Machine Learning, Fall 2019: Project 3
out on 10/15/19 - due on 10/29/19 by noon on Blackboard

Your Name Here

Header: List the major resources you used to complete this project and the programming language you used. The LaTex that generated this page is available here: https://www.overleaf.com/read/pknftzdpkzzc. Please submit a pdf file and your code for project 3.

You may use any programming language you like (Matlab, C++, C, Java, Python ...). **You may use any toolkit that performs machine learning functions but do NOT collaborate with your classmates. Cite any resources that were used.**

1 Project Description

In this exercise, you will implement support vector machines (SVM) to solve the Adult Census Income kaggle task: https://www.kaggle.com/uciml/adult-census-income to predict whether income exceeds $50K/yr based on 1994 census data. Instead of using the complete dataset, use the dataset linked on the syllabus: https://drive.google.com/file/d/17WwuiehBv79FQUsNz-W2IZfXk31f_FsM/view?usp=sharing. I am giving you 10% of the original dataset in order to save you time during training.

2 Dataset preprocessing and interpretation

1. **5 points** Some samples have missing features: There are several rows of data containing '?'. Replace the missing feature values for nominal and numeric attributes with the modes and means from the training data. Provide your code in the report (pdf file).

2. **5 points** Dealing with discrete (categorical) features: There are some categories that contain discrete features. For example, *marital.status* can have different values: "Widowed", "Divorced", "Never-married", and so on. Find a good representation for them so that they can be used to train a support vector machine and explain your methodology.

3. **5 points** Split the dataset for stratified 10-fold-cross validation. Provide your code in the report (pdf file).

4. **5 points** Analyze the features and make a scatter plot with the two features that have the highest information gain. Which features are these and what are their information gain values?

3 Using a linear soft-margin SVM

1. **15 points** Train your SVM with *stratified 10-fold-cross-validation* on the 2 features with the highest information gain and visualize your boundary. i.e. plot the support vectors (list which data points they are), the margin, and draw the decision boundary.

2. **15 points** Change the hyper-parameter C from small to larger values. Report your observations on how the value of C would affect SVM's performance. Draw the decision boundaries and margins with smaller and larger values of C to explain its effect in two separate figures.

3. **15 points** Train the SVM using all the features. Find a way to determine the optimal value of C. Report your methodology and accuracy from *stratified 10-fold-cross-validation* by using learning curves.

4 Using SVM with a kernel

1. **15 points** Compare the performance (precision, recall, f1-score, and variance) of different kernels: Linear, RBF, and polynomial.

2. **20 points** Try your best to get higher performance! You can design your own kernel, use bagging or boosting methods, logistic regression, decision trees, Naïve Bayes, or whichever method you prefer. Provide your code and your evaluation method, then explain why the performance is better with your method of choice by using learning curves.