Assignment 2: Ridge Regression, SVM, and Model Selection

UVA CS 6316 :
Machine Learning (Fall 2015)

Out: Sept. 21 / Mon / 11pm, 2015
Due: Oct. 6 / Tue midnight 11:55pm, 2015 @ Collab

a *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) into PDF form.*

b *For questions and clarifications, please post on piazza. TA Ritas (rs3zz@virginia.edu) or Beilun (bw4mw@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.*

Please provide proper steps to show how you derive the answers.

Question 1. Ridge Regression (TA: Beilun)

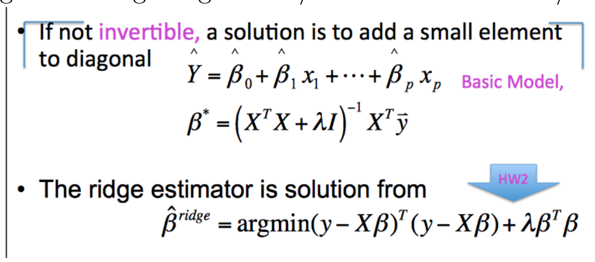
- Purpose 1: To emphasize the importance of selecting the right model through k-folds CV when using supervised regression.
- Purpose 2: To show a real case in which linear regression learns badly and adding regularization is necessary.

This problem provides a case study in which just using a linear regression model for data fitting is not enough. Adding regularization like ridge estimator is necessary for certain cases.

- Here we assume $X_{n\times p}$ represents a data sample matrix which has p features and n samples. $Y_{n\times 1}$ includes target variable's value of n samples. We use β to represent the coefficient. (Just a different notation. We had used θ for representing coefficient before.)
- 1.1 Please provide the math derivation procedure for ridge regression (shown in Figure)

  (Hint1: provide a procedure similar to how linear regression gets the normal equation through minimizing its loss function.)
  (Hint2: $\lambda|\beta|_2 = \lambda\beta^T\beta = \lambda\beta^T I\beta = \beta^T(\lambda I)\beta)$)
  (Hint3: Linear Algebra Handout Page 24, first two equations after the line "To recap,")

Figure 1: Ridge Regression / Solution Derivation / 1.1



- 1.2 Suppose $X = \begin{bmatrix} 1 & 2 \\ 3 & 6 \\ 5 & 10 \end{bmatrix}$ and $Y = [1, 2, 3]^T$, could this problem be solved through linear regression? Please provide your reasons.
  (Hint: just use the normal equation to explain)

- 1.3 If you have the prior knowledge that the coefficient β should be **sparse**, which regularized linear regression method should be chosen to use ? (Hint: sparse vector)

- A data file named "RRdata.txt" is provided. For this data, you are expected to write programs to compare between linear regression and ridge regression.

- Please submit your python code as "ridgeRegression.py" . Please use the following instructions and use required function names. Please use Numpy or other related package to implement the ridge regression. Other requirements or recommendations are the same as Homework1.

- Notation: The format of each row in data file is $[1, x_1, x_2, y]$, where x_1, x_2 are two features and y is the target value.

- 1.4 For "ridgeRregression.py",
  - Load the data file and assume the last column is the target value. You should use $xVal$ to represent the data sample matrix and $yVal$ to represent the target value vector.
  - 1.4.1 The first function is to implement the ridge regression and return the coefficient β with the hyperparameter $\lambda = 0$. (i.e. when $\lambda = 0$, it's just the standard linear regression). Please plot the data points and the learned plane [1]. Please submit the result into the writing part of this assignment. You are required to provide the following function (and module) for grading:

    $$betaLR = ridgeRegression.ridgeRegress(xVal, yVal, lambda = 0)$$

  - 1.4.2 The second function is to find the best λ by using a 10-fold cross validation procedure. The function should be,

    $$lambdaBest = ridgeRegression.cv(xVal, yVal)$$

  - You don't need to regularize the β_0. Instead, you can estimate β_0 by center the input(i.e. $\hat{\beta}_0 = \frac{\sum y_i}{n}$).
  - (Hint1: you should implement a function to split the data into ten folds; then loop over the folds; use one as test, the rest train)
  - (Hint2: for each fold, on the train part, perform ridgeRegress to learn β_k; Then use this β_k on all samples in the test fold to get predicted $\hat{y}$; Then calculate the error (difference) between true y and $\hat{y}$, sum over all testing points in the current fold k.)

[1] http://matplotlib.org/mpl_toolkits/mplot3d/tutorial.html#surface-plots

- Try all the λ values from the set: $\{0.02, 0.04, 0.06, \ldots, 1\}$ (i.e. $\{0.02i | i \in 1, 2, \ldots, 50\}$). Pick the λ achieving the best objective criterion from the 10-fold cross validation procedure. Our objective criterion is just the value of the loss function (i.e. $J(\theta)$ MSE in the slides) on each test fold. Please plot the λ versus $J(\beta)$ graph (which is also called path of finding the best λ) and provide it into the writing.
- Note : To constrain the randomness, please set seed to be 37. [2]
- Then run the ridge regression again by using the best λ calculated from 1.4.2. Please include the result into writing.

$$betaRR = ridgeRegression.ridgeRegress(xVal, yVal, lambdaBest)$$

- Please plot the data points and the learned plane from best ridge regression. Please include the result into writing. [3].

- 1.5 If assuming the true coefficient in problem 1.4 is $\beta = (3, 1, 1)^T$, could you compare and conclude whether linear regression or ridge regression performs better ? Explain why this happens based on the data we give.
  - (Hint: 1. Please implement a standard linear regression between x_1, x_2 and plot the x_1 versus x_2 graph;)
  - (Hint: 2. Guess the relationship between the two features and consider the problem 1.2.)

Question 2. Support Vector Machines with Scikit-Learn

(TA:Ritas)

Purpose: To emphasize the importance of selecting the right model when performing the SVM learning pipeline.

- (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 adult.data from collab.
- (2) For this assignment, you will create a program using scikit-learn's C-Support Vector Classifier.[4]
- 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 HW1 to help you import the data. Bear in mind you will also need to do some preprocessing of the data before applying the SVM.
- 2.1 You are required to provide the following function (and module) for grading:
  predictions = svmIncomeClassifier.processDataSet(adult.test)
- The 'adult.test' - a text file in the same format as the training "adult.data" file is supplied to you.
- 2.2 We will evaluate your output 'predictions' - an array of strings (">50K" or "<=50K") corresponding to the test file (i.e. 'adult.test'). Your models will be compared using the same test data set. So try to submit the best performing model that you can!
- 2.3 You need to report the classification accuracy results on train set and test set on 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 set into writing. For instance, you can summarize the results into a table with each row containing kernel choice, kernel parameter, train accuracy and test accuracy.

[2] More about random in python, please see, `https://docs.python.org/2/library/random.html`
[3] `http://matplotlib.org/mpl_toolkits/mplot3d/tutorial.html#surface-plots`
[4] Documented here: http://scikit-learn.org/stable/modules/classes.html#module-sklearn.svm

- (Hint: you can choose SVM kernels like, basic linear kernel / polynomial kernel, varying its parameters / RBF kernel, varying its parameters).

- 2.4 Please provide a one-sentence justification for the following TRUE/FALSE questions.

  (**True/False**) We would expect the support vectors to remain the same in general as we move from a linear kernel to higher order polynomial kernels.

Submission Instructions: You are required to submit the following :

1. A code that inputs both train and test data and includes the function:

   predictions = svmIncomeClassifier.processDataSet(adult.data,adult.test) (This processes the train and test data)

   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, it should be able to classify the test data and print out the classification score using function :

   print(yourClassifier.score())

2. A table in your PDF submission reporting classification accuracies (score) on the test data, along with details of the kernels, best performing hyperparameters used for training/testing 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 Q2.