Assignment 4: SVM, and Model Selection

UVA CS 4501-03 :
Machine Learning (Spring 2018)

Out: Mar. 27 2018
Due: Apr. 11, Wed midnight 11:59pm, 2018 @ Collab

a *The assignment should be submitted in the PDF format through Collob. If you prefer hand-writing QA parts of answers, please convert them (e.g., by scanning or using PhoneApps like officeLens) into PDF form.*

b *For questions and clarifications, please post on piazza.*

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.*

1 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".
- 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!

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

- 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 :
(The starting code, 'income_classifier.py', has been provided in Collab.)

1. A python program that includes the statements:

   ```
   clf = SvmIncomeClassifier()
   trained_model, cv_score = clf.train_and_select_model(``salary.labeled.csv")
   ```

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

   Next, we should be able to use a trained model to classify an unlabeled test set using the following function:

   ```
   predictions = clf.predict(``salary.2Predict.csv",trained_model)
   ```

2. A file "predictions.txt" generated by:

   ```
   clf.output_results(predictions)
   ```

   Please do not archive the file or change the file name for the automated grading.

3. 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.

2 Sample Exam Questions:

Each assignment covers a few sample exam questions to help you prepare for the midterm and the final. (Please do not bother by the information of points in some the exam questions.)

Question 1. Support Vector Machine

Soft-margin linear SVM can be formulated as the following constrained quadratic optimization problem:

$$argmin_{\{w,b\}} \frac{1}{2} w^T w + C \sum_{i=1}^{m} \epsilon_i$$

such that

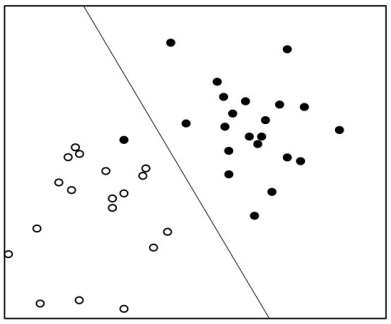
$$y_i(w^T x_i + b) \geq 1 - \epsilon_i$$
$$\epsilon_i \geq 0 \qquad \forall i$$

where C is the regularization parameter, which balances the margin (smaller $w^T w$) and the penalty of mis-classification (smaller $\sum_{i=1}^{m} \epsilon_i$).

(a) (**True/False**) Number of support vectors do not depend on the selected value of C. Please provide a one-sentence justification.

Answer: False. Changing C will change the max-margin boundary lines.

In the figure below,



C=1

$n_{C=0}$ is the number of support vectors for C getting close to 0 ($\lim_{C \to 0}$) and $n_{C=\infty}$ is the number of support vectors for $C = \infty$.

(b) Select the correct option:( )

(1) $n_{C=0} > n_{C=\infty}$

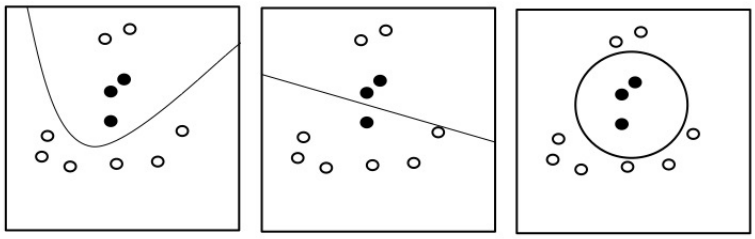
(2) $n_{C=0} < n_{C=\infty}$

(3) $n_{C=0} = n_{C=\infty}$

(4) Not enough information provided

Answer: (1). less penalty indicates wider margin in the figure.

(c) Match the following kernels used for SVM classification with the following figures:



(1) Linear Kernel : $K(x, x') = x^T x'$

(2) Polynomial Kernel (order = 2) : $K(x, x') = (1 + x^T x')^2$

(3) Radial Basis Kernel : $K(x, x') = exp(-\frac{1}{2}||x - x'||^2)$

Answer: (2) ; (1) ; (3)

Rank the best to the worse performing kernels in (c) (through just visual-inspection)

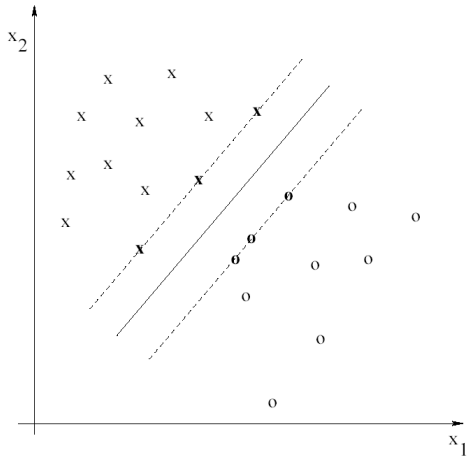
(1) Best:

(2) Middle:

(3) Worst:

Answer: (3) ; (2) ; (1)

(d) What is the leave-one-out cross-validation error for maximum margin separation in the following figure ? (we are asking for a number) Please provide a (at least one-sentence) justification.



Answer: 0. Removing any single point will not change the max-margin boundary. All points are correctly classified. LOOCV error is zero.

Question 2. Another Support Vector Machine

Consider a supervised learning problem in which the training examples are points in 2-dimensional space. The positive examples are $(1, 1)$ and $(-1, -1)$. The negative examples are $(1, -1)$ and $(-1, 1)$.

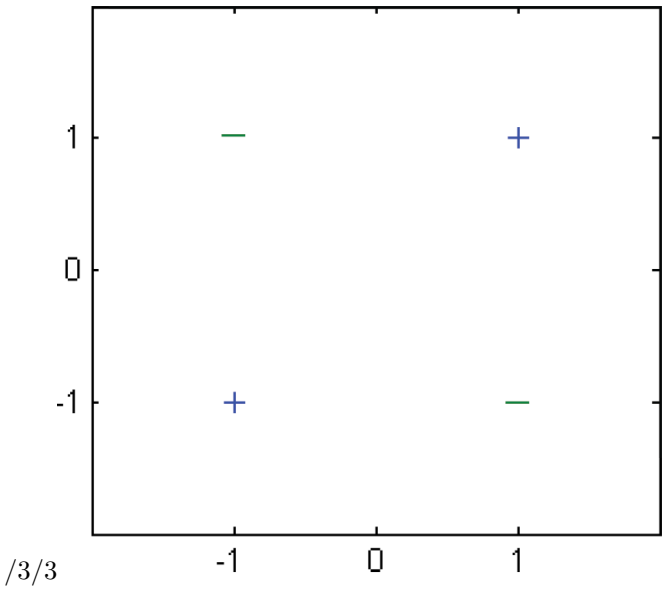
1. (1 pts) Are the positive examples linearly separable from the negative examples in the original space?

   Answer: **SOLUTION:** NO

2. (4 pts) Consider the feature transformation $\phi(x) = [1, x_1, x_2, x_1x_2]$, where x_1 and x_2 are, respectively, the first and second coordinates of a generic example x. The prediction function is $y(x) = w^T * \phi(x)$ in this feature space. Give the coefficients , w, of a maximum-margin decision surface separating the positive examples from the negative examples. (You should be able to do this by inspection, without any significant computation.)

   Answer: **SOLUTION:** $w = (0, 0, 0, 1)^T$

3. (3 pts) Add one training example to the graph so the total five examples can no longer be linearly separated in the feature space $\phi(x)$ defined in problem 5.2



/3/3

4. (4 pts) What kernel $K(x, x')$ does this feature transformation ϕ correspond to?

   Answer: **SOLUTION:** $1 + X_1X_1' + X_2X_2' + X_1X_2X_1'X_2'$