**Machine Learning Exam**

**Honor Statement:**

**I, Wesley Allen, 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** Wesley Allen **Date: 12/04/2024**

**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 your code in Jupyter, as .py, in 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. Do not place code into this document.

**Please save this document as: YourName\_ML\_Exam.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. **It is not possible to submit this Exam late**. Please submit early to avoid any issues. If you have an ADA Accommodation, please email me ASAP.

**The Exam**

**Part 1: The Interview**.

To answer the following questions, pretend you are on 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. Using AI will create answers that sound exactly like everyone else’s and therefore can **and will**  lose points. Do not Google. Just answer.

Answer each of the following questions **using only 2 – 4 sentences (never more than 4).** Be precise, concise, and succinct. Each topic here has books written about it. 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?
   1. Ensemble Learning refers to the process of using a committee of different models to make predictions instead of one individual model. One popular Ensemble Learning method is the Random Forest Model. This uses a committee of decision trees, each trained on different and potentially overlapping subsets of data. To predict, each individual tree makes a prediction, and we then assign the data point to the class that had the most “votes” from the individual trees
2. Explain how boosting is different from bagging.
   1. Boosting aims to reduce bias in our prediction by sequentially training models, where each aims to improve upon the errors of its predecessor. On the contrary, bagging aims to reduce the variance in our prediction, by taking the average prediction from a set of models that are each trained independently.
3. Briefly explain the difference between supervised and unsupervised learning?
   1. Supervised Learning is concerned with prediction and uses statistical models to learn data patterns based on provided labels. These patterns are then evaluated on a testing set without labels. Unsupervised Learning is more about discovering patterns or some sort of organizational structure of the dataset. We use statistical models on the entire dataset, without labels, to gain a deeper understanding of these underlying patterns.
4. What is a “Training set” and a “Testing Set” in supervised learning and how are they used?
   1. We initially begin with a labeled dataset. A subset of the data (70%-90%) is used to help the model make prediction, and the labels allow the model to correct erroneous predictions. The labels are removed from the remaining data and this is used as the testing/evaluation set.
5. How would you handle or manage missing or incorrect data in a record dataset?
   1. First of all, the answer to this question is entirely dependent on the context. For missing or incorrect data, it might make sense to interpolate using the mean, median, or a quantile value, but each will have differing effects on the distribution of the data. In addition, if a significant amount of data is missing, it might also make sense to delete the row altogether.
6. In a confusion matrix, what is a false negative, and give an example where a false negative could cause a very serious issue.
   1. A false negative or type 2 error occurs when a record was incorrectly labeled as negative, when it is actually positive. In a confusion matrix, the number or proportion of false negatives is given in either the top right or bottom left corners. Consider a scenario, where a machine learning model is predicting whether or not someone has cancer. A false negative would imply that the model predicted an individual was cancer free, but instead, they actually do have cancer.
7. 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 are the other two labels (like dog and mouse). Describe the issue here and a possible resolution.
   1. Without adjusting the data, the model will incorrectly favor the label cat because this label contributes more to the loss function than the others. Secondly, typical metrics like accuracy may be misleading because if a model simply classifies each record as cat, then the model is guaranteed to have 90% accuracy despite the obvious problems with the model. To circumvent this, we could apply weights to our loss function, increasing the significance of less numerous labels like dog and mouse. Alternatively, we could oversample the minority classes by simply duplicating the corresponding records or generating synthetic data.
8. 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 😊
   1. Suppose we model an email as a Bag of Words, meaning we represent an email as a high dimensional vector where the number of entries is the number of unique words in the vocabulary, and each element corresponds to the frequency of the word in that email. If we have a dataset of emails labeled as spam or not spam, we could then create testing and test sets after applying the Bag of Words to each email. From here, the process doesn’t differ much standard supervised learning processes. We train the model using a labeled training set, and evaluate using an unlabeled testing set.
9. Name two unsupervised learning methods and briefly describe what they do.
   1. K-means clustering is a model that seeks to identify unknown groupings or clusters within the data, helping us to understand the underlying structure of the data. It does this by initializing k centroids, calculating the distance between each data point and each cluster using some distance measure, assigning the datapoint to the closest cluster, update the centroids, and repeat.
   2. The Apriori Algorithm takes as input a transactional dataset, and seeks to find association rules between items. It iteratively generates sets of items that meet a minimum support threshold and leverages the fact that if a itemset is frequent, then all subsets must also be frequent, along with the corollary that if an itemset is infrequent, all supersets of this itemset are also infrequent.
10. Name three supervised learning methods and briefly describe what they do.
    1. Each model adjusts parameters in order to more accurately fit the data while undergoing training. Linear regression involves fitting a linear equation to a dataset and adjusts the weights of the linear equation using gradient descent. Support vector machines aim to compute the optimal hyperplane that separates the data points and classifies datapoints based on where it is located in relation to the hyperplane. Decision trees aim to identify partitions in the dataset that maximize the information gain.
11. 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.
    1. I would use the Apriori algorithm. This would enable us to understand which items are commonly purchased together. Suppose we work at Amazon, and we know that Raspberry Pi computers are often purchased with a book titled Raspberri Pi for Dummies. If a user is purchasing a Raspberry Pi computer, it would make sense to recommend the book to the user as well.

**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. You may type, write, or whatever. Just be sure all work and answers are here.

1. Choose any **three** datapoints that are 2D. [do not use the same values that I use on the slides in class 😉]
2. Create a plot (you can use “draw” for this or whatever) 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.**
   1. **A graph with numbers and lines

      Description automatically generated**
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. Be clear and show each step. Make it easy for us to read and to see what you are doing. When you are done, draw your calculated separator line (your SVM model) onto the cartesian coordinate system to illustrate that it does in fact separate your labeled points correctly.
5. Again – there is a complete example of this on the slides. You are welcome to use it to guide you.

A math equations and formulas

Description automatically generated with medium confidence

A graph of a graph

Description automatically generated

**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) = (a^Tb)^3

1. Next, write your K as a dot product between two vectors **a** and **b**. Show all the work.

A math equations with black text

Description automatically generated with medium confidence

1. Choose any 2D point. Use your K and show what your 2D point would be in that new kernel space. Show all the steps and work.

A math problem with black text

Description automatically generated with medium confidence

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.

**Part 3: Doing Data Science**

This part of the test is testing whether your skill at “being a data scientist” and your ability to:

1. Understand a new problem
2. Determine how best to solve it on your own
3. Present results with excellence
4. Create conclusions for 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 have asked 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 any/all of Amazon’s User data – anything you need.

1. 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.
   1. For the first problem I used the following dataset

A screenshot of a black screen

Description automatically generated

* 1. For the 2nd problem I used the following dataset.

A screenshot of a black screen

Description automatically generated

1. 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**.

For problem 1, I utilized a decision tree to base my prediction due to it’s high interpretability. The resulting tree is shown below.

A diagram of a number of numbers

Description automatically generated with medium confidence

The root of the tree partitions the applicants based on whether or not the monthly income is greater than or equal to 7253. If it is not, we then check if their credit score is greater than or equal to 564.5. If neither are true, the applicant is denied. If the income check is false, but the credit score check is true, we then evaluate if the monthly income is less than or equal to 2808.5. If true, the applicant is approved, and if false, the applicant is denied. Suppose the initial income check is true. We then consider if their monthly expenditure on amazon is less than or equal to 200.5. If false, they are denied, and if true, they are approved. I split the dataset into training and testing sets when constructing the tree. We were only able to have 60% accuracy on the testing set. While the accuracy is low, this is likely due to the fact that the question asks us to make our own dataset instead of using one of the many available datasets online. Further, since my initial labels were created based on if I felt they should get approved or not, there was no reason to expect any sort of accuracy from this model. If instead we were provided a dataset with sufficient rows and real-life numbers, we would have likely seen greater accuracy.

For the 2nd problem, I utilized the Apriori algorithm to generate frequent itemsets to determine recommendations. The following itemsets were generated.

A computer screen with white text

Description automatically generated

Suppose an amazon user purchases a laptop. Since, the laptop belongs to a frequent itemset that includes headphones, users that purchase a laptop are more likely to also purchase headphones.

Discuss and illustrate your technical results and non-technical 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 both technical and **non-technical** people.