how much money could be earned or lost if certain actions are taken. *TRUE*

   (f) 2 points TRUE / FALSE : The AUC is useful when a single number is needed to summarize performance, or when nothing is known about the operating conditions. *TRUE*

   (g) 2 points TRUE / FALSE : It is more important to have good data scientists than it is to have good management of the data science team. *FALSE - DS4B Ch 13*

   (h) 2 points TRUE / FALSE : Selenium enables R to extract information from interactive websites. *TRUE*

   (i) 2 points TRUE / FALSE : The naive Bayes assumption is that events are conditionally independent, and thus can be multiplied together to estimate the probability of a combination of events. *TRUE DS4B Ch 9*

   (j) 2 points TRUE / FALSE : Text is an example of unstructured data. *TRUE - DS4B Ch 10 p252*

   (k) 2 points TRUE / FALSE : With topic models, we perform supervised classification of documents. *FALSE - DS4B Ch 10*

   (l) 2 points TRUE / FALSE : A profit curve is the plot of expected profit at each possible threshold that can be applied to a ranking classifier. *TRUE - DS4B Ch 8 p212*

   (m) 2 points TRUE / FALSE : The rbind() function in R determines whether the R object is bound or not. *FALSE - R Twitter lecture - row binding*

   (n) 2 points TRUE / FALSE : The lift of a classifier is calculated as the classifier’s performance divided by the performance of a randomly guessing classifier. *TRUE - DS4B Ch 8 p220*

   (o) 2 points TRUE / FALSE : A shapefile contains the points and contours of a geographical region. *TRUE - R Mapping lecture*

2. Multiple choice. **Circle the best single answer**.

   (a) 3 points When word order is important, which text mining strategy should we use?

   1. Bag of Words
   2. N-grams ***
   3. Named Entity Extraction
   4. Topic Models
   5. All of the above are valid models that respect word order

   (b) 3 points If a model is created for the classification of students as freshmen, sophomores, juniors, or seniors, what would be the maximum dimensions of the confusion matrix?

   1. 4x4 ***
   2. 4x2
   3. 4x1
   4. 2x2
   5. It depends on what data the model is built on

   (c) 3 points Confusion matrices can be used to analyze the performance of which of the following models:

   1. decision trees
   2. Naive Bayes
   3. k-Nearest Neighbors
   4. all of the above ***
   5. 1) and 3) only

   (d) 3 points In the context of preparing for textual analysis, what is meant by “stop words?”

   1. words that have the same meaning as the word “stop”
   2. words such as and, or, and the ***
   3. words specific to the topic of a corpus, such as “rebound” or “offense” when talking about a basketball game
   4. the words at the end of each sentence
   5. none of the above

   (e) 3 points Which of these organizations would have the most challenge in applying supervised predictive modeling?

   1. A business school that wants to start a new Master’s degree program in Business Analytics, and would like to estimate the likely number of applicants. ***
   2. A grocery store that is trying to identify which of its loyalty-card-carrying customers will spend more than $100 next month.
   3. A city government that is trying to predict which neighborhoods will see the most new businesses open up next quarter.
   4. An online marketing company that wants to estimate the number of clicks that the ads it serves will receive when shown to a particular population.
   5. All of the above are equally challenging.

   (f) 3 points Which of the following terms will have the lowest IDF score in a typical (general purpose) corpus?

1. bug
2. car
3. she ***
4. spaghetti
5. vertebrae

(g) $\boxed{\text{3 points}}$ Which of the following is not a step in generating an ROC curve?

1. Sort the test set by the model predictions
2. Start with the cutoff = min(prediction) ***
3. Decrease cutoff, and then count the number of true positives and false positives
4. Calculate the TP rate and the FP rate
5. Plot current number of TP/P as a function of the current FP/N

(h) $\boxed{\text{3 points}}$ The points on a model's ROC graph

1. represent the performance of different thresholds ***
2. represent different rankings of examples
3. represent the cost of different classifications
4. all of the above
5. none of the above

(i) $\boxed{\text{3 points}}$ In a marketing environment, the expected benefit of not targeting is typically:

1. a negative value
2. zero ***
3. a positive value
4. all of the above
5. none of the above

(j) $\boxed{\text{3 points}}$ Which of the following is NOT an advantage of the Naive Bayes classifier?

1. very simple implementation
2. efficient in terms of storage space
3. efficient in terms of computation time
4. performs well in many real-world applications
5. generally accurate class probability estimation ***

(k) $\boxed{\text{3 points}}$ The bag of words model

1. treats every document as a collection of individual words
2. ignores grammar, word order and sentence structure
3. is a straightforward representation that is inexpensive to generate
4. all of the above ***
5. none of the above

3. Short answer

(a) $\boxed{\text{4 points}}$ Give two distinct examples of words that can be "stemmed" and what their stems would be.

(b) $\boxed{\text{2 points}}$ If the probability of a burrito having cheese on it is 7/8 and the probability of a burrito having beans on it is 1/4, what is the probability of a burrito having both cheese and beans on it assuming that these two events are independent of each other?

*7/8 * 1/4*

(c) $\boxed{\text{2 points}}$ R provides a mechanism to allow computation to continue, even when an error (e.g., some kind of exception) has occurred. What is that function's name?

try()

(d) $\boxed{\text{2 points}}$ Name a learning algorithm we have seen that can operate incrementally (i.e., can be updated easily without re-examining the rest of the training data).

Naive Bayes / kNN

(e) $\boxed{\text{3 points}}$ Write a valid for loop in R that sums the numbers 1 through 100.

(f) $\boxed{\text{3 points}}$ Given the following confusion matrix, calculate precision, recall and accuracy

		Actual	
		p	n
	Y	5	7
Predicted	N	4	6

precision= 5/12 recall= 5/9 accuracy= 11/22

4. TF-IDF

(a) $\boxed{\text{3 points}}$ Define TF (using formulas or words)

TF(t,d) is the (normalized) count of the number of times term t appears in document d.

(b) $\boxed{\text{3 points}}$ Define IDF (using formulas or words)

IDF(t) is 1 + log(number of docs in corpus / number of docs containing t)

(c) $\boxed{\text{3 points}}$ Define TF-IDF as a combination of TF and IDF as defined above.

TFIDF(t,d) = *TF(t,d)* × *IDF(t)*

5. Expected value framework. Note that Part (b) builds on Part (a), and Part (c) builds on Part (b). (Show your work to get partial credit.)

(a) $\boxed{\text{4 points}}$ In online advertising, a CPC advertisement is one in which the advertising network is paid by the advertiser only when the user clicks on the ad ("cost-per-click"). If you know that 30% of advertising attempts fail because of browser ad blockers, and only 1% of impressions are clicked on, what is the expected income of displaying a page with a CPC ad to the advertising network when it earns $1 per click?

Start: $EV = p_R(x) \times v_R + (1 - p_R(x)) \times v_{NR}$

So... $EI = (.7 * .01) \times 1 + 0 = .007$

(b) $\boxed{\text{3 points}}$ If it costs the advertising network $.002 every time it attempts to show an advertisement (e.g., for bandwidth, computers, etc.), what is the expected value of displaying a page with an ad?

So... $EV = (.7 * .01) \times (1 - .002) + (1 - .7 * .01) \times (-.002)$

$= .007 \times .998 - .993 \times .002$

$= .006986 - .001986 = .005$

$= EI - .002$

(c) 3 points A second kind of pay model in advertising is called CPA ("cost-per-action"). Assuming the advertising network above also received an additional two dollars for every purchase made by someone who clicked, and the rate of purchases per click was 1 out of 20, what is the company's new expected value for displaying a page with an ad?

Now... $EV = .005 + .007/20 \times 2 = .0057$

6. Opinion, full points regardless of which answers you choose.

(a) 1 point In general, the exam questions were

1. too easy
2. just right
3. too hard

(b) 1 point In this exam, I

1. had plenty of time
2. had just enough time
3. did not have enough time to finish the exam questions