CS 182: Introduction to Machine Learning, Fall 2022
Homework 3

(Due on Wednesday, Nov. 16 at 11:59pm (CST))

Notice:

- Please submit your assignments via Gradescope. The entry code is G2V63D.
- Please make sure you select your answer to the corresponding question when submitting your assignments.
- Each person has a total of five days to be late without penalty for all the assignments. Each late delivery less than one day will be counted as one day.

1. [20 points] [*SVM*]

   (a) In hard-margin SVM, the problem of maximizing margin can be converted into the following equivalent problem

$$\begin{aligned} \underset{\mathbf{w}, w_0}{\text{minimize}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 \\ \text{subject to} \quad & r^{(i)}(\mathbf{w}^\mathsf{T}\mathbf{x}^{(i)} + w_0) \geq 1, \ i = 1, \ldots, N. \end{aligned}$$

   (i) By introducing Lagrange multipliers $\{\alpha_i\}$, please give the Lagrangian function and the dual representation of the problem above. [6 points]

   (ii) Please show that the maximum of the margin $\gamma = \frac{1}{\|\mathbf{w}\|}$ is given by

$$\frac{1}{\gamma_{\max}^2} = \sum_{i=1}^{N} \alpha_i.$$

   (Hint: $\{\alpha_i\}$ can be obtained by solving the dual representation of the maximum margin problem.) [7 points]

   (b) The dual problem of soft-margin SVM is

$$\begin{aligned} \underset{\{\alpha_i\}}{\text{maximize}} \quad & \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j r^{(i)} r^{(j)} \mathbf{x}^{(i)\mathsf{T}} \mathbf{x}^{(j)} \\ \text{subject to} \quad & \sum_{i=1}^{N} \alpha_i r^{(i)} = 0, \\ & 0 \leq \alpha_i \leq C, \ i = 1, \ldots, N, \end{aligned} \tag{1}$$

   where $\{\alpha_i\}$ are the dual variables and $(\mathbf{x}^{(i)}, r^{(i)}) \in \mathbb{R}^d \times \{-1, 1\}$ are feature-label pairs. Let $f(\mathbf{x}) = \mathbf{w}^\mathsf{T}\mathbf{x} + w_0$ be the prediction function, where $\mathbf{w} \in \mathbb{R}^d$, $b \in \mathbb{R}$ are primal variables. Argue from KKT conditions why the following hold:

$$\begin{aligned} \alpha_i = 0 &\Longrightarrow r^{(i)} f(\mathbf{x}^{(i)}) \geq 1, \\ 0 < \alpha_i < C &\Longrightarrow r^{(i)} f(\mathbf{x}^{(i)}) = 1, \\ \alpha_i = C &\Longrightarrow r^{(i)} f(\mathbf{x}^{(i)}) \leq 1. \end{aligned}$$

   [7 points]

2. [20 points] [SVM] Given a dataset $\left\{(\mathbf{x}^{(i)}, r^{(i)})\right\}_{i=1}^{N}$. We wish to find a linear function f such that it can predict an input approximately correct, i.e, $f(\mathbf{x}) = \mathbf{w}^\mathsf{T}\mathbf{x} \approx r$ (already aborbed the bias, and $\mathbf{x} \in \mathbb{R}^d$). As in slides, we use the ϵ-insentive loss function $p_\epsilon(u) = \max(0, |u| - \epsilon)$. Accordingly, the SVR cost function is

$$J(\mathbf{w}) = \frac{1}{2}\left\|\mathbf{w}\right\|^2 + C\sum_{i=1}^{N} p_\epsilon(r^{(i)} - \mathbf{w}^\mathsf{T}\mathbf{x}^{(i)}).$$

Here the term $\frac{1}{2}\left\|\mathbf{w}\right\|^2$ acts as a regularizer.

(a) By following essentially the same procedure as for SVM, write down the dual problem as a quadratic programming (QP). [13 points]

(b) Develop a kernelized version of the Dual Problem. Also specify how you may obtain the prediction for a new point $\mathbf{x}^{(t)}$. [5 points]

(b) How do you define “Support Vectors” for this problem? [2 points]

3. [20 points] [*Dimension Reduction*]

(a) Principal components analysis (PCA) reduces the dimensionality of the data by finding projection direction(s) that minimizes the squared errors in reconstructing the original data or equivalently maximizes the variance of the projected data. On the other hand, Fisher's linear discriminant is a supervised dimension reduction method, which, given labels of the data, finds the projection direction that maximizes the between-class variance relative to the within-class variance of the projected data. In the following Figure 1, draw the first principal component direction in the left figure, and the first Fisher's linear discriminant direction in the right figure (Note: for PCA, ignore the fact that points are labeled (as round, diamond or square) since PCA does not use label information. For linear discriminant, consider round points as the positive class, and both diamond and square points as the negative class). [7 points]

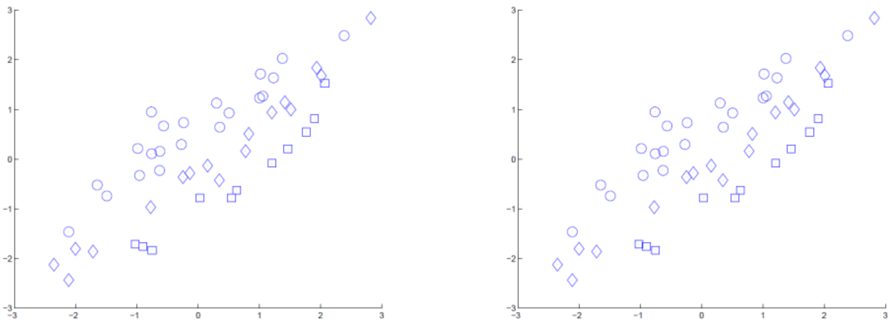

Figure 1: Draw the first principal component and linear discriminant component, respectively

(b) Canonical correlation analysis (CCA) handles the situation that each data point (i.e., each object) has two representations (i.e., two sets of features), e.g., a web page can be represented by the text on that page, and can also be represented by other pages linked to that page. Now suppose each data point has two representations $\mathbf{x}$ and $\mathbf{y}$, each of which is a 2-dimensional feature vector (i.e., $\mathbf{x} = [x_1, x_2]^T$ and $\mathbf{y} = [y_1, y_2]^T$). Given a set of data points, CCA finds a pair of projection directions $(\mathbf{u}, \mathbf{v})$ to maximize the sample correlation $\hat{\text{corr}}\left(\mathbf{u}^T\mathbf{x}\right)\left(\mathbf{v}^T\mathbf{y}\right)$ along the directions $\mathbf{u}$ and $\mathbf{v}$. In other words, after we project one representation of data points onto$\mathbf{u}$ and the other representation of data points onto $\mathbf{v}$, the two projected representations $\mathbf{u}^T\mathbf{x}$ and $\mathbf{v}^T\mathbf{y}$ should be maximally correlated (intuitively, data points with large values in one projected direction should also have large values in the other projected direction).

Now we can see data points shown in Figure 2, where each data point has two representations $\mathbf{x} = [x_1, x_2]^T$ and $\mathbf{y} = [y_1, y_2]^T$. Note that data are paired: each point in the left figure corresponds to a specific point in the right figure and vice versa, because these two points are two representations of the same object. Different objects are shown in different gray scales in the two figures (so you should be able to approximately figure out how points are paired). In the right figure we've given one CCA projection direction $\mathbf{v}$, draw the other CCA projection direction $\mathbf{u}$ in the left figure.[5 points]

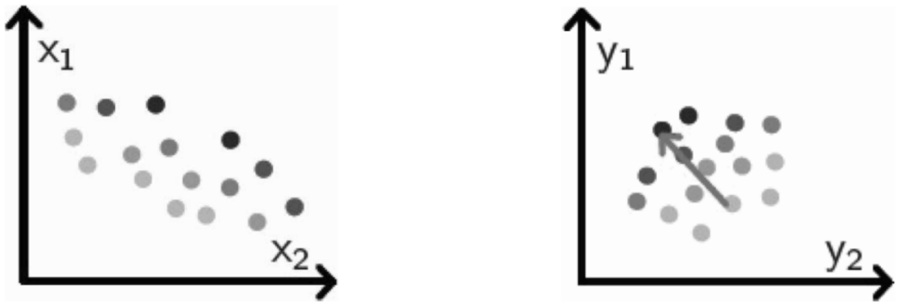

Figure 2: Draw the CCA projection direction in the left figure

(c) Consider 3 data points in the 2-d space:$(-1,-1), (0,0), (1,1)$. What is the first principal component (write down the actual vector)? Besides, If we project the original data points into the 1-d subspace by the principal component you choose, what are their coordinates in the 1-d subspace? And what is the variance of the projected data? [8 points]

4. [20 points] [*Dimension Reduction*]

(a) Prove that $\mathbf{w}_1$ is the principal component in principal component analysis (PCA) in that the projection onto direction $\mathbf{w}_1$ leads to the maximum variance for $\mathbf{X}$, i.e., $\mathbf{w}_1$ maximizes $\mathbf{w}^T\mathbf{\Sigma}\mathbf{w}$ for any $\mathbf{w}$ satisfying $\|\mathbf{w}\|_2 = 1$. [10 points]

(b) Suppose the sample $\mathbf{X}$ is labeled into two classes, briefly describe the idea of linear discriminant analysis (LDA) and discuss the similarities and differences between PCA and LDA. [5 points]

(c) Show that $\mathbf{w}_1$ minimizes the reconstruction error $\left\|\mathbf{X} - \mathbf{X}\mathbf{w}\mathbf{w}^T\right\|_F^2$ for any $\mathbf{w}$ satisfying $\|\mathbf{w}\|_2 = 1$. [5 points]

5. [20 points] [*Coding: MLP*] Complete “HW3-Coding.ipynb”. After completion, you should convert your notebook to PDF, and concatenate the writing part and the coding part into one PDF which is the file to submit. Download the dataset from `http://pan.shanghaitech.edu.cn/cloudservice/outerLink/decode?c3Vnb24xNjY3NzEzMzQ3ODg0c3Vnb24=`