Shape

Description automatically generated with medium confidence School of Engineering  
Department of Electrical Engineering  
Instructor: Matthew Beckler

SENG4400: Data Science & Machine Learning

Final Exam

Name:

**Open book, open notes, open laptop, open internet. Calculators ok****. Write your name at the top.  
All words must be your own.**

1. [10 points] In your own words, describe how data science differs from statistics

To me they can be very similar math wise but where it differs comes down to the data. Data science focuses on the collection, organization, analyzation, and visualization of large amounts of data. During this process is where I think they are similar because sometimes in data science you use the same (or similar) equations as you learned in stats, giving you the same process through the math. Now stats, on the other hand, use mathematical models to quantify relationships between variables and outcomes, then make predictions based on said relationships. That definition itself can make them sound very similar, so in conclusion, the fields differ in the modeling process, size of their data, the types of problems, and the “language” used.

2. [10 points] In your own words, compare and contrast supervised learning and unsupervised learning.

To compare and contrast we must first understand the definition of both sides. Supervised learning is a machine learning technique. This technique is used to “train” or “supervise” data into predicting data accurately. Using labeled inputs and outputs the supervised data model can measure the accuracy and learn over time (pretty cool). Unsupervised learning is used for analyzing and clustering data sets. It is useful to discover patterns in the data that us as humans couldn’t notice before. Off the bat from the definitions, there are some differences here. Supervised learning: predictions and accuracy, Unsupervised learning: clustering and patterns. They also have different problems or algorithms used when data mining. Supervised uses classification and regression to predict. Unsupervised uses clustering, association, and dimension reduction to catch onto patterns. Supervised data uses labeled datasets, unsupervised uses unlabeled datasets. Where they are similar is they both help us as data scientist understand data better and allow us to visualize it.

3. [10 points] In your own words, describe why we often split data into a training set and a testing set. For which types of problem do we use this approach. How do we use each set?  
  
We often split data these ways to increase our accuracies and understand the data model. I like to compare it to sports, imagine you started playing soccer and before any training you get thrown into a game, you aren’t going to be good right? You need training before being tested (in a game), that is the same with accuracy with the data, it isn’t going to be high without training it first. We use train/test split when we want to predict the data, like in linear regression we can see where the trend in the data is going. We use each set like I mentioned in the beginning to train it first to predict the outcome and then test it on the testing set in the data.  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
4. [10 points] In your own words, describe potential ethical implications of machine learning.

Since machine learning must deal with data, I always think a big ethical concern is privacy of the data. Companies like Facebook, for example, got caught in a scandal basically selling their user’s data without their consent. That is more of a just data concern, but with machine learning a concern is some algorithms are “black boxes” where it is impossible to see how they really work, so can we really trust the outcome? Lastly, an ethical concern (mix of A.I and machine learning) is the robots taking over/getting smarter than humans. Of course, it sounds outlandish but is a very big possibility that we take A.I and machine learning too far and cause bad things to happen to us as humans.

5. [10 points] In your own words, compare and contrast K-Means clustering with DBSCAN clustering.

K-means clustering is a partitioning algorithm and recreates datasets to (k) clusters formed. DBSCAN is a density-based clustering algorithm. K-means doesn’t care about outliers or “noise” and well cluster all data objects in the set. DBSCAN will discard any objects that it defines as “noise”. K-means is much more efficient in large datasets, DBSCAN cannot efficiently handle high dimensions of datasets. There are plenty more differences but the last one I will mention will be the parameters. K-means takes one parameter (k) for clustering and DBSCAN takes two parameters Radius and Minimum Points. Where they are similar is they both cluster data and they both help us see patterns (through visualization) in the datasets.

6. [10 points] So many k values! In your own words, compare and contrast K-Means and K-NearestNeighbor. What is each used for? What does the k value represent for each?

Highest level difference is k means is an unsupervised learning algorithm and KNN is a supervised, so as said before unsupervised = clustering and patterns, supervised = predictions and accuracies. KNN is based on feature similarity between the objects and K-means is based on the division of objects into clusters. KNN is a classification technique while k-means is a clustering technique. With KNN you can differentiate data through its characteristics, like a people dataset with hair colors, height, weight, etc. With k-means, for example, would then cluster similar people into k amount of clusters.

7. [3 points] What is the entropy of a set containing 0 red, 12 blue, and 0 green items?

**\*\*For these next 3 questions I used items as x (1 row) and color as y (3 columns)\*\***

Entropy of Y: 0.0000

Entropy of Y Given X: 0.0000

Information Gain: 0.0000

8. [3 points] What is the entropy of a set containing 3 red, 9 blue, and 0 green items?

Entropy of Y: 0.8113

Entropy of Y Given X: 0.0000

Information Gain: 0.8113

9. [3 points] What is the entropy of a set containing 2 red, 4 blue, and 6 green items?

Entropy of Y: 1.4591

Entropy of Y Given X: 1.4591

Information Gain: 0.0000

10. [11 points] For the dataset below, A-D are input features, and E is the binary output class. Without doing any calculations, which input feature would be best for our first decision tree decision node? Why? What would you use for a threshold value in that decision node?

Index A B C D E

1 4.8 3.4 1.9 0.2 positive

2 5.0 3.0 6.6 1.2 positive

3 5.0 3.4 1.6 0.2 positive

4 5.2 3.5 1.5 0.2 positive

5 5.2 3.4 4.4 0.2 positive

6 4.7 3.2 1.6 0.2 positive

7 4.8 3.1 1.6 0.2 positive

8 5.4 3.4 1.5 0.4 positive

9 7.0 3.2 4.7 1.4 negative

10 6.4 3.2 4.7 1.5 negative

11 6.9 3.1 1.4 1.5 negative

12 5.5 2.3 4.0 1.3 negative

13 6.5 2.8 4.6 1.5 negative

14 5.7 2.8 4.5 1.3 negative

15 6.3 3.3 4.7 1.6 negative

16 4.9 2.4 3.3 1.0 negative

I would say input C because it has the highest overall difference when comparing positive to negative ~3 point difference. I would use 4 as the threshold.

11. [10 points] Given that Decision Trees often over-fit the data, in your own words, describe two approaches to prune a decision tree to avoid over-fitting.

Cross validation which is a resampling method that uses different portions of the data to test and train a model on different iterations

Also by tuning the hyperparameters of the decision tree model we can prune the trees and prevent them from overfitting.

12. [10 points] In your own words, describe two approaches to combine multiple decision trees to build a better overall model.

One way you can accomplish this is by summing the spectra of each model and then transform the results back into to the decision tree domain.

The other way is using the random Forests model which combines different decision trees, these models are done on random subsets of the features in the trees and in some cases random subsets of the data.