

## A. Kernel PCA

Read the *Dimensionality Reduction* Chapter 15 in the course textbook Foundations of ML with a focus on PCA and Kernel PCA. Sections 15.1 and 15.2 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  $S = \{x_1, \dots, x_m\}$  be a sample of  $m$  points. When  $\Pi$  is the rank- $r$  KPCA projection, we define the (regularized) hypothesis set of linear separators in the RKHS  $\mathbb{H}$  of kernel  $K$  as

$$\mathcal{H} = \{x \mapsto \langle \mathbf{w}, \Pi \Phi_K(x) \rangle_{\mathbb{H}} : \|\mathbf{w}\|_{\mathbb{H}} \leq 1\}. \quad (1)$$

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  $h \in \mathcal{H}$  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  $\lambda_1, \dots, \lambda_r$  be the top  $r$  (nonzero) eigenvalues of  $\mathbf{K}$  with the corresponding eigenvectors  $\mathbf{v}_1, \dots, \mathbf{v}_r$ . 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  $h \in \mathcal{H}$ .

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  $\mathbf{u}_1, \dots, \mathbf{u}_r$  are the eigenvectors of  $\Sigma$  corresponding to its top  $r$  eigenvalues.
2. Prove that  $\mathbf{u}_i = \mathbf{X} \frac{\mathbf{v}_i}{\sqrt{\lambda_i}}$ , where  $\mathbf{X} = [\Phi_K(x_1), \dots, \Phi_K(x_m)]$ .
3. Using the result above, prove that any function  $h \in \mathcal{H}$  can be represented as

$$h(x) = \sum_{i=1}^r \sum_{j=1}^m \frac{\alpha_i}{\sqrt{\lambda_i}} K(x_j, x) [\mathbf{v}_i]_j,$$

for some  $\alpha_i \in \mathbb{R}$ .

4. Bonus question: derive the Rademacher complexity bound on the hypothesis set  $\mathcal{H}$  defined in this problem.

## B. Boosting

1. Implement AdaBoost with boosting stumps and apply the algorithm to the `spambase` dataset

<http://archive.ics.uci.edu/ml/datasets/Spambase>.

Download a shuffled version of that dataset (will be sent by email). Scale the features of all the data. Use the first 3450 examples for training, the last 1151 for testing. The scaling parameters should be computed only on the training data and then applied to the test data.

Consider the binary classification that consists of predicting if the e-mail message is a spam using the 57 features. Randomly split the training data into ten equal-sized disjoint sets. Plot the average cross-validation error plus or minus one standard deviation as a function of the number of rounds of boosting  $T$  by selecting the value of this parameter out of  $\{10, 10^2, \dots, 10^k\}$  for a suitable value of  $k$ . Let  $T^*$  be the best value found for the parameter. Plot the error on the training and test set as a function of the number of rounds of boosting for  $t \in [1, T^*]$ .

2. Consider the following variant of the classification problem where, in addition to the positive and negative labels  $+1$  and  $-1$ , points may be labeled with  $0$ . This can correspond to cases where the true label of a point is unknown, a situation that often arises in practice, or more generally to the fact that the learning algorithm incurs no loss for predicting  $-1$  or  $+1$  for such a point. Let  $\mathcal{X}$  be the input space and let  $\mathcal{Y} = \{-1, 0, +1\}$ . As in standard binary classification, the loss of  $f: \mathcal{X} \rightarrow \mathbb{R}$  on a pair  $(x, y) \in \mathcal{X} \times \mathcal{Y}$  is defined by  $1_{yf(x) < 0}$ .

Consider a sample  $S = ((x_1, y_1), \dots, (x_m, y_m)) \in (\mathcal{X} \times \mathcal{Y})^m$  and a hypothesis set  $\mathcal{H}$  of base functions taking values in  $\{-1, 0, +1\}$ . For a base hypothesis  $h_t \in \mathcal{H}$  and a distribution  $\mathcal{D}_t$  over indices  $i \in [1, m]$ , define  $\epsilon_t^s$  for  $s \in \{-1, 0, +1\}$  by  $\epsilon_t^s = \mathbb{E}_{i \sim \mathcal{D}_t} [1_{y_i h_t(x_i) = s}]$ .

- Derive a boosting-style algorithm for this setting in terms of  $\epsilon_t^s$ s, using the same objective function as that of AdaBoost. You should carefully justify the definition of the algorithm.
  - What is the weak learning condition in this setting?
  - Write the full pseudocode of the algorithm.
  - Give an upper bound on the training error of the algorithm as a function of the number of rounds of boosting and  $\epsilon_t^s$ s.
3. Let  $\mathcal{X}$  be the input space and  $\mathcal{Y} = \{-1, +1\}$  the binary label space. Consider the exponential loss used in AdaBoost:  $\ell(h, x, y) = \exp(-yh(x))$ . Given a distribution  $\mathcal{D}$  over  $\mathcal{X} \times \mathcal{Y}$ , the Bayes error for the exponential loss  $\ell$  is defined as the infimum of the errors achieved by measurable functions  $h: \mathcal{X} \rightarrow \mathbb{R}$ :

$$R_\ell^* = \inf_{h: \mathcal{X} \rightarrow \mathbb{R} \text{ measurable}} R_\ell(h),$$

where  $R_\ell(h) = \mathbb{E}_{(x, y) \sim \mathcal{D}} [\ell(h, x, y)]$ . A hypothesis  $h_{\text{exp}}$  with  $R_\ell(h_{\text{exp}}) = R_\ell^*$  is called a Bayes optimal solution. Define  $\eta(x) = \mathbb{P}[y = +1|x]$ .

- Give the expression of the Bayes optimal solution  $h_{\text{exp}}$  for the exponential loss in terms of  $\eta(x)$ .
- Define the generalization error and the Bayes error for the binary classification loss as follows:

$$R(h) = \mathbb{E}_{(x, y) \sim \mathcal{D}} [1_{\text{sign}(h(x)) \neq y}], \quad R^* = \inf_{h: \mathcal{X} \rightarrow \mathbb{R} \text{ measurable}} R(h),$$

where  $\text{sign}(t) = 1_{t \geq 0} - 1_{t < 0}$ . Show that  $R(h_{\text{exp}}) = R^*$ .