Sample Exam #2

Answers are not provided. Discuss the questions (and answers) on Piazza.

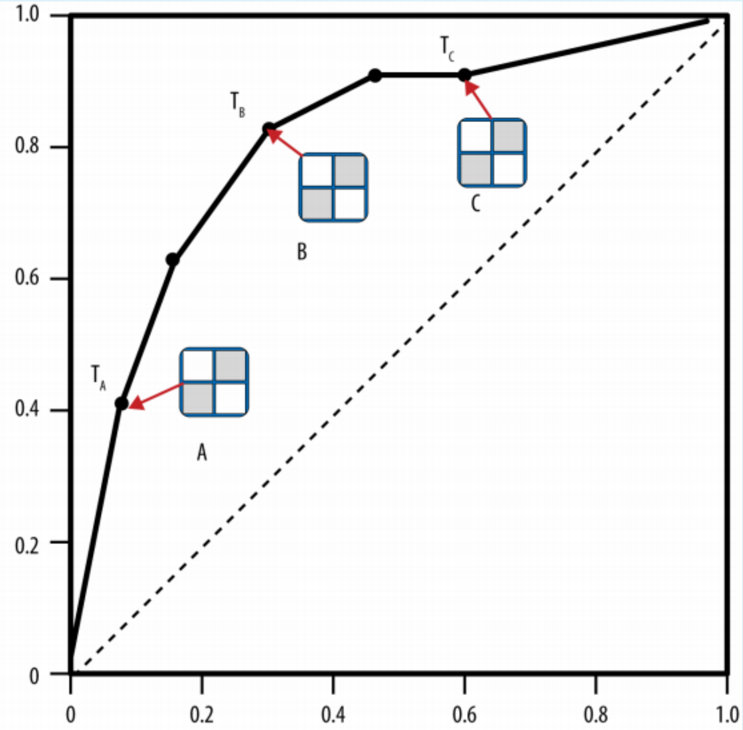
Question **1**

a, 3 pts) Explain why an ROC curve always starts from (0, 0) and ends at (1, 1).

b, 3 pts) Suppose point C is (0.6, 0.9), and our result comes from 100 positive and 100 negative instances, then what is the confusion matrix at point C.

c, 2 pts) Compute the corresponding F1 and Accuracy at point C.

d, 2 pts) Explain how we use an ROC graph to evaluate and compare two models.



Question **2**

If we have a situation with three possible outcomes, what is the general equation to calculate expected value for this situation?

Question **3**

Probability is an important interpretation for a model's prediction. Please explain how the following models provide a probability of a class for a testing instance Xi. You can either use math formulation or natural language.

Hint: you may wish to start from how the model gives a prediction for a testing instance, and then explain how the probability is generated from the prediction.

a, 6pts) A decision tree trained for 3-class classification
b, 6pts) A logistic regression model trained for binary classification

c, 6pts) A k-NN model for 3-class classification

Question **4**

A ROC graph is a two-dimensional plot of a classifier with what on the x and y axes?

Question **5**

Given below is a scenario for Training Error (TE) and Validation Error (VE) for a machine learning algorithm. You want to choose a hyperparameter (H) based on TE and VE. Which value of H is best?

Select one:

H=1 resulting in TE=105, VE=90
H=2 resulting in TE=200, VE=85
H=3 resulting in TE=250, VE=96
H=4 resulting in TE=105, VE=85
H=5 resulting in TE=300, VE=100

Question **6**

The cost and benefits of each decision are needed to calculate the overall expected value of a model.

Question **7**

For any classifier, its confusion matrix is always a 2 by 2 matrix.

Question **8**

By definition, a profit curve cannot go negative.

Question **9**

Suppose we accidentally swap the columns with the rows when generating a confusion matrix, then all the correct cases are still on the diagonals.

Question **10**

A certain private university has 80,000 living alumni, from which it depends on donations to provide scholarships, support operations and grow its endowment. In a typical year, 3% of its alumni will randomly make a donation with an average value of $60. However, over the years, the university has developed a semi-automated targeted marketing process that classifies alumni into three groups: A, B, and C. Group A is the largest, representing 95% of alumni, and is not targeted. Group B is 4.5% of alumni and is targeted via student phone calls, soliciting donations. Group C (.5% of alumni) gets personalized treatment with university representatives making phone calls throughout the year and holding occasional in-person meetings with them.

The cost to pay students to make phone calls to group B is $10,000 per year, and they reach 80% of the alumni they attempt, and 50% of those they reach will donate, on average, $50 each.

The cost to employ fundraising staff to work with group C is $300,000 per year. 5% of the alumni they attempt to contact are deceased, but they reach the rest of them. 20% of those alumni will donate, on average, $25,000 each. (The others do not donate this year.)

a: 4 pts) If the university were to stop all targeted solicitations, after all expenses, how much money do they expect to raise next year? (Show work as well as final amount)

b: 14 pts) If the university continues as it has in past years, after all expenses, how much money do they expect to raise next year? (Show work as well as final amount)

c: 2 pts) Describe one assumption made above that might negatively affect fundraising next year?

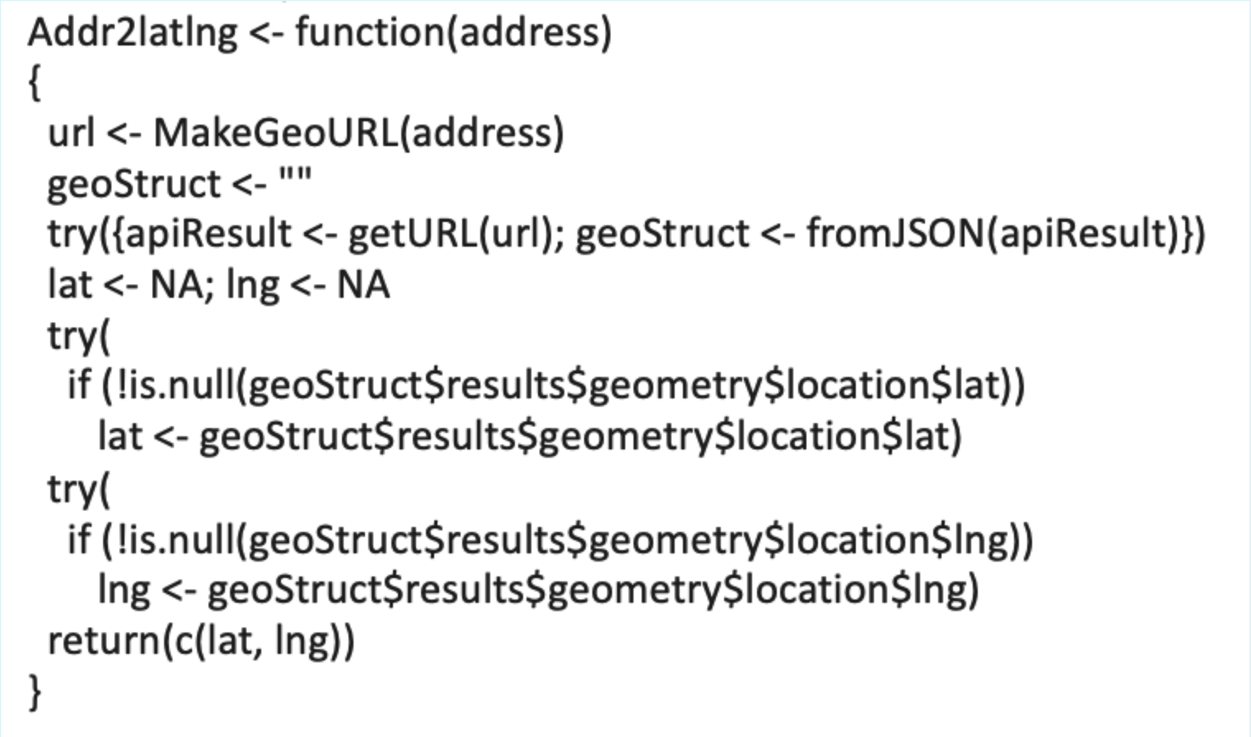
Question **11**

In data science, a good method often depends on intuitive assumptions. Please explain what are the key assumptions for the following methods (that is, one key assumption each):

1. Naive Bayes
2. Bag of Words
3. IDF (inverse document frequency)

Question **12**

The following is code similar to what we have shown in the lecture that uses a web service for geocoding. Explain what the benefit of calling function try() is in this scenario.



```
Addr2latlng <- function(address)
{
  url <- MakeGeoURL(address)
  geoStruct <- ""
  try({apiResult <- getURL(url); geoStruct <- fromJSON(apiResult)})
  lat <- NA; lng <- NA
  try(
    if (!is.null(geoStruct$results$geometry$location$lat))
       lat <- geoStruct$results$geometry$location$lat)
  try(
    if (!is.null(geoStruct$results$geometry$location$lng))
       lng <- geoStruct$results$geometry$location$lng)
  return(c(lat, lng))
}
```

Question **13**

Name two things discussed in our primary text that can help a business sustain a competitive advantage in their respective space and explain why and how they are effective?

Question **14**

We saw evidence lifts when talking about building models based on Facebook likes. At Meta, you are responsible for building a model to predict whether a person is a high technology consumer or not (as it is a desirable segment of the market). Based on a sample of 1 million users, you know that only 20% of them are high tech consumers. Of the 20,000 people who have visited Apple's website, two-thirds of them are considered high tech consumers.

Calculate the evidence lift for a potential high tech consumer visiting Apple's website.

Question **15**

Suppose we are doing bag-of-words text classification on a document.

(a. 6 points) The raw input is a single string containing the text of each document. Describe the necessary steps to go from the raw input to TF-IDF feature form. You may describe the pipeline step by step. Note that we are not looking for R code.
(b. 4 points) To ensure the features we have learned on the training documents can be correctly applied on the testing documents of various length, what is needed when computing TF-IDF?

Question **16**

The point of analytical engineering is to:

Select one:

a. Promote thinking about problems data analytically
b. Develop complex solutions by addressing every possible contingency
c. Convert all students into engineers
d. Analyze engineers
e. None of the others are appropriate answers

Question **17**

p(C|E) stands for:

Select one:

a. The probability of neither C nor E happening
b. The joint probability of both C and E happening
c. The probability of an event E happening given C happened
d. The probability of an event C happening given E happened
e. None of the above

Question **18**

Negative instances classified as positive are referred to as what?

Question **19**

Which of the following is not one of the three factors that characterize the errors a model could make?

Select one:

a. Inherent Randomness
b. Human error
c. Bias
d. Variance

Question **20**

Term frequency (TF) is document specific.

Question **21**

With naïve Bayes, we assume that the predictor variables are conditionally independent of one another given the response value.

Question **22**

An incremental learner is an induction technique that can update its model one training example at a time.

Question **23**

Naive Bayes Classifier classifies a new example by estimating the probability that the example doesn't belong to each class and reports the class with the lowest probability.

Question **24**

When the events are independent, p(AB) = p(A) / p(B).

Question **25**

Using n-grams greatly decreases the size of the feature set.

Question **26**

Like its location, the font size of each word in a word cloud is random

Question **27**

Stemming removes suffixes and transforms plural nouns to their singular forms

Question **28**

Correlation implies that there must be some cause and effect relationship between the two variables.

Question **29**

Bayes' Rule calculates the probability of a hypothesis given some evidence (the posterior probability)

Question **30**

Stop words are typically very common words and include functional words like prepositions

Question **31**

Firms are more and more considering whether and how they can obtain competitive advantage from their data and/or their data science capability.

Question **32**

Data scientists often struggle under a management that sees the potential benefit of predictive modeling, but does not have enough appreciation for the process to invest in proper training data or in proper evaluation procedures.

Question **33**

Managers should be able to expect data scientists to have deep expertise in business solutions.

Question **34**

A new web search engine called “Lehigh-Finds” is at a competitive disadvantage against Google and Bing because Lehigh-Finds lacks the click-through data to learn models of what answers are better than others.

Question **35**

Ensembles have been observed to improve generalization performance in many situations

Question **36**

Referring to term frequency, the importance of a term in a document should decrease with the number of times that term occurs.

Question **37**

The conditional independence assumption allows us to calculate the probability of an event as the product of all of the probabilities of each component of the event

Question **38**

In this course, we talked about various methods related to K. Explicitly explain what K means in the following methods.
a: (2 pts) K-means
b: (2 pts) K-nearest neighbor
c: (2 pts) K-fold cross-validation

Question **39**

Initialization of K-Means will affect the clustering results.

Question **40**

Given points (-2, -4), (5, -10), (10, 2), (-1, 2) belonging to the C1 cluster and (0,0) to the C2 cluster. At what location is the new centroid of the cluster C1 using k-means clustering?