Homework 5

Due February 19th, 2020 by 11:59pm

Instructions: Upload your answers to the questions below to Canvas. Submit the answers to the questions in a PDF file and your code in a (single) separate file. Be sure to comment your code to indicate which lines of your code correspond to which question part. The homework assignment consists of 1 reading assignment, 2 graded exercises, and 1 optional non-graded exercise. Note that the homework assignment is shorther to give you more time to read and review before the midterm exam.

Reading Assignment

Read the article entitled "Kernel Methods for Machine Learning" by Hofmann, Schölkopf, and Smola [1].

1 Exercise 1

We consider the supervised multi-class classification problem with c classes. Assume you have a training sample $(x_1, y_1), ..., (x_n, y_n) \in \mathcal{X} \times \{-1, +1\}^c$. The goal is predict the labels on unseen examples using *your own nearest-mean-element classifier* for multi-class classification.

Consider the pre-processed version of the Labeled Faces in the Wild dataset (LFW). Pick several classes of your choice with $c \geq 3$. Create your own dataset from these classes (training set and test set).

We view here each image as a *bag of patches* that is a bag of all small square regions of size $p \times p$ pixels where p is an odd number. Consider two images I and I', with resp. $\{P(z_1), \ldots, P(z_N)\}$ and $\{P^{'}(z_1), \ldots, P^{'}(z_N)\}$ the corresponding bags of patches and $z_1, \ldots, z_N$ the corresponding patch locations. Define the kernel between two images

$$\kappa\left(I, I^{'}\right) = \sum_{i=1}^{N} \sum_{j=1}^{N} h_{\text{loc}}(z_i, z_j) h_{\text{patch}}\left(\tilde{P}(z_i), \tilde{P}^{'}(z_j)\right) \ ,$$

where $\tilde{P}_i = P_i / \|P_i\|_2$ and $\tilde{P}'_i = P'_i / \|P'_i\|_2$ resp., $h_{\text{patch}}(\cdot, \cdot)$ is the Gaussian kernel with hyper-parameter σ_p and $h_{\text{loc}}(\cdot, \cdot)$ is the the Gaussian kernel with hyper-parameter σ_l.

Write a function that compute that evaluates the kernel κ for any pair of images from the LFW dataset. You may pick the patch-size p and the set of patch locations $\{z_1, \ldots, z_N\}$ as you feel best.

Plot the (multi-class) misclassification error on the test set of *your own nearest-mean-element classifier* with the kernel κ on the test set versus the hyper-parameters σ_p and σ_l. What do you observe?

2 Exercise 2

You will generate simulated data and then perform Kernel Principal Component Analysis (KPCA) on the data. You will write *your own Power Iteration algorithm* for KPCA.

(a) Generate a simulated dataset consisting of data sampled around 2 concentric ellipses. Each class consist of 40 datapoints sampled around an ellipse. To generate a datapoint from a class, you may add Gaussian noise to a point lying on the corresponding ellipse.

(b) Write *your own Power Iteration algorithm*. You may set maxiter = 700 in your stopping criterion. Run your algorithm on the simulated dataset to compute the leading kernel principal component with your favorite kernel.

(c) Write *your own KPCA algorithm* that computes the k kernel principal components from a sample with your favorite kernel. Run your KPCA algorithm with your favorite kernel on the 100 observations to compute the two leading kernel principal components.

(d) Visualize the dataset projected on the two leading *kernel principal components* (KPCA), highlighting the datapoints in each of the two classes. Play with the kernel hyperparameter. Compare with the visualization of the dataset projected on the two leading *principal components* (PCA). What do you observe?

3 Exercise 3

Consider $\mathcal{Y} = \mathbb{R}$ with linear kernel $h(y, y') = yy'$ and $\mathcal{X}$ to be a set equipped with some psd kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Denote by $\mathcal{F}$ and $\mathcal{G}$ the RKHS associated to resp. k and h. Consider two random variables X, Y on resp. $\mathcal{X}, \mathcal{Y}$. Kernel ridge regression aims to solve a regularized least squares using non-linear functions defined by the kernel k. Given a sample of i.i.d pairs $(x_i, y_i)_{i=1}^n$ drawn from the distribution of (X, Y), kernel ridge regression amounts to solving

$$\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda \|f\|_{\mathcal{F}}^2 \tag{1}$$

In the following we assume the data to be centered i.e. such that $\hat{\mu}_X = 0$ and $\hat{\mu}_Y = 0$ with $\hat{\mu}_X$ and $\hat{\mu}_Y$ the empirical mean elements of X and Y respectively.

1. Derive the solution of this problem in terms of $(k(x_i, \cdot))_{i=1}^n$, the gram matrix $K = (k(x_i, x_j))_{i,j=1}^n$ and $y = (y_i)_{i=1}^n$. *Hint: Can we invoke the Representer Theorem?*

2. Write problem (1) in terms of the empirical covariance operators of X and Y.

3. Solve the problem in terms of the function $f \in \mathcal{F}$ using the formulation in terms of covariance operators. *Hint: You can admit that the solution of the problem is the unique function whose functional gradient is zero.*

4. Formulate kernel CCA on the samples $(x_i, y_i)_{i=1}^n$ using the empirical covariance operators (with the regularization).

5. Solve kernel CCA.

   Hints:

   (a) Without loss of generality you can assume that the supremum on $g \in \mathcal{G}$ is reached for $g^* = 1$. (optional: prove why)

   (b) Make a change of variables as in the lecture to express the problem as the maximization of an inner product on normalized functions.

   (c) Derive the solution in terms of the original parametrization

6. Assume the variance of the samples y_i to be equal to one and replace $\left(\langle g, \widehat{S}_{YY} g\rangle_{\mathcal{G}} + \lambda \|g\|^2_{\mathcal{G}}\right)^{1/2}$ by $\left(\langle g, \widehat{S}_{YY} g\rangle_{\mathcal{G}}\right)^{1/2}$ in the expression of kernel CCA (i.e. no regularization when dividing by the empirical variance of y). What do you observe?

References

[1] T. Hofmann, B. Schölkopf, and A. J. Smola. Kernel methods in machine learning. *The Annals of Statistics*, 36(3):1171–1220, 2008.