**CSCI 5622 ML Exam (Fall 2023)**

**Honor Statement:**

**I, \_\_Amith Nandivada\_\_, understand that I will do this exam alone and without the help of other humans. I understand exactly what it means to “do my own work” and I will do my own work. If I fail to do this and/or my work looks similar to other work, I understand that very bad things can happen.**

**Signature \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Date\_\_12/08/2023\_\_**

**Directions: Please complete this Exam using THIS DOCUMENT.**

**Download and save this document.** Place all answers, work, illustrations, images, etc. that you want graded into this document.

If you are asked to code, you can create you code in Jupyter, as .py, on git, or whatever. You just need to have a link to your code that you will place on this document. Therefore, when you submit this exam, you will submit only this document.

**Please save this document as: YourName\_ML\_Exam\_2023.docx.**

**Notes and Rules:**

1. **No questions are permitted** by anyone for any reason during this exam. Follow the instructions, do not overcomplicate things, and make assumptions (that you clearly write down) if/as needed. Part of the test is your ability to “do the test”.
2. **It is not permitted to work with any other humans** on this Exam. While I am not concerned about this as I feel that all of you are very ethical, I am required to note that anyone who works together must get a “0” grade and may both fail the class and potentially have further issues with the program. Just do your own work 😊
3. **This Exam is open** – meaning you can use the web, class notes, my website, my code, your code, etc. **If you are in doubt about something – do not use it**. For example, if your buddy Bob posts code on the web that answers one of these questions and then you use it – that’s cheating. Using my code or your code is fine.
4. **This Exam will be due no later than 12/10 (Sunday) by 11:59 pm MT**. If you wait until the last hour to submit and run into a problem, there will be no solution. This Exam cannot be late for any reason. Please do not test this. Submit EARLY – like 24 hours early! I have set up the submission area so that you can **submit as many times as you want.** So, you can submit in advance (like Saturday), then you can do more work on Sunday if you wish, and then you can submit again if you want. **We will use the LAST (latest) submission** for grading. This way, there is no need to wait until the last minute to submit. Again, to be clear, the **system will close and will lock at 11:59pm MT on 12/10. Exams not submitted by that time will not be graded.** Also – for those of you who always push things – if you submit the “wrong version” or a “poem for your honey” instead of the Exam – that will **not matter!** **BE CAREFUL**. **Submit early and submit correctly.** (It makes me sad that I have to say all of this 😐)

**Part 1: The Interview**.

To answer the following questions, pretend you are in an interview. Interviewers are generally looking for brief answers (at least at first) and are trying to see what you do and do not know. Writing “more” does not show that you know more – and often – it suggests that you cannot articulate your answer succinctly.

Therefore, answer each of the following questions using 2 – 4 sentences (never more than 4). Be precise, concise, and succinct. Each topic here has books written about them. I am not looking for a book – just a few sentences that directly, quickly, and clearly answer the question. (All questions below were taken from interviews)

1. What is Ensemble Learning and give an example?

Ensemble Learning is the process of combining decisions from multiple models to obtain better performance. This can help improve the performance and robustness of our model and help us provide a generalized model overall. Example: In heart attack analysis and prediction, we can use various ensemble techniques like simple average, weighted average, max voting, stacking, bagging, boosting etc., to find the patients with higher and lower chances of heart attack.

1. Briefly explain the difference between supervised and unsupervised learning.

The main difference between supervised and unsupervised learning is the approach to train the model. In supervised learning, we train the model on the labeled data. The algorithm learns the patterns by mapping input to that corresponding output. Here the target variable is known. We use supervised learning to solve classification and regression problems. On the other hand, in unsupervised learning, we are working with unlabeled data and the algorithm tries to learn the patterns and structure from the data without labeled outputs. We use this technique for clustering or dimensionality reduction. While we use Supervised learning for making predictions, unsupervised learning focuses on understanding the underlying nature of the data.

1. What is a “Training set” and a “Testing Set” in supervised learning and how are they used?

In supervised learning, we train models on training data set to teach the models to yield desired output. Using this trained dataset, the model would be able to learn the patterns and relationships as the data would include inputs and correct outputs. The testing dataset is a dataset that the model has not seen during training and is used to evaluate the model’s performance on the new and unseen data. The testing set helps us to measure how accurate the model is while the training set helps the model learn from examples.

1. How would you handle or manage missing or incorrect data in a record dataset?

There are numerous ways to handle or manage missing or incorrect data in a record dataset. Few of them are:

1. Data Imputation (using Mean, Median and Mode imputation to replace missing values)
2. Deleting the rows with missing values completely.
3. Using algorithms that specifically handle such cases.
4. Interpolation
5. Handling incorrect data by removing outliers.
6. By implementing data validation checks to identify and correct incorrect data that falls outside of the expected ranges or formats.
7. In a confusion matrix, what is a false negative, and give an example where a false negative could cause a very serious issue.

A false negative in a confusion matrix is when the model predicts a negative result (such a condition not occurring) when, in fact, the condition is positive (it did occur). A false negative in heart attack analysis and prediction could have major consequences, because it suggests the model was unable to identify a person who was at risk, which would have prevented prompt intervention or preventive measures.

1. Suppose you have a labeled dataset where the labels are one of three categories (like dog, cat, mouse). However, suppose that 90% of the data is one label (like cat) and 10% of the data is the other two labels (like dog and mouse). Describe the issue here and a possible resolution.

One class (cat) makes up 90% of the labeled dataset in this situation, whereas the other two classes (dog and mouse) each make up only 5%. This represents a class imbalance. Machine learning models may favor the majority class as a result of this imbalance, which would lead to inaccurate forecasts for the minority classes. Using techniques such as under sampling, oversampling, weighted loss functions etc., we can ensure a balanced representation of the classes which can in turn enhance the model’s capacity to predict all the classes accurately.

1. Describe how supervised learning can be used to detect email spam. You describe this – this question is not asking you to paste from the web 😊

By using a labeled dataset that indicates which emails are spam and which are not, we may use supervised learning to detect spam in emails. The procedure is choosing a classification technique (e.g., SVM or Naive Bayes) and using the labeled data to train the model. All the required information is taken from the email text, including word counts, specific keywords, sender information, subject lines, email identifiers, external tags, attachments, and if the sender is known or unknown. The training data teaches the model unique patterns that allow it to distinguish between emails that are spam and those that are not. Based on the patterns it has learnt, the model can successfully detect whether or not fresh emails are spam once it has been trained.

1. Name two unsupervised learning methods and briefly describe what they do.

In unsupervised learning, we are working with unlabeled data and the algorithm tries to learn the patterns and structure from the data without labeled outputs. We use this technique for clustering or dimensionality reduction. Two unsupervised learning methods are:

1. **K-Means Clustering:**

It is a clustering algorithm that divides the data into ‘k’ number of clusters, each of which is identified by a centroid. In order to minimize the total squared distances inside each cluster, it allocates data points to the nearest centroid and modifies the centroids iteratively. This method is useful for identifying clusters or groups within our dataset.

1. **Association Rule Mining**

The goal of association rule mining is to identify or discover correlations between a group of items in large datasets. One of the popular ARM methods is the Apriori algorithm, which finds frequent items and determines association rules based on their occurrence. These rules show how likely it is for particular items to be related to one another.

1. Name three supervised learning methods and briefly describe what they do.

In supervised learning, we train the model on the labeled data. The algorithm learns the patterns by mapping input to that corresponding output. Here the target variable is known. We use supervised learning to solve classification and regression problems. Three supervised learning methods are:

1. **Naïve Bayes**

The probabilistic classification algorithm Naïve Bayes is founded on the Bayes theorem. Given a collection of input features, the method determines the likelihood of each class and forecasts the class with the highest probability.

1. **Decision Trees**

Decision Trees are a tree-like model where each internal node represents a decision based on a feature, and each leaf node represents the predicted outcome. The data is recursively partitioned based on the most informative features, with information gain or Gini impurity optimization being used.

1. **Support Vector Machines**

It offers various kernel options like linear, gamma, sigmoid, and Gaussian. SVM can transform features into higher-dimensional spaces, making input vectors separable by a hyperplane in this expanded space. If you have more than two groups to separate, multiple SVMs are needed. The crucial training points, known as support vectors, define the hyperplane. Quadratic optimization algorithms identify which training points are support vectors with non-zero Lagrangian multipliers for precise classification.

1. Suppose you work for a store, and you want to recommend items to customers based on past purchases. Which machine learning method would you use? Describe an example.

Association Rule Mining is what I would use if I worked for a retail store and my goal was to give clients individualized product recommendations based on their past purchases (ARM). ARM method is used in a real-world situation, which is comparable to the methodology used in Amazon's product recommendations. In this situation, I would be responsible for evaluating past user purchase data for Amazon.com. Using ARM, I would be able to spot trends like the things that customers typically purchase in tandem or the order in which they are made. With the usage of this data, consumers' overall buying experiences are improved through the creation of recommendations and ads that are specifically tailored to them.

**Part 2: SVMs**

**Question 1:** Here, you will illustrate the steps needed to solve for the Lagrange multipliers and then for the equation (w and b) of the SVM separating line. You may refer to the SVM slides for an example. Show all of your work. Do this right here in Word on this document. There is no need to write things out by hand 😊 If you must, you can use the “draw” tool in Word - but only as needed for images.

1. Choose any **three** datapoints that are 2D. (Example, the point (2,3) is a 2D datapoint) [do not use the same values that I use on the slides in class 😉]
2. Create a plot (you can use “draw” for this) that shows the cartesian coordinate system, the three points, and their labels as +1 or -1. You can choose the labels as long as you represent both labels.
3. Your next goal is to solve for the SVM model – which you will need **w** and b for.
4. Show all of your work and steps to do this. When you are done, draw your calculated separator line (your SVM model) onto the cartesian coordinate system so illustrate that it does in fact separate your points correctly.
5. Again – there is a complete example of this on the slides. You are welcome to use it to guide you.

**Solution:**

We have to use Lagrange multipliers to determine the parameters (w and b) of the SVM model while working with our selected data points and their labels. To keep things straightforward, we'll refer to the data points and their labels as follows:

Point x1 is (-1, 1) with label 1 y1 as -1

Point x2 is (1, 2) with label 2 y2 as +1

Point x3 is (2, 3) with label 3 y3 as -1

A graph with a number of points

Description automatically generated

3, 4, 5.

Let us find Lagrange multipliers λi and use it to calculate w and b.

We know that,

W = i \* \* ----- eq (a)

L(λ) = i – ½ ----- eq (b)

yi(w^T xi + b) – 1 = 0 ----- eq (c)

Let us calculate the dot product for our points x1, x2 and x3:

*x*1​⋅*x*1​=(−1)2+12=2  
*x*1​⋅*x*2​=(−1⋅1)+(1⋅2)=1  
*x*1​⋅*x*3​=(−1⋅2)+(1⋅3)=1  
*x*2​⋅*x*2​=12+22=5  
*x*2​⋅*x*3​=(1⋅2)+(2⋅3)=8  
*x*3​⋅*x*3​=22+32=13

We know that

λ1y1 + λ2y2 + λ3y3 = 0

Substituting values of y1, y2 and y3, we get

λ1(-1) + λ2(1) + λ3(-1) = 0

λ2 = λ1 + λ3 ------ eq (d)

From eq (b) we know

L(λ) = i – ½

Substituting x1, x2, x3 and y1, y2, y3 in eq (b) we get

L(λ) = λ1 + λ2 + λ3 – ½(2 λ1^2 + 5 λ2^2 + 13 λ3^2 - 2 λ1 λ2 + 2 λ1 λ3 - 16 λ2 λ3)

L(λ) = λ1 + λ2 + λ3 – λ1^2 – (5/2) λ2^2 - (13/2) λ3^2 + λ1 λ2 - λ1 λ3 + 8 λ2 λ3

Substituting eq (d) in above equation, we get

L(λ) = 2 λ1 + 2 λ3 – (5/2) λ1^2 + 3 λ1 λ3 – λ3^2

Let us differentiate L(λ) w.r.t λ1 and λ3

dL(λ)/d λ1 = 2 - 5 λ1 + 3 λ3 = 0

dL(λ)/d λ3 = 2 - 2 λ3 + 3 λ1 = 0

Solving these equations, we will get λ1 = 10 and λ3 = 16

From eq (d) we know,

λ2 = λ1 + λ3 = 26

Hence, we have λ1 = 10, λ2 = 26 and λ3 = 16

Now, we can calculate weighted vector.

From eq(a) we know that,

W = i \* \*

Substituting x1, x2, x3 and y1, y2, y3 and λ1, λ2, λ3 we get

W = 10 \* (-1) \* (-1, 1) + 26 \* (1) \* (1, 2) + 16 \* (-1) \* (2, 3)

W = (4, -6)

Now, let us calculate b. We know that

yi(w^T xi + b) – 1 = 0

let us substitute point x2, y2, we will get

y2(w^T x2 + b) – 1 = 0

(1)((4, -6)^T (1, 2) + b) – 1 = 0

((4, -6)(1, 2) + b) – 1 = 0

(4\*1 + (-6)\*2 + b) – 1 = 0

b – 8 = 1

b = 9

Let us cross verify,

let us substitute x3, y3, we will get

y3(w^T x3 + b) – 1 = 0

(-1) ((4, -6)(2, 3) + b) – 1 = 0

(-1)(8 – 18 + b) = 1

(-1)(b – 10) = 1

b – 10 = -1

b = 9

Hence the calculated value for b is 9.

Therefore, the SVM models parameters are: Weight vector w is (4, -6) and offset b is 9.

So, the equation of optimal hyperplane separating the classes is

(4)\*x1 + (-6)\*x2 + 9 = 0

4x1 – 6x2 + 9 = 0

A line graph with a dotted line and a dotted line

Description automatically generated

**SVMs Question 2:**

1. A general polynomial kernel can be written as K=(**a**T**b** + r)^d where **a** and **b** are any two points (vectors) in your dataset. Suppose you have a polynomial kernel K specifically with r=0 and d=3. What is **your K**?

Your K (with r=0 and d=3) = ?

1. Write your K as a dot product between two vectors. Show all the work.
2. Choose any 2D point. [For example, (2,3) is a 2D point]. Use your K and show what your 2D point would be in that new kernel space. Show all the steps and work.

Note: Here again, you can find many examples of this type of thing on the SVM slides. Feel free to use them as a guide.

**Solution:**

1. We know that a general polynomial function can be written as K=(**a**T**b** + r)^d ---- eq(a)

Given that, r=0 and d=3

Now if we substitute in eq(a) we get

K=(aTb + r)^d = (aTb + 0)^3 = (aTb)^3

In 2D, let a is (a1, a2) and b is (b1, b2)

Now K = (aTb)^3 = (a1b1 + a2b2)^3

Implies,

K = (a1b1)^3 + 3(a1b1)^2 (a2b2) + 3(a1b1)(a2b2)^2 + (a2b2)^3

If we are choosing 1 point, then let a=b, we will get

K = (a1a1)^3 + 3(a1a1)^2 (a2a2) + 3(a1a1)(a2a2)^2 + (a2a2)^3

K = a1^6 + 3a1^4 a2^2 + 3a1^2 a2^4 + a2^6

1. We can represent K as a dot product of two vectors by taking square root of each term in K.

Hence, for K = (aTb)^3 the dot product would be

K = (

=

=

1. Let us take a 2D point P = (3, 4)

Hence a = b = (3, 4)

a1 = 3, a2 = 4

We know that K = (aTb)^3 = ( (3, 4) . (3, 4) )^3 = (3.3+4.4)^3 = (9 + 16)^3 = 25^3 = 15625

We can also substitute in K = a1^6 + 3a1^4 a2^2 + 3a1^2 a2^4 + a2^6 = 3^6 + 3\*(3)^4\*(4)^2 + 3\*(3)^2\*(4)^4 + 4^6 = 729 + 3(81)(16) + 3(9)(256) + 4096 = 15625.

**Part 3: Doing Data Science**

This part of the test is testing whether you can “be a data scientist”:

– understand a problem

– determine how best to solve it on your own

– and present results and conclusions to a non-technical audience.

There is no one to tell you what to do or how to do this job.

You will make the decisions, determine what to do and how you want to do it, and use your judgement, etc.

**The Problem**

Suppose you work for Amazon and they ask you to (1) determine if a User should get a credit card (yes or no), and (2) determine which products to advertise to the User.

Imagine that you have all of Amazon’s User data – anything you need.

Design/create a small dataset that you can use to address the questions above. Paste it here. Keep it small with 3 – 4 columns and 25 – 30 rows. You decide the column/variable names and what the data would look like. It can be anything you want that also makes sense with respect to the question. You create the dataset you need.

Next, describe, illustrate, and perform (in Python or R) whatever you think you need to do so that you can address the questions asked properly.

The reason why you are not being told what tools/methods/models to use is because on the job – you are the data scientist and you determine what to use, what to do, what the results mean, and how to present it. Please do that here.

Discuss and illustrate your results and conclusions. YOU decide what is needed to do this.

**\*\*Do not include or paste in any code on this document**.

To submit code used for this question, place your code online (however you want) and include a link to the code here.

Your audience/readers include only **non-technical** people.

Git Link:

Please download the csv files for both use cases.

For use case 1 use data file exam\_data.csv and run AmithNandivada\_ML\_U1.ipynb.

For use case 2 use data file transactional\_exam\_data.csv and run AmithNandivada\_ML\_U2.R