The University of Western Ontario

Computer Science 2035B

Instructor: Dr. David Champredon

Final Examination
Take-Home Version

Handed out: April 7, 2020 at 00:01AM

Due Date: April 20, 2020 at 11:55PM

Marking Scheme

- This exam consists of 7 questions worth a total of 100 points.
- This exam comprises 22% of your *overall* mark for this course.
- There is an additional *relative* 10% bonus point for this exam if your *entire* work is submitted before April 16, 2020, 11:55PM.
  For example, if your grade for this exam was 80 points, you could earn an extra 8 points. There is no cap, so the maximum mark is potentially 110 points for entire work submitted that early.
- This is an exam, so no late work will be accepted. A late submission will be marked zero.

GOOD LUCK!

Important

- Submit your assignment on OWL in the form of *well commented* Matlab script files in the "Assignments" section. The file names *must* follow this convention: `Final_STUDENTNUMBER_EXn.m` where `STUDENTNUMBER` is your 9-digit student number and `n` represents the exercise number (`n` is `1` to `7`). Functions (if any) can either be in the same script ("all-in-one" style) or in separated scripts (with the script name the same as the function name).
- Make sure your code runs. **Your grade will be based on what the grader can run. A program that does not run (i.e., stops because of any error) will be graded zero**. Make sure you submit all the necessary files such that the grader can run your programs. This includes the **csv data files** (the ones provided with this exam). Unlike assignments, there will be no exception or second chance.
- Your submission must reflect you own work. **Programs that are suspiciously similar will be graded zero**.
- You are encouraged to submit regularly to OWL. This will avoid unnecessary last-minute stress.

(7 points) 1. **Speed of Light.** An astrophysics laboratory has purchased a new set of instruments to connect to their telescope. The scientists performed a set of independent experiments to measure the speed of light with the new instruments to verify if they are well calibrated. The speed of light measurements, in meter per second (m/s), are recorded in the file `speed-light.csv`. The exact theoretical value for the speed of light is 299,792,458 m/s.

Are the instruments well calibrated? Justify your answer with a Matlab code that implements a rigorous statistical analysis and a short explanation.

(7 points) 2. **Algorithm.** Write a Matlab function named `find_pos_occ` that returns the number of occurrences and the positions of a single character in a text. For example, the text
`I have an apple in my bag.  You can take it!`
contains 6 occurrences of the character `a` at the positions 4, 8, 11, 24, 33 and 37.
You can *only* use any of the following Matlab commands in your algorithm:
`for, while, if, then, sum(), size(), length(), zeros()`.

(12 points) 3. **Decathlon.** The results of a decathlon where 33 athletes competed are saved in the file `deca.csv`. The first ten columns indicate the performance for each of the ten events (100 meters run, long jump, etc.). The last (11th) column is the total score calculated for each athlete from his 10 performances .

a) Create a 2-by-5 panels figure that plots the athletes' total score against the performance of the event, for all events.

b) Create a 10-row table named `cor_event_score` where the first column is the name of the event and the second column is the correlation between the total score and the numeric performance of the event. The rows should be sorted in descending order from the highest absolute value of the correlation to the lowest.

c) Perform a principal component analysis on this dataset. Provide a short explanatory paragraph and a figure that illustrate your interpretation of the PCA.

(16 points) 4. **Influenza Evolution.** The file `h3n2.csv` contains $n = 950$ genetic sequences of the H3N2 strain of influenza viruses that have circulated among human since 1968. The first column represents the name of the genetic sample and the second column its molecular sequence expressed with amino acids. Each letter of the molecular sequence represents an amino acid. The third column is the year of the sample collection. Each sequence has 566 amino acids. We refer to the position of an amino acid simply as its position in the genetic sequence. For example, in the sequence `NGTMVK`, the amino acid `G` is in position 2. We want to focus this analysis only on the amino acids that are between positions 100 and 500 (inclusive). We define the distance between two sequences as the number of amino acids they differ by. For example, the sequences `NGTMVK` and `NGAMTK` have a distance equal to 2 because their amino acids in positions 3 and 5 differ.

a) Calculate (in a $n \times n$ matrix) the pairwise distances between all sequences.

b) Perform a classical multi-dimensional scaling (MDS) in dimension 2. Create a scatter plot of the projected data points, colouring each projected point by the decade of its collection year (for example, the 1970s is the decade for the years 1970, 1971, ..., 1979). Make a legend to identify the decades.

c) Antigenic drift is a kind of genetic variation in viruses resulting from the accumulation, over time, of mutations in the virus genes that code for virus-surface proteins that antibodies of humans recognize. This results, year after year, in new strains of influenza virus that "look" different from the strains of previous years (they are more "distant") making it easier for the changed virus to spread throughout a partially immune population. Does the MDS performed in b) illustrate the antigenic drift of influenza? Explain with one short paragraph.

(14 points) 5. **Ice creams and Weather.** An ice cream truck is a commercial vehicle that serves as a mobile retail outlet for ice cream. The owner of an ice cream truck has noticed that, unsurprisingly, her sales are better during sunny warm days. She wants to better understand this relationship between the weather and her business. Her daily sales revenues (in dollars) are presented in the file `sales.csv`. Daily records of temperature (in Celsius) and rainfall (in mm) for the 2019 summer season were downloaded from a weather website to the file `weather.csv` for the location where she usually sells ice cream. Note that the ice cream truck owner does not operate every day because of various reasons (sickness, technical problems with the truck, etc.).

a) Perform a multivariate linear regression that models the ice cream sales revenues as a function of the two weather variables. Make one figure of your choice that illustrates this linear regression (possibly with multiple panels).

b) The 5-day weather forecast for next week is:

Day	**Temperature (^{o}C)**	**Rainfall (mm)**
Monday	32	0
Tuesday	31	5
Wednesday	26	10
Thursday	23	15
Friday	23	2

What is the total expected revenues for the next 5 days (assuming the owner works every day)? What is the 90% confidence interval for your estimate?

(22 points) 6. **Spam Emails.** A charity has contacted you to ask if you could help them filtering the hundreds of spam emails they receive everyday. Spams affects directly the quality of the service offered by the charity because they crowd out important and urgent emails received from people in need. One of your friend thinks spams can be detected by calculating the frequency of certain words and characters in the text of an email. Your friend developed a program that calculates the frequencies of 48 selected keywords and 9 other variables that look at other various metrics from the content of the email. Hence, in total, there are 57 metrics extracted from a given email. The charity gave you a random sample of 1,000 emails they received last month, and your friend ran the program on those emails. Then, your friend read through all of the 1,000 emails and annotated each email to indicate if it was indeed a spam or not. The result of this hard work is saved in the file `spam-train.csv`, where the first 57 columns represent the various metrics calculated by your friend's program, and the last (58th) column is the spam annotation: `1` if the email was a spam, `0` else.

a) Based on this "training" dataset presented in `spam-train.csv`, develop a predictive model based on a logistic regression coupled with a ROC analysis, that classifies emails as spam or not. The charity is willing to accept that not more than 2 out of 100 non-spam emails can be wrongly classified as "spam".

b) Now that your predictive model is developed, your friend has retrieve new emails received yesterday, ran the 57 metrics on them and saved the results in the file `spam-test.csv`. Run you predictive model on this new data set and classify each email as a spam or not. What is the proportion of spams that were identified with your model in this new data set?

c) A software company has approached the charity, claiming they have a new state-of-the-art software that can detect spams like never before. The director of the charity is tempted to buy this software but hesitates because of its hefty price. The software company has run its state-of-the-art program on the same training set `spam-train.csv` as you did. The software gives a numerical score to an email: the higher the score, the more likely the email is a spam. The scores of the 1,000 emails of the training dataset are saved in `spam_comp.txt`. The director asks you if the software company does better than the method you and your friend provided from your benevolent (free) work. Answer the director with a short paragraph that explains your comparative analysis along with one single figure of your choice.

(22 points) 7. **Bike Sharing.** A city put in place a bike sharing system a year ago. Through this system, users are able to easily rent a bike from a particular station and return it back at another station (possibly the same). There are complaints that some stations often have no bike available to rent. The logistics to make sure there are enough bikes available at all renting stations is complex and partially based on the duration of the bike ride for each user. Municipal employees have noticed that the ride duration tends to be longer when the weather is nice. If this is true, the municipal staff that moves bikes to empty stations could plan its activity in advance, based on weather forecasts, to improve bike availability. The manager of the bike sharing program wants to be sure that weather influences the bike ride duration and hires you as a data analyst consultant to study this.

The manager provides a dataset of 2,000 bike rides randomly selected over the last year in the file `bike_trips.csv` as well as the file `bike_stations.csv` that translates in English the bike stations names from numerical codes. You also have access to weather data for the city in the file `bike_weather.csv`. The file `bike_INFO.txt` contains important additional information about those three files.

a) Merge the information from all three datasets (about bike rides, station names and weather) in a single table that *must* have the following format (note that the codes for weather and stations are replaced by their "names"):

day	station_start	station_end	duration	weather
1	BotanicalGardens	MainStreetSouth	56.78	Sunny
1	MontagueStreet	MontagueStreet	12.34	Storm
1	BakerStreet	AdelaideStreet	8.76	Storm
1	TrainStation	BotanicalGardens	6.31	Light Rain
2	ShoppingMall	CravenAvenue	69.31	Sunny
...	...	...	...	...

The table above is illustrative and does not represent the values of the actual data contained in the file. The variable `duration` is the duration of the bike ride in minutes.

Hint: this question involves several steps of joining tables.

b) Produce and display a table that summarizes the average bike ride duration for each of the four types of weather. Create a well-annotated boxplot that shows the distribution of bike ride durations by weather type.

c) Conduct a rigorous statistical analysis that determines if the bike ride durations differs with the type of weather. Write a short paragraph that summarizes your analysis.