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?

    label
    b. What is the inductive bias?

    the assumption made by the algorithm to generalize it's training data
    c. Give an example of any of the inductive biases we've gone over in class.

    linear regression, shallow decision trees in random forest
    d. Give an example of a binary classification problem.

    is dog or not
    e. Give an example of a multi-class classification problem.

    what kind of animal
    f. Give an example of a regression problem.

    temperature tomorrow
    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?

    over fit to the training set
    h. What is the difference between training error and training loss?

    errors are percentage of incorrect prediction, loss is measured by specific self defined loss functions
    i. List and describe three parameters of a RandomForest you used to tune a model in your homeworks:
        i. n_estimators
        ii. max_depth
        iii. max_leaf
        max_feature
    j. I re-ran the identical model training on the identical dataset, and got slightly different accuracy scores. What might have happened?

    there are certain randomness in the model (e.g. random forest)
    k. What am I trying to prevent when pruning a decision tree model?

    over fitting
    l. What is bagging in ensemble learning?

    use multiple models together to predict the result
    m. List two ways to reduce the noise/complexity in your features (without adding or deleting samples)
        i. remove highly correspondence features
        ii. change minor values in high cardinal features to be 'others'
    n. Give an example of feature engineering

    fixing missing value
    scale input feature

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

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

so model will not prioritize on some features

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

compare the accuracy score on holdout vs validation

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

early stop, use weaker model

s. What is model bias?

the overall error of the model

t. What is model variance?

how stable the model is

how many perdictions are true

u. What is precision?

how many the model predicted divide samples are supposed to be predicted

v. What is recall?

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

allows you to figure out what target classes are confusing the model

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

some times range value doesn't make sense, e.g.: ice cream flavor

2. **Two-sentence** answers (4 pts each == 28 total):
   a. What does it mean for a model to generalize?

the model is trained to be able to predict on other datasets of the same problem.

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

we need to figure out if the model has really generalized.(rather than overfitting to validation set)

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

cross-validation is to use different combinations of dataset to train and validate the model, this is used to check if the model has generalized on the problem or it's just overfitting on the training set

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

there is a linear relationship between the features and the target, like height vs weight. The induction bias of regression has better performance on this type of problems.

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).

when a problem is not linear, like predict favourite color. Liner regression is not powerful enough for this kind of problem.

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

the split that cause the lowest entropy(the most pure split)

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

decision trees are bad at too many features, can cause the model to overfit

----------END OF EXAM----------

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