Statistical Machine Learning GU4241/GR5241
Spring 2019
https://courseworks.columbia.edu/

Homework 4

Due: Thursday, Mar. 28th, 2019

Homework submission: Please submit your homework electronically through Gradescope (pdf file) and Canvas (code) by 11:59pm on the due date. You need to submit both the pdf file and your code (either in R or Python).

Problem 1 (Kernelized Nearest Neighbor Classification, 4 points)

Since k-nearest-neighbor classifier only computes distances between points, we will exploit this fact to make a kernel version of this classifier, using the same kernel trick that we did in SVM. You may now assume k is given. To derive *kernelized* nearest neighbor classifier, you need follow these steps:

1. Consider the squared Euclidean distance between any two points $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^p$. Expand the expression $d^2(\mathbf{x}, \mathbf{x}') = \|\mathbf{x} - \mathbf{x}'\|_2^2$ in terms of vector inner products.
2. Assume you are given a kernel $K(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle_{\mathcal{F}} : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$. Use the kernel trick to make the squared Euclidean distance from the previous part into a kernel squared distance. Write down this expression and call it $d_K^2(\mathbf{x}, \mathbf{x}')$.
3. d_K is itself a distance measure. What distance does it calculate?

Problem 2 (Subset Selection Methods, 6 points)

In this problem, we will compare different subset selection methods. We will study the `Credit` data set, which can be downloaded from CourseWorks. The data set records `balance` (average credit card debt) as well as several quantitative predictors: `age`, `cards` (number of credit cards), `education` (years of education), `income` (in thousands of dollars), `limit` (credit limit), and `rating` (credit rating). There are also four qualitative variables: `gender`, `student` (student status), `status` (marital status), and `ethnicity` (Caucasian, African American or Asian). We want to fit a regression model of `balance` on the rest of the variables.

- *(Best subset selection)* The `regsubsets()` function in R (part of the `leaps` library) performs the best subset selection by identifying the best model that contains a given number of predictors, where *best* is defined to be the one which minimizes the residual sum-of-squares (RSS).

  Here we need to represent the qualitative predictors by dummy variables. `gender`, `student` and `ethnicity` are all two-level categorical variables, and each of them is coded by one dummy variable. `ethnicity` takes on tree values and is coded by two dummy variables. Therefore, we have 11 predictors in total.

- *(Forward stepwise selection)* We can also use the `regsubsets()` function to perform forward stepwise selection, using the argument `method=''forward''`.

- *(Backward stepwise selection)* The `regsubsets()` function can be used to perform backward stepwise selection as well (`method=''backward''`). Here we start from the full model and at each step remove a predictor which leaves a model having smallest RSS.

- *(Choosing the optimal model)* After obtaining a set of models by using the subset selection approaches, we will choose a single best model which minimizes the prediction error. For this problem, we use C_p or BIC statistic as estimates of the prediction error (We will talk about this later in class). C_p statistic is defined by

$$C_p = \frac{1}{n}(\text{RSS} + 2p\hat{\sigma}^2),$$

where $\hat{\sigma}^2$ is an estimate of the variance of the error. BIC is defined by

$$\text{BIC} = \frac{1}{n}(\text{RSS} + \log(n)p\hat{\sigma}^2).$$

The summary() function returns RSS, C_p and BIC. You DO NOT need to compute them by yourselves.

Homework problems.

1. Apply the three subset selection methods mentioned above to Credit data set. Plot the RSS as a function of the number of variables for these three methods in the same figure.
2. Each subset selection method results in a set of models. For each approach, choose a single optimal model by using C_p and BIC statistics respectively. Report the optimal models for each approach (i.e. specify the predictors in the optimal model).

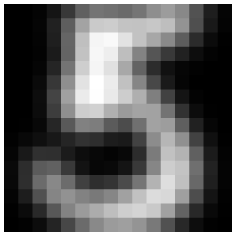
Remark. From this problem, you may notice that BIC tends to select a model with less predictors when compared to C_p.

Problem 3 (SVM, 10 points)

In this problem, we will apply a support vector machine to classify hand-written digits. You do not have to implement the SVM algorithm: The R library e1071 provides an implementation, see

`http://cran.r-project.org/web/packages/e1071/index.html`

Download the digit data set from the course website. The zip archive contains two files: Both files are text files. Each file contains a matrix with one data point (= vector of length 256) per row. The 256-vector in each row represents a 16×16 image of a handwritten number. The data contains two classes—the digits 5 and 6—so they can be labeled as -1 and $+1$, respectively. The image on the right shows the first row, re-arranged as a 16×16 matrix and plotted as a gray scale image.



- Randomly select about 20% of the data and set it aside as a test set.
- Train a linear SVM with soft margin. Cross-validate the margin parameter.
- Train an SVM with soft margin and RBF kernel. You will have to cross-validate both the soft-margin parameter and the kernel bandwidth.
- After you have selected parameter values for both algorithms, train each one with the parameter value you have chosen. Then compute the misclassification rate (the proportion of misclassified data points) on the test set.

Homework questions:

1. Plot the cross-validation estimates of the misclassification rate. Please plot the rate as
   (a) a function of the margin parameter in the linear case.
   (b) a function of the margin parameter and the kernel bandwidth in the non-linear case (you are encouraged to use heat map here).
2. Report the test set estimates of the misclassification rates for both cases, with the parameter values you have selected, and compare the two results. Is a linear SVM a good choice for this data, or should we use a non-linear one?