Mehryar Mohri
Foundations of Machine Learning 2014
Courant Institute of Mathematical Sciences
Homework assignment 3
October 28, 2014
Due: November 14, 2014

A. Kernels

1. Show that the kernel K defined by

$$\forall (x, y) \in \mathbb{R}^N \times \mathbb{R}^N, \; K(x, y) = \frac{1}{1 + \frac{\|x-y\|^2}{\sigma^2}} \;, \tag{1}$$

where $\sigma > 0$ is a parameter, is PDS (*hint*: the function $x \mapsto \int_0^{+\infty} e^{-sx} e^{-s} ds$ defined for all $x \geq 0$ could be useful for the proof).

2. Show that the kernel K defined by

$$\forall (x, y) \in \mathbb{R}^N \times \mathbb{R}^N, \; K(x, y) = \exp\left(\frac{\sum_{i=1}^N \min(|x_i|, |y_i|)}{\sigma^2} \right) \;, \tag{2}$$

where $\sigma > 0$ is a parameter, is PDS (*hint*: the function $(x_0, y_0) \mapsto \int_0^{+\infty} 1_{t \in [0, |x_0|]} 1_{t \in [0, |y_0|]} dt$ defined over $\mathbb{R} \times \mathbb{R}$ could be useful for the proof).

B. Support Vector Machines

1. Download and install the `libsvm` software library from:

`http://www.csie.ntu.edu.tw/~cjlin/libsvm/` ,

and briefly consult the documentation to become more familiar with the tools.

2. Consider the `splice` data set

`http://www.cs.toronto.edu/~delve/data/splice/desc.html`.

Download the already formatted training and test files of a noisy version of that dataset from

`http://www.cs.nyu.edu/~mohri/ml14/splice_noise_train.txt`
`http://www.cs.nyu.edu/~mohri/ml14/splice_noise_test.txt`.

Use the `libsvm` scaling tool to scale the features of all the data. The scaling parameters should be computed only on the training data and then applied to the test data.

3. Consider the corresponding binary classification which consists of distinguishing two types of splice junctions in DNA sequences using about 60 features. Use SVMs combined with polynomial kernels to tackle this problem.

   To do that, randomly split the training data into ten equal-sized disjoint sets. For each value of the polynomial degree, $d = 1, 2, 3, 4$, plot the average cross-validation error plus or minus one standard deviation as a function of C (let other parameters of polynomial kernels in `libsvm` be equal to their default values), varying C in powers of 5, starting from a small value $C = 5^{-k}$ to $C = 5^k$, for some value of k. k should be chosen so that you see a significant variation in training error, starting from a very high training error to a low training error. Expect longer training times with `libsvm` as the value of C increases.

4. Let (C^*, d^*) be the best pair found previously. Fix C to be C^*. Plot the ten-fold cross-validation error and the test errors for the hypotheses obtained as a function of d. Plot the average number of support vectors obtained as a function of d. How many of the support vectors lie on the marginal hyperplanes? Plot the soft margin of the solution as a function of d.

5. Now, combine SVMs with Gaussian kernels to tackle the same task. Use cross-validation as before to determine the best value of C and σ, varying C in powers of 5, and σ in powers of 2 for a reasonable range so that you see a significant variation in training error, as before. Fix C and σ to the best values found via cross-validation. How does the test error of the solution compare to the best result obtained using polynomial kernels? What is the value of the soft margin?

C. Boosting

As discussed in class, AdaBoost can be viewed as coordinate descent applied to an exponential objective function. Here, we consider an alternative ensemble method algorithm, HingeBoost, that consists of applying coordinate descent to an objective function based on the hinge loss. Using the same notation as in class, consider the function F defined for all $\boldsymbol{\alpha} \in \mathbb{R}^N$ by

$$F(\boldsymbol{\alpha}) = \sum_{i=1}^{m} \max\left(0, 1 - y_i \sum_{j=1}^{N} \alpha_j h_j(x_i)\right), \qquad (3)$$

where the h_js are base classifiers belonging to a hypothesis set H of functions taking values -1 or $+1$.

1. Show that F is convex and admits a right- and left-derivative along any direction.

2. For any $j \in [1, N]$, let $\mathbf{e}_j$ denote the direction corresponding to the base hypothesis h_j. Let $\boldsymbol{\alpha}_t$ denote the vector of coefficients $\alpha_{t,j}$, $j \in [1, N]$ obtained after $t \geq 0$ iterations of coordinate descent and $f_t = \sum_{j=1}^{N} \alpha_{t,j} h_j$ the predictor obtained after t iterations.

   Give the expression of the right-derivative $F'_+(\boldsymbol{\alpha}_{t-1}, \mathbf{e}_j)$ and the left-derivative $F'_-(\boldsymbol{\alpha}_{t-1}, \mathbf{e}_j)$ after $t-1$ iterations in terms of f_{t-1}.

3. For any $j \in [1, N]$, define the maximum directional derivative $\delta F(\boldsymbol{\alpha}_{t-1}, \mathbf{e}_j)$ at $\boldsymbol{\alpha}_{t-1}$ as follows:

$$\delta F(\boldsymbol{\alpha}_{t-1}, \mathbf{e}_j) =$$
$$\begin{cases} 0 & \text{if } F'_-(\boldsymbol{\alpha}_{t-1}, \mathbf{e}_j) \leq 0 \leq F'_+(\boldsymbol{\alpha}_{t-1}, \mathbf{e}_j) \\ F'_+(\boldsymbol{\alpha}_{t-1}, \mathbf{e}_j) & \text{if } F'_-(\boldsymbol{\alpha}_{t-1}, \mathbf{e}_j) \leq F'_+(\boldsymbol{\alpha}_{t-1}, \mathbf{e}_j) \leq 0 \\ F'_-(\boldsymbol{\alpha}_{t-1}, \mathbf{e}_j) & \text{if } 0 \leq F'_-(\boldsymbol{\alpha}_{t-1}, \mathbf{e}_j) \leq F'_+(\boldsymbol{\alpha}_{t-1}, \mathbf{e}_j). \end{cases}$$

   The direction $\mathbf{e}_j$ considered by the coordinate descent considered here is the one maximizing $|\delta F(\boldsymbol{\alpha}_{t-1}, \mathbf{e}_j)|$. Once the best direction j is selected, the step η can be determined by minimizing $F(\boldsymbol{\alpha}_{t-1} + \eta \mathbf{e}_j)$ using a grid search. Give the pseudocode of HingeBoost.

4. Bonus question: implement HingeBoost and AdaBoost using as base classifiers boosting stumps and compare their performance on the data set of Problem B. The number of rounds of boosting T can be determined via cross-validation varying T in powers of 10. Report the test errors of each algorithm and compare them with those obtained for SVMs in problem B.