CS 59970 - Intro to Data Science
Midterm Exam
March 25, 2019

Exam Answer Key (Version 1)

This is a 90 minute exam. Please use your time wisely.

Multiple Choice: ____ / 58
Short Answer: ____ / 36

Academic dishonesty is prohibited in The City University of New York. Penalties for academic dishonesty include academic sanctions, such as failing or otherwise reduced grades, and/or disciplinary sanctions, including suspension or expulsion.

Exam 8891

Part I: Multiple Choice. (2 points per question)

Circle the best answer(s).

1. Which of the following is ***NOT*** true of regular expressions?
    a) They are helpful in cleaning raw text data.
    b) They are useful in dropping outliers from continuous data.
    c) They can be used to generate features from loosely structured text data.
    d) They rely of sequences of characters to specify search patterns.

2. Which of the following is true of the R-squared value? *Select all that apply.*
    a) It represents the share of variation in the target explained by the features.
    b) A model is not good if it has an R-squared of 0.35.
    c) It is always between 0 and 1.
    d) It represents the share of variation in the features explained by the target.

3. Which of the following, when done by itself and without other calculations, could help you determine whether your data is normally distributed?
    a) Calculating the mean of the data.
    b) Calculating the standard deviation of the data.
    c) Plotting a histogram of the data.
    d) Calculating the median of the data.

4. Which of the following is the best example of continuous data?
    a) The level of education achieved by potential voters.
    b) The political affiliation of potential voters.
    c) The marital status of potential voters.
    d) The annual income earned by potential voters.

5. Which of the following is true of the normal distribution? *Select all that apply.*
    a) It occurs often in nature.
    b) It is often skewed.
    c) It is an important part of the Central Limit Theorem.
    d) It is symmetric.

6. Which of the following is ***least helpful*** in describing the distribution of a population of data?
    a) Median
    b) Mean
    c) Percentiles
    d) Number of observations

The *law of large numbers* tells us that the average of the results obtained from a large number of trials should be [ANSWER TO QUESTION XX], and will tend to [ANSWER TO QUESTION XX].

7.
    a) heteroskedastic
    b) identical among most trials
    c) decreasing with the number of trials
    d) close to the expected value

8.
    a) diminish in significance as more trials are performed
    b) become closer to the expected value as more trials are performed
    c) be uniformly distributed as more trials are performed
    d) become more difficult to expect as more trials are performed

9. According to the central limit theorem, when independent random variables are added:
    a) their properly normalized sum tends toward a uniform distribution
    b) their properly normalized sum tends toward a normal distribution
    c) their properly normalized sum tends toward zero
    d) their variance tends to increase

10. Which of the following is a python package commonly used for statistical modeling and machine learning:
    a) matlab
    b) stata
    c) sklearn, aka sci-kit learn
    d) learnable

11. Which of the following is a python package commonly used for data handling:
    a) handlr
    b) pandas
    c) dataSlayer
    d) giraffes

12. Which of the following is a python package commonly used for data visualization:
    a) showpy
    b) stata
    c) meowviz
    d) matplotlib

13. Which of the following offers access to a wide range well-documented demographic and socioeconomic indicators through a publicly accessible API?
    a) Facebook Newsfeed
    b) Kaggle
    c) Opendata.org
    d) U.S. Census Bureau

14. Which of the following was ***NOT*** a famous or well-publicized application of data science we discussed in class?
    a) Target predicting pregnancy using customers' purchase history
    b) Facebook's use of technology to eavesdrop on users through Instagram
    c) Tracking gold exports from Venezuela using publicly available flight data
    d) LinkedIn's creation of an algorithm to suggest "people you may know"

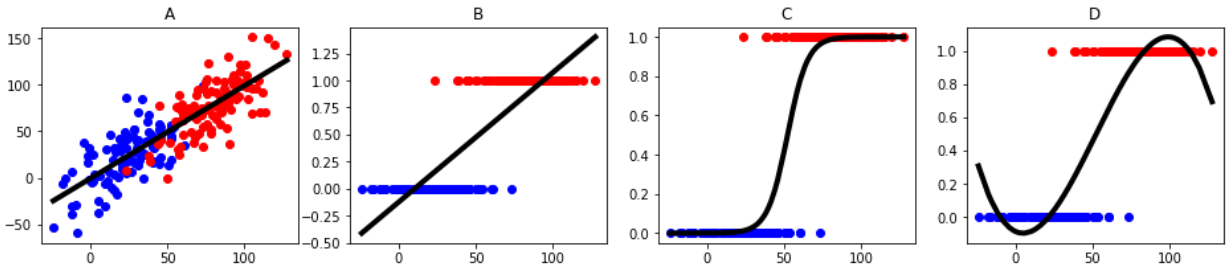
15. Which of the following are assumptions of linear regression? *Select all that apply*.
    a) The data are linear in form.
    b) The underlying data are normally distributed.
    c) The underlying data are uniformly distributed.
    d) The errors have constant variance.

16. In statistics, a sequence or a vector of random variables is homoskedastic if:
    a) all variables in the sequence or vector have increasing variance.
    b) one predictor variable can be linearly predicted from the others with a substantial degree of accuracy.
    c) all variables in the sequence or vector are drawn randomly.
    d) all variables in the sequence or vector have the same finite variance.

17. In statistics, a sequence or a vector of random variables exhibits multicollinearity if:
    a) all variables in the sequence or vector have increasing variance.
    b) one predictor variable can be linearly predicted from the others with a substantial degree of accuracy.
    c) all variables in the sequence or vector are drawn randomly.
    d) all variables in the sequence or vector have the same finite variance.

18. Which of the following are true of the *Netflix Prize*? *Select all that apply.*
    a) The prize awarded $1 million to the team able to predict which ten streaming series would be the most watched in 2017.
    b) The prize awarded $1 million to the first team able generate a 10 percent improvement to Netflix's content recommendation algorithm.
    c) Netflix never implemented the winning algorithm, citing technical complexity and changing business priorities.
    d) The winning algorithm has since been adopted by data science teams at LinkedIn and Target.

19. Which of the follow figures below depicts the output of logistic regression?



    a) A
    b) B
    c) C
    d) D

20. Which of the following statements are true of linear regression ***but not*** logistic regression.
    a) The target is a continuous variable.
    b) The target is a binary variable.
    c) The features may contain continuous variables.
    d) The model is linear in form.

21. Which of the following statements are true of logistic regression ***but not*** linear regression.
    a) The target is a continuous variable.
    b) The target is a binary variable.
    c) The features may contain continuous variables.
    d) The model is linear in form.

22. Which of the following best describes the function of DAGs, or Directed Acyclic Graphs, in the *Airflow* package:
    a) A collection of models, from which the most predictive is selected.
    b) A collection of all of the tasks you want to run, organized in a way that reflects their relationships and dependencies.
    c) A shortest path algorithm problem with a depth-first solution.
    d) A control system that Airbnb utilizes to switch between heating and air-conditioning.

23. Airflow, which was created by Airbnb, and Luigi, which was created by Spotify, are both examples of which of the following? *Select all that apply*.
    a) Machine learning models.
    b) Frameworks for automating and monitoring ETL tasks.
    c) Open source projects that have since been adopted by several other prominent tech firms.
    d) Distributed event streaming platforms capable of handling large volumes of events.

24. Which of the following are examples of data scientists applying the principle of Occam's Razor? *Select all that apply*.
    a) Alicia seeks more data to train her model.
    b) Brian decides to deploy a simple model that performs well rather than a significantly more complex model the performs marginally better on his training data.
    c) Chinu forces her model to use just the 20 most predictive features out a possible 110 input to the model training process.
    d) Dao uses a deep learning system to forecast housing prices for the 20 largest metropolitan areas in the U.S.

25. Which of the following is a ***NOT*** metric for evaluating the performance or fit of linear regression?
    a) Recall
    b) R-squared
    c) P-value
    d) Coefficient

26. Which of the following are true with regard to evaluating stock price data? *Select all that apply*.
    a) It is relatively simple to predict future stock market prices.
    b) It is best to look at absolute levels of stock prices when comparing performance between different companies, because those are what is bought and sold.
    c) It is best to look at daily, weekly, or monthly price changes when comparing performance between different companies, because that reflects what is earned from investment.
    d) Individual stock price changes are likely to be correlated with broader market movements, though to varying degrees.

Figure

OLS Regression Results

Dep. Variable:	rent	R-squared:	0.698
Model:	OLS	Adj. R-squared:	0.698
Method:	Least Squares	F-statistic:	2310.
Date:	Tue, 19 Mar 2019	Prob (F-statistic):	0.00
Time:	21:10:45	Log-Likelihood:	-44013.
No. Observations:	5000	AIC:	8.804e+04
Df Residuals:	4994	BIC:	8.808e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-843.2363	59.763	-14.110	0.000	-960.399	-726.074
bedrooms	-325.8528	35.599	-9.153	0.000	-395.642	-256.063
size_sqft	5.7808	0.078	74.086	0.000	5.628	5.934
floor	50.2127	2.197	22.856	0.000	45.906	54.520
min_to_subway	-9.4129	4.333	-2.172	0.030	-17.908	-0.918
has_doorman	228.7888	54.955	4.163	0.000	121.053	336.525

Omnibus:	1054.785	Durbin-Watson:	2.019
Prob(Omnibus):	0.000	Jarque-Bera (JB):	10191.796
Skew:	0.727	Prob(JB):	0.00
Kurtosis:	9.841	Cond. No.	2.77e+03

27. Based on the figure above, which of the following numbers below best reflects the overall goodness of fit of the model?
    a) 5000
    b) 0.097
    c) 0.698
    d) 5.7808

28. Based on the figure above, which value reflects the likelihood that the coefficient for **`min_to_subway`** is equal to zero?
    a) 0.000
    b) 4.333
    c) -9.4129
    d) 0.030

29. Based on the figure above, how much more in monthly rent should we expect to pay in a building with a doorman?
    a) $5.78
    b) $228.79
    c) $4.16
    d) $300

Part II: Short Answer. (6 points per question)

Answer each of the following in a few short sentences.

1. Many firms where information plays a large role have critical ETL processes. Give an example of one of these processes, and identify the extract, transform, and load elements of the pipeline.

   Spotify would like to produce customized "Daily Mix" playlists for all of their customers on a daily basis. They first collect all of their customer's listening histories from their user activity database (extract), and then run a collaborative filtering algorithm to produce personalized recommendations based on the listening history of similar users (transform). Finally, they load these personalized recommendations into a new database so they can be pushed to customer's apps (load).

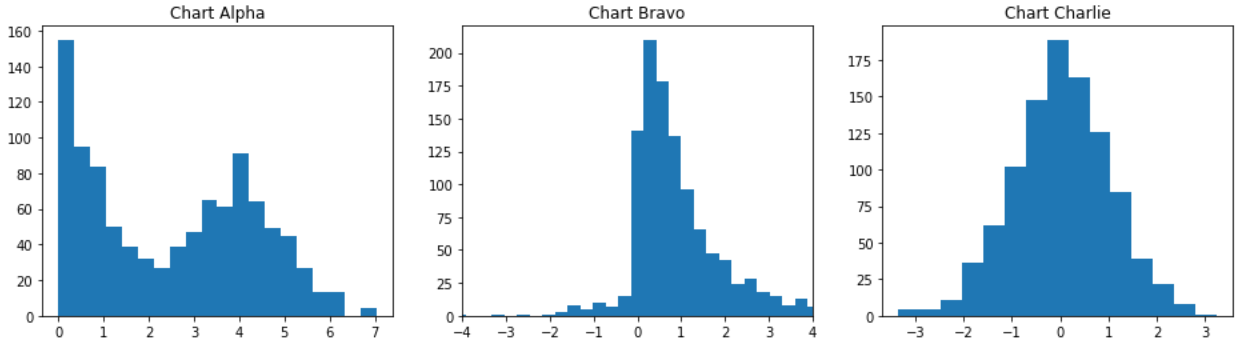
2. Suppose you are a data scientist at Airbnb and would like to create a model to suggest the most appropriate nightly rents for properties in a particular area.
    a. What type of model would you use? Why?
    b. What would the target be?
    c. What features would you use? (name at least two)

    a. A linear regression would allow us to predict a continuous target using a wide range of information about the listings.
    b. Our target is nightly rent (as measured by past and current listing rents).
    c. Our features might include number of bedrooms, size in square feet, and distance to public transit, the city center, or nightlife options.

3. George Box famously said all models are wrong, but some are useful.
    a. Why are all models wrong?
    b. What makes some models useful?
    c. Give an example of a model that is wrong, but useful.

    a. All models account for some degree of measurement error stemming from the inherent randomness in the data gathering and sampling process.
    b. While all models have some degree of error, models that are more often right than not are often quite useful in planning for the future.
    c. Weather forecasts are quite often off by a degree or two, but quite convenient for planning whether to wear a coat or bring an umbrella with you.

4. Which of the following charts is most likely to be drawn from normally distributed data? For each of the charts, describe why the chart does or does not appear to display normally distributed data?



- **Chart Alpha is bimodal (it has two peaks) and thus unlikely to be drawn from a normal distribution.**
- **Chart Bravo is unimodal, but heavily skewed to the right, and thus also unlike to be drawn from a normal distribution.**
- **Chart Charlie is symmetric and evenly distributed in a bell curve, and is indeed drawn from a random sample of a normal distribution.**

5. Explain the difference between regression and classification models and give an example of a modeling problem (ie, using ***X*** to predict ***Y,*** where ***X*** and ***Y*** are real life examples) for each.

   Regression analysis estimates the conditional expectation of the dependent variable given the independent variables. Classification is the problem of identifying to which of a set of categories a new observation belongs.

   An example of a regression problem is predicting NYC rents using apartment size and distance to the subway.

   An example of a classification problem is classifying an email as a spam or not spam.

6. In class we said scraped HTML code could be considered either structured data or unstructured data, depending on your approach.
    a. Explain the difference between structured and unstructured data.
    b. In what way could scraped HTML code be considered unstructured data?
    c. In what way could scraped HTML code be considered structured data?

    Structured data is data that follows a well defined model. It is easy to use for analysis and typically well documented. Unstructured data follows little or no data model, and requires substantial processing for use in data.

    Scraped HTML code could contains large amounts of unstructured text, which could be useful when structured, but requires substantial processing and thus could be considered unstructured.

    HTML code is itself, highly structured, containing a number of different element tags. These tags give the data inherent structure.