Mehryar Mohri
Foundations of Machine Learning 2018
Courant Institute of Mathematical Sciences
Homework assignment 3
11/10, 2018
Due: 11/26, 2018

A. Kernel PCA

Read the *Dimensionality Reduction* Chapter 12 in the course textbook Foundations of ML with a focus on PCA and Kernel PCA. Sections 12.1 and 12.2 are recommended. In this problem we will analyze a hypothesis set based on KPCA projection. Let $K(x, y)$ be a kernel function, $\Phi_K(x)$ be its corresponding feature map and $S = \{x_1, \ldots, x_m\}$ be a sample of m points. When Π is the rank-r KPCA projection, we define the (regularized) hypothesis set of linear separators in the RKHS $\mathbb{H}$ of kernel K as

$$H = \left\{ x \to \langle w, \Pi\Phi_K(x)\rangle_{\mathbb{H}} : \|w\|_{\mathbb{H}} \leq 1 \right\}. \tag{1}$$

This hypothesis set essentially means that the input data is projected onto a smaller dimensional subspace of the RKHS before fitting a separation hyperplane. This problem will show that we can use the eigenvectors and eigenvalues of the sample kernel matrix to give a closed form expression for the functions $h \in H$ without a need for explicit representation of the RKHS itself.

Let $\mathbf{K}$ be the sample kernel matrix for kernel K evaluated on m points of sample S, that is $\mathbf{K}_{i,j} = K(x_i, x_j)$. Let $\lambda_1, \ldots, \lambda_r$ are the top r (nonzero) eigenvalues of $\mathbf{K}$ with the corresponding eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_r$. Denote the j-th element of vector $\mathbf{v}_i$ as $[\mathbf{v}_i]_j$. Follow the subproblems below to derive the explicit representation of $h \in H$.

1. Assume that the feature maps $\Phi_K(x)$ are centered on sample S and recall that the sample covariance operator is $\Sigma = \sum_{i=1}^m \frac{1}{m} \Phi_K(x_i)\Phi_K(x_i)^\top$. Prove that $h(x) = \sum_{i=1}^r \alpha_i \langle \mathbf{u}_i, \Phi_K(x)\rangle_{\mathbb{H}}$ for some $\alpha_i \in \mathbb{R}$, where $\mathbf{u}_1, \cdots, \mathbf{u}_r$ are the eigenvectors of Σ corresponding to its top r eigenvalues.

2. Prove that $\mathbf{u}_i = \mathbf{X}\frac{\mathbf{v}_i}{\sqrt{\lambda_i}}$, where $\mathbf{X} = [\Phi_K(x_1), \ldots, \Phi_K(x_m)]$

3. Using the result above, prove that any function $h \in H$ can be represented as
$$h(x) = \sum_{i=1}^{r} \sum_{j=1}^{m} \frac{\alpha_i}{\sqrt{\lambda_i}} K(x_j, x)[v_i]_j,$$
for some $\alpha_i \in \mathbb{R}$.

4. Bonus question: derive the Rademacher complexity bound on the hypothesis set H defined in this problem.

B. Multi-class boosting

Lecture 10 introduces the AdaBoost.MH algorithm, which is AdaBoost for multi-class classification. (Consult with Lecture 10's slides if you are unfamiliar with multi-class learning setting.) AdaBoost.MH is defined by objective function $F(\alpha)$:

$$F(\alpha) = \sum_{l=1}^{k} \sum_{i=1}^{m} e^{-y_i[l] \sum_{t=1}^{n} \alpha_t h_t(x_i, l)},$$

where $y_i \in \mathcal{Y} = \{-1, +1\}^k$, and $y_i[l]$ denotes the l-th coordinate of y_i for any $i \in [m]$ and $l \in [k]$. The base classifiers come from $H = \{h : \mathcal{X} \times [k] \to \{-1, +1\}\}$. Consider an alternative objective function for the same problem:

$$G(\alpha) = \sum_{i=1}^{m} e^{-\frac{1}{k} \sum_{l=1}^{k} y_i[l] \sum_{t=1}^{n} \alpha_t h_t(x_i, l)}.$$

1. Compare $G(\alpha)$ with $F(\alpha)$. Show that $F(\alpha) \geq kG(\alpha)$.

2. Let $g_n(x_i, l) = \sum_{t=1}^{n} \alpha_t h_t(x_i, l)$. Assume that $|g_n(x_i, l)| \leq 1$ for all $x_i \in \mathcal{X}, l \in [k]$. Show that $kG(\alpha)$ is a convex function upper bounding the multi-label multi-class error:
$$\sum_{i=1}^{m} \sum_{l=1}^{k} 1_{y_i[l] \neq \text{sgn}(g_n(x_i, l))} \leq kG(\alpha).$$

3. Drive an algorithm defined by the application of coordinate descent to $G(\alpha)$. You should give a full description of your algorithm, including the pseudocode, details for the choice of the step and direction, as well as a generalization bound.