Mehryar Mohri
Foundations of Machine Learning 2015
Courant Institute of Mathematical Sciences
Homework assignment 2
October 23, 2015
Due: November 09, 2015

A. VC-dimension of convex combinations

1. Let H be a family of functions mapping from an input space $\mathcal{X}$ to $\{-1, +1\}$ and let T be a positive integer. Give an upper bound on the VC-dimension of the family of functions $\mathcal{F}_T$ defined by

$$\mathcal{F} = \left\{ \operatorname{sgn}\left(\sum_{t=1}^{T} \alpha_t h_t \right) : h_t \in H, \alpha_t \geq 0, \sum_{t=1}^{T} \alpha_t \leq 1 \right\}.$$

   (*Hint*: you can use Problem C. of (Foundations of Machine Learning, HW2, 2014, `http://www.cs.nyu.edu/~mohri/ml14/hw2.pdf` and its solution).

B. Growth function

1. A *linearly separable labeling* of a set X of vectors in $\mathbb{R}^d$ is a classification of X into two sets X^+ and X^- with $X^+ = \{\mathbf{x} \in X \colon \mathbf{w} \cdot \mathbf{x} > 0\}$ and $X^- = \{\mathbf{x} \in X \colon \mathbf{w} \cdot \mathbf{x} < 0\}$ for some $\mathbf{w} \in \mathbb{R}^d$.

   Let $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ be a subset of $\mathbb{R}^d$.

   (a) Let $\{X^+, X^-\}$ be a dichotomy of X and let $\mathbf{x}_{m+1} \in \mathbb{R}^d$. Show that $\{X^+ \cup \{\mathbf{x}_{m+1}\}, X^-\}$ and $\{X^+, X^- \cup \{\mathbf{x}_{m+1}\}\}$ are linearly separable by a hyperplane going through the origin if and only if $\{X^+, X^-\}$ is linearly separable by a hyperplane going through the origin and $\mathbf{x}_{m+1}$.

   (b) Let $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ be a subset of $\mathbb{R}^d$ such that any k-element subset of X with $k \leq d$ is linearly independent. Then, the number of linearly separable labelings of X is $C(m, d) = 2 \sum_{k=0}^{d-1} \binom{m-1}{k}$. (*Hint*: prove by induction that $C(m+1, d) = C(m, d) + C(m, d-1)$.

   (c) Let $f_1, \ldots, f_p$ be p functions mapping $\mathbb{R}^d$ to $\mathbb{R}$. Define $\mathcal{F}$ as the family of classifiers based on linear combinations of these

functions:

$$\mathcal{F} = \left\{ x \mapsto \operatorname{sgn}\left(\sum_{k=1}^{p} a_k f_k(x) \right) : a_1, \ldots, a_p \in \mathbb{R} \right\}.$$

Define Ψ by $\Psi(x) = (f_1(x), \ldots, f_p(x))$. Assume that there exists $x_1, \ldots, x_m \in \mathbb{R}^d$ such that every p-subset of $\{\Psi(x_1), \ldots, \Psi(x_m)\}$ is linearly independent. Then, show that

$$\Pi_{\mathcal{F}}(m) = 2 \sum_{i=0}^{p-1} \binom{m-1}{i}.$$

C. 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/ml15/splice_noise_train.txt`
   `http://www.cs.nyu.edu/~mohri/ml15/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, $d = 1, 3, 5$, 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?

6. Here, use as a kernel the sum of the best polynomial kernel (degree d^*) and the Gaussian kernel with the best parameter σ you found in the previous question. Use cross-validation as before to determine the best value of C. How does the test error of the solution compare to the best result obtained in the previous questions?

D. Kernels

Show that the following kernels are PDS.

1. Let n be a positive integer. K is defined by $K(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{N} \cos^n(x_i^2 - y_i^2)$ for all $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^N \times \mathbb{R}^N$.

2. Let σ be a positive real number. K is defined by $K(x, y) = e^{-\frac{\|\mathbf{x}-\mathbf{y}\|}{\sigma}}$ for all $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^N \times \mathbb{R}^N$ (*Hint*: you could show that K is the normalized kernel of a kernel K' and show that K' is PDS using the following equality: $\|\mathbf{x} - \mathbf{y}\| = \frac{1}{2\Gamma(\frac{1}{2})} \int_0^{+\infty} \frac{1-e^{-t\|\mathbf{x}-\mathbf{y}\|^2}}{t^{\frac{3}{2}}} dt$ valid for all $\mathbf{x}, \mathbf{y}$).