Assignment 3: SVM, and Model Selection

UVA CS 6316/4501 :
Machine Learning (Fall 2016)

Out: Oct. 3 / Mon, 2016
Due: Oct. 16 / Sun midnight 11:55pm, 2016 @ Collab

a *The result part of the assignment should be submitted in the PDF format through Collab. If you prefer hand-writing the writing part of answers, please convert them (e.g., by scanning or using apps like OfficeLens) into PDF form.*

b *For questions and clarifications, please post on piazza. TA Weilin (xuweilin@virginia.edu) and Kamran (kk7nc@virginia.edu) will try to answer there.*

c *Policy on collaboration:*

Homework should be done individually: each student must hand in their own answers. It is acceptable, however, for students to collaborate in figuring out answers and helping each other solve the problems. We will be assuming that, with the honor code, you will be taking the responsibility to make sure you personally understand the solution to any work arising from such collaboration.

d *Policy on late homework: Homework is worth full credit at the midnight on the due date. Each student has three extension days to be used at his or her own discretion throughout the entire course. Your grades would be discounted by 15% per day when you use these 3 late days. You could use the 3 days in whatever combination you like. For example, all 3 days on 1 assignment (for a maximum grade of 55%) or 1 each day over 3 assignments (for a maximum grade of 85% on each). After you've used all 3 days, you cannot get credit for anything turned in late.*

Question 1. Support Vector Machines with Scikit-Learn

- (1) Install the latest stable version of scikit-learn following directions available at `http://scikit-learn.org/stable/install.html` Also make sure to download "salary.labeled.csv" from collab.
- (2) For this assignment, you will create a program using scikit-learn's C-Support Vector Classifier.[1]
- Given a proper set of attributes, the program will be able to determine whether an individual makes more than 50,000 USD/year. You may use code from HW2 to help you import the data. Bear in mind you will also need to do some preprocessing of the data before applying the SVM.
- Two sample files are provided. The unlabeled sample set "salary.2Predict.csv" is a text file in the same format as the labeled dataset "salary.labeled.csv", except that its last column includes a fake field for class labels.
- 2.1 You are required to provide the predicted labels for samples in "salary.2Predict.csv".

[1] Documented here: http://scikit-learn.org/stable/modules/classes.html#module-sklearn.svm

- 2.2 We will evaluate your output 'predictions' - an array of strings (">50K" or "<=50K") corresponding to the true labels of these test samples (ATT: you don't have these labels !!!). This simulates a Kaggle-competition in which test labels are always held out and only team-ranking be released after all teams have submitted their predictions. When grading this assignment, we will rank all students' predictions. So please try to submit the best performing model that you can!
- 2.3 You need to report the classification accuracy results from 3-fold cross validation (CV) on the labeled set using at least three different SVM kernels you pick. Please provide details about the kernels you have tried and their performance (e.g. classification accuracy) on train and test folds into the writing. For instance, you can summarize the results into a table with each row containing kernel choice, kernel parameter, CV train accuracy and CV test accuracy.
- (Hint: you can choose SVM kernels like, basic linear kernel / polynomial kernel, varying its parameters / RBF kernel, varying its parameters).

Submission Instructions: You are required to submit the following :

1. A python program that includes the functions:

   trainedModel, CVscores = svmIncomeClassifier.processLabelSet("salary.labeled.csv")

   It should be able to train the selected model using a set of hyperparameters on the train data, these hyperparameters can be hard coded or be input by the user.

   Next, we should be able to print out the CV classification score using the following function:

   print(svmIncomeClassifier.score())

   Next, we should be able to use a saved (trained) model to classify an unlabeled test set using the following function: predictions = svmIncomeClassifier.predict("salary.2Predict.csv", trainedModel)

2. A table in your PDF submission reporting classification accuracy (score) averaged over the test folds, along with details of the kernels, best performing hyperparameter C for each case etc

Classes: >50K, <=50K.
Attributes:
age: continuous. workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked. fnlwgt: continuous. education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool. education-num: continuous. marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse. occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces. relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried. race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black. sex: Female, Male. capital-gain: continuous. capital-loss: continuous. hours-per-week: continuous. native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

Table 1: About the data in Q1.