Homework #5

CS 260: Machine Learning Algorithms
Prof. Ameet Talwalkar
Due: 3/6/17, 10am
Please abide by the Academic Integrity Policy

1 Bias-Variance Tradeoff

Consider a dataset with n data points $(\mathbf{x}_i, y_i)$, $\mathbf{x}_i \in \mathbb{R}^{p\times 1}$, drawn from the following linear model:

$$y = \mathbf{x}^\top \boldsymbol{\beta}^\star + \varepsilon,$$

where ε is a Gaussian noise and the star sign is used to differentiate the true parameter from the estimators that will be introduced later. Consider the L_2 regularized linear regression as follows:

$$\widehat{\boldsymbol{\beta}}_\lambda = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2 + \lambda \|\boldsymbol{\beta}\|_2^2 \right\},$$

where $\lambda \geq 0$ is the regularization parameter. Let $X \in \mathbb{R}^{n\times p}$ denote the matrix obtained by stacking $\mathbf{x}_i^\top$ in each row. Properties of an affine transformation of a gaussian random variable will be useful throughout this problem.

a. Find the closed form solution for $\widehat{\boldsymbol{\beta}}_\lambda$ and its distribution.

b. Calculate the bias term $\mathbb{E}[\mathbf{x}^\top \widehat{\boldsymbol{\beta}}_\lambda] - \mathbf{x}^\top \boldsymbol{\beta}^\star$ as a function of λ and some *fixed* test point $\mathbf{x}$.

c. Calculate the variance term $\mathbb{E}\left[\left(\mathbf{x}^\top \widehat{\boldsymbol{\beta}}_\lambda - \mathbb{E}[\mathbf{x}^\top \widehat{\boldsymbol{\beta}}_\lambda]\right)^2\right]$ as a function of λ and some *fixed* test point $\mathbf{x}$.

d. Use the results from parts (b) and (c) and the bias–variance theorem to analyze the impact of λ in the squared error. Specifically, which term dominates when λ is small or large?

2 Kernels

Mercer's theorem implies that a bivariate function $k(\cdot,\cdot)$ is a positive definite kernel function iff, for any N and any $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N$, the corresponding kernel matrix K is positive semidefinite, where $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$. Recall that a matrix $A \in \mathbb{R}^{n\times n}$ is positive semidefinite if all of its eigenvalues are non-negative, or equivalently, if $\mathbf{x}^\top A \mathbf{x} \geq 0$ for arbitrary vector $\mathbf{x} \in \mathbb{R}^{n\times 1}$.

Suppose $k_1(\cdot,\cdot)$ and $k_2(\cdot,\cdot)$ are positive definite kernel functions with corresponding kernel matrices K_1 and K_2. Use Mercer's theorem to show that the following kernel functions are positive definite.

a. $K_3 = a_1 K_1 + a_2 K_2$, for $a_1, a_2 \geq 0$.

b. K_4 defined by $k_4(\mathbf{x}, \mathbf{x}') = f(\mathbf{x}) f(\mathbf{x}')$ where $f(\cdot)$ is an arbitrary real valued function.

c. K_5 defined by $k_5(\mathbf{x}, \mathbf{x}') = k_1(\mathbf{x}, \mathbf{x}') k_2(\mathbf{x}, \mathbf{x}')$.

3 Soft Margin Hyperplanes

The function of the slack variables used in the optimization problem for soft margin hyperplanes has the form: $\xi \mapsto \sum_{i=1}^{n} \xi_i$. Instead, we could use $\xi \mapsto \sum_{i=1}^{n} \xi_i^p$, with $p > 1$.

a. Give the dual formulation of the problem in this general case.

b. How does this more general formulation ($p > 1$) compare to the standard setting ($p = 1$) discussed in lecture? Is the general formulation more or less complex? Justify your answer.

4 Programming

In this problem, you will experiment with SVMs on a real-world dataset. You will implement a linear SVM (i.e., an SVM using the original features. You will also use a widely used SVM toolbox called LibSVM to experiment with kernel SVMs.

Dataset: We have provided the *Splice Dataset* from UCI's machine learning data repository.[1] The provided binary classification dataset has 60 input features, and the training and test sets contain 1,000 and 2,175 samples, respectively (the files are called `splice_train.mat` and `splice_test.mat`).

4.1 Data preprocessing

Preprocess the training and test data by

a. computing the mean of each dimension and subtracting it from each dimension

b. dividing each dimension by its standard deviation

Notice that the mean and standard deviation should be estimated from the *training data* and then applied to both datasets. Explain why this is the case. Also, report the mean and the standard deviation of the third and 10th features on the test data.

4.2 Implement linear SVM

Please fill in the Matlab functions `trainsvm` in `trainsvm.m` and `testsvm.m` in `testsvm.m`.

The input of `trainsvm` contain training feature vectors and labels, as well as the tradeoff parameter C. The output of `trainsvm` contain the SVM parameters (weight vector and bias). In your implementation, you need to solve SVM in its primal form

$$\begin{aligned} \min_{\boldsymbol{w},b,\boldsymbol{\xi}} \ & \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_i \xi_i \\ \text{s.t.} \quad & y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1 - \xi_i, \forall i \\ & \xi_i \geq 0, \forall i \end{aligned}$$

Please use the `quadprog` function in Matlab to solve the above quadratic problem.

For `testsvm`, the input contains testing feature vectors and labels, as well as SVM parameters. The output contains the test accuracy.

[1] https://archive.ics.uci.edu/ml/datasets/Molecular+Biology+%28Splice-junction+Gene+Sequences%29.

4.3 Cross validation for linear SVM

Use 5-fold cross validation to select the optimal C for your implementation of linear SVM.

a. Report the cross-valiation accuracy (averaged accuracy over each validation set) and average training time (averaged over each training subset) on different C taken from $\{4^{-6}, 4^{-5}, \cdots, 4, 4^2\}$. How does the value of C affect the cross validation accuracy and average training time? Explain your observation.

b. Which C do you choose based on the cross validation results?

c. For the selected C, report the test accuracy.

4.4 Use linear SVM in LibSVM

LibSVM is widely used toolbox for SVMs, and it has a Matlab interface. Download LibSVM from `http://www.csie.ntu.edu.tw/~cjlin/libsvm/` and install it according to the README file (make sure to use the Matlab interface provided in the LibSVM toolbox). For each C from $\{4^{-6}, 4^{-5}, \cdots, 4, 4^2\}$, apply 5-fold cross validation (use `-v` option in LibSVM) and report the cross validation accuracy and average training time.

a. Is the cross validation accuracy the same as that in 4.3? Note that LibSVM solves linear SVM in dual form while your implementation does it in primal form.

b. How does LibSVM compare with your implementation in terms of training time?

4.5 Use kernel SVM in LibSVM

LibSVM supports a number of kernel types. Here you need to experiment with the polynomial kernel and RBF (Radial Basis Function) kernel.

a. **Polynomial kernel**. Please tune C and `degree` in the kernel. For each combination of (C, `degree`), where $C \in \{4^{-4}, 4^{-3}, \cdots, 4^6, 4^7\}$ and `degree` $\in \{1, 2, 3\}$, report the 5-fold cross validation accuracy and average training time.

b. **RBF kernel**. Please tune C and `gamma` in the kernel. For each combination of (C, `gamma`), where $C \in \{4^{-4}, 4^{-3}, \cdots, 4^6, 4^7\}$ and `gamma` $\in \{4^{-7}, 4^{-6}, \cdots, 4^1, 4^2\}$, report the 5-fold cross validation accuracy and average training time.

Based on the cross validation results of Polynomial and RBF kernel, which kernel type and kernel parameters will you choose? Report the corresponding test accuracy for the configuration with the highest cross validation accuracy.

Submission Instructions

- Provide your answers to problems 1-3, 4.1, and 4.3-4.5 in hardcopy. The papers need to be stapled, and submitted at the beginning of class on the due date.
- For problem 4, please put all of your code in a single folder named `[lastname]_[firstname]_hw5`, and submit a single `.zip` file containing this folder called `[lastname]_[firstname]_hw5.zip` to **CCLE** by the due date. The only acceptable programming languages are MATLAB and Octave.
- You are encouraged to collaborate, but collaboration must be limited to discussion only and you need to write down / implement your solution on your own. You also need to list with whom you have discussed the HW problems.