CS 412 Take-home Final

Due: May 6th - 11:59 PM CDT

Name: _____

UIC NetID: _____

UIN: _____

Please make sure all of your questions are properly labeled before you submit. No late submissions will be accepted

Question	Max Points	Earned Points
1	15	
2	40	
3	20	
4	15	
5	30	
Total	**120**	

- You are allowed to use your notes with this exam
- All work that is given must be your own
- You will submit your assignment to gradescope when you are finished.
- For the non-coding portions, you may answer the questions using any format you'd like, e.g, typed, handwritten, tablet etc.
- Good luck! I will be on slack to answer clarification questions. Please use the #exam channel

1 True/False/Contingent

Determine whether the following propositions are always true, always false or contingent. Explain your answer with a sentence or two.

a) Classification problems are easier than regression problems

b) An SVM with a higher value of C will have more support vectors if run on the same dataset

c) Regularization reduces error within the training set.

d) The kernel transformation allows SVM to make non-linear discriminators

e) For kNN, high values of k results in longer training times.

2 Short Answer

Answer the following short prompts

a) Other than examples from class, give an example of a ML model which should raise ethical questions around autonomy.

b) What is leave-one-out-cross-validation? Give an example of when it might be used.

c) Give three indicators of overfitting.

d) What is the vanishing gradient? Give a reason why it might occur.

e) How do the random start points impact the kMeans algorithm?

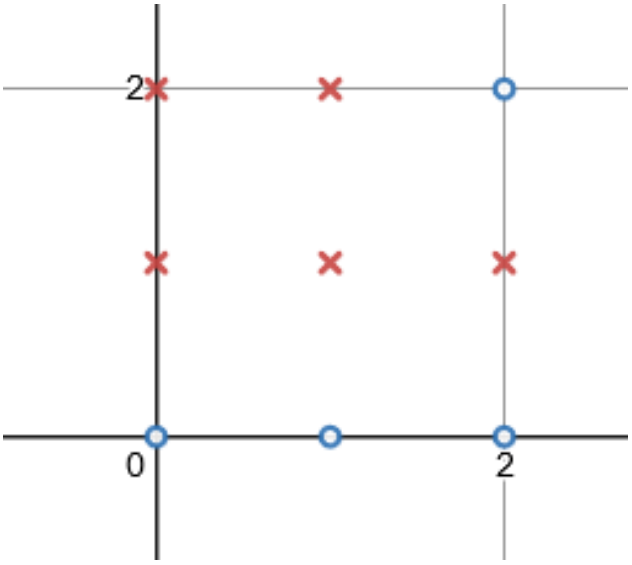
f) Give two models which use gradient descent. Do these models converge to local or global optima?

g) Explain the importance of separating data into training and testing.

h) How do ensemble methods reduce error?

3 Decision Trees

Consider the following dataset.



a) Draw the binary decision tree for the above data with one internal node that has lowest error

b) Draw the binary decision tree for the data on the previous page with two internal nodes that has lowest error

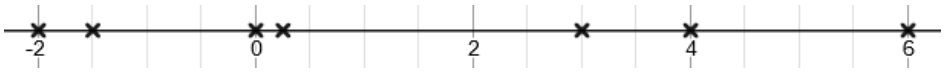
c) Draw the binary decision tree which has zero training error.

d) Give a reason why you might not want to select the model from part c)

e) If almost all of the features in a dataset are noise, i.e. they do not explain the data, why might a random forest model perform poorly? What can be done to correct it?

4 Unsupervised Learning

For this problem use the following linear dataset. It may be helpful for you to label the points



a) Draw the agglomerative clustering dendogram for the following data set.
Use complete linkage. Indicate the order in which the clusters were formed

b) What is semi-supervised learning and when is it most useful?

c) Since there is no accuracy, give two methods by which we can judge the quality of an unsupervised learning approach.

5 Client Specification

A client wants you to build a model on their dataset using a **neural network**.You can import the data with the following.

```
import numpy as np
features = np.loadtxt("features.csv")
labels = np.loadtxt("labels.csv")
```

Find what you think is the best model and answer the following questions. There will be a separate submission for your code

a) What were the hyperparamers for your final neural network? How did you select those hyperparameters? Did anything unusual happen while you were searching?

b) Give the 95% confindence interval for your final error on the test set. Use the Hoeffding bound.

c) Give three ways in which randomness impacted the model or your process

d) Give evidence that your model is not overfit. Figures maybe helpful here