CS4364/6364 Intro to Machine Learning
Dr. Dobolyi

Exam 1

Please **DO NOT START** the exam until instructed, out of fairness to all students. 60 minutes.

Score: _____ / 76pts

Name: ___________________________

GWID: ___________________________

1. Short answer: write a **<u>phrase or less</u>** (no full sentences please) for each item below. (2pts each == 48pts)
    a. What is the difference between supervised learning and unsupervised learning?

    b. What is the inductive bias?

    c. Give an example of any of the inductive biases we've gone over in class.

    d. Give an example of a binary classification problem.

    e. Give an example of a multi-class classification problem.

    f. Give an example of a regression problem.

    g. Imagine I have a model that achieves near-perfect performance on some arbitrary dataset, but I know this is too high. What is something that I could have done wrong in training this model to cause this result?

    h. What is the difference between training error and training loss?

    i. List and describe three parameters of a RandomForest you used to tune a model in your homeworks:
        i.
        ii.
        iii.

    j. I re-ran the identical model training on the identical dataset, and got slightly different accuracy scores. What might have happened?

    k. What am I trying to prevent when pruning a decision tree model?

    l. What is bagging in ensemble learning?

    m. List two ways to reduce the noise/complexity in your features (without adding or deleting samples)
        i.
        ii.

    n. Give an example of feature engineering

o. List two ways to handle missing columns/values in your features (i.e. what you would do with `NaN` values):

   i.

   ii.

p. List one reason why you would want to scale/normalize your input features.

q. List one way you can tell you have overfit your model to your dataset.

r. List one way to help solve/reduce overfitting (other than adding more data).

s. What is model bias?

t. What is model variance?

u. What is precision?

v. What is recall?

w. Why is a confusion matrix helpful for multi-class classification performance evaluation, specifically for multiple classes?

x. Why would I want to one-hot encode a categorical variable?

2. **<u>Two-sentence</u>** answers (4 pts each == 28 total):

   a. What does it mean for a model to generalize?

   b. Why do we often use a validation set in addition to a holdout set?

   c. What is cross-validation, and why do we use it?

d. In what scenario is it better to choose a linear regression over a RandomForest? Why?

e. In what scenario is it better to choose a RandomForest over a linear regression? Why? (Note: don’t just ‘take the inverse’ of the previous question).

f. How does a decision tree choose the best split at each node?

g. Why is it bad to have too many features?

--END OF EXAM--

Extra credit: List one interesting application of machine learning to a real-world problem