ECE M146 Homework 4
Introduction to Machine Learning Monday, April 22, 2019
Instructor: Lara Dolecek Due: Wednesday, May 1, 2019
TA: Zehui (Alex) Chen, Ruiyi (John) Wu

1. In this problem, we will derive the least square solution for multi-class classification. Consider a general classification problem with K classes, with a 1-of-K binary encoding scheme (defined latter) for the target vector $t, t \in \mathbb{R}^K$. Suppose we are given a training data set $\{x_n, t_n\}, n = 1, \cdots, n$ where $x_n \in \mathbb{R}^D$. For the 1-of-K binary encoding scheme, t_n has the k-th element being 1 and all other elements being 0 if the n-th data is in class k. We can use the following linear model to describe each class:

$$y_k(x) = w_k^T x + w_{k0},$$

where $k = 1, \cdots, K$. We can conveniently group these together using vector notation so that

$$y(x) = \tilde{W}^T \tilde{x},$$

where $\tilde{W}$ is a matrix whose k-th column comprises the $D+1$-dimensional vector $\tilde{w} = [w_{k0}, w_k^T]^T$ and $\tilde{x}$ is the corresponding augmented input vector $[1, x^T]^T$. For each new input with feature x, we assign it to the class for which the output $y_k = \tilde{w}_k^T \tilde{x}$ is largest. Define a matrix T whose n-th row is the vector t_n^T and together a matrix $\tilde{X}$ whose n-th row is $\tilde{x}_n^T$, the sum-of-squares error function can be written as

$$J(\tilde{W}) = \frac{1}{2} Tr \left\{ (\tilde{X}\tilde{W} - T)^T (\tilde{X}\tilde{W} - T) \right\}.$$

(a) Find the closed form solution of $\tilde{W}$ that minimizes the objective function $J(\tilde{W})$. Hint: You many use the following two matrix derivative about trace, $\frac{\partial}{\partial Z} Tr(AZ) = A^T$ and $\frac{\partial}{\partial Z} Tr(Z^T A Z) = (A^T + A)Z$.

(b) Show that $J(\tilde{W})$ has a unique minimum. Hint: show that the double derivative of $J(\tilde{W})$ with respect to $\tilde{W}$ is positive semi-definite.

2. Show that a kernel function $K(x, x')$ satisfies the following generalization of the Cauchy-Schwartz inequality:

$$K(x_1, x_2)^2 \leq K(x_1, x_1)K(x_2, x_2).$$

Hint: The Cauchy-Schwartz inequality states that: for two vectors u and v, $|u^T v|^2 \leq \|u\|^2\|v\|^2$.

3. Given valid kernels $K_1(x, x')$ and $K_2(x, x')$, show that the following kernels are also valid:

   (a) $K(x, x') = K_1(x, x') + K_2(x, x')$.

   (b) $K(x, x') = K_1(x, x')K_2(x, x')$.

   (c) $K(x, x') = \exp(K_1(x, x'))$. Hint: use your results in (a) and (b).

4. In class, we learned that the soft margin SVM have the primal problem:

$$\begin{aligned}
\min_{\xi,w,b} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i \\
s.t. \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i, \quad i = 1, \cdots, m \\
& \xi_i \geq 0, \quad i = 1, \cdots, m
\end{aligned}$$

and the dual problem:

$$\begin{aligned}
\max_{\alpha} \quad & W(\alpha) = \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{m} y^{(i)}y^{(j)}a_i a_j \langle x^{(i)}, x^{(j)}\rangle \\
s.t. \quad & 0 \leq \alpha_i \leq C, i = 1, \cdots, m \\
& \sum_{i=1}^{m}\alpha_i y^{(i)} = 0.
\end{aligned}$$

Now suppose we have solved the dual problem and have the optimal α. Show that the parameter b can be determined using the following equation:

$$b = \frac{1}{N_{\mathcal{M}}}\sum_{n\in\mathcal{M}}\left(y^{(n)} - \sum_{m\in\mathcal{S}}\alpha_m y^{(m)}\langle x^{(n)}, x^{(m)}\rangle\right). \tag{1}$$

In (1), $\mathcal{M}$ denotes the set of indexes of data points having $0 < \alpha_n < C$ and $\mathcal{S}$ denotes the set of indexes of data points having $\alpha_n \neq 0$.

5. Suppose you are given 6 one-dimensional points: 3 with negative labels $x_1 = -1, x_2 = 0, x_3 = 1$ and 3 with positive labels $x_4 = -3, x_5 = -2, x_6 = 3$. In this question, we first compare the performance of linear classifier with or without kernel. Then we solve for the maximum margin classifier using SVM.

   (a) Consider a linear classifier of form $f(x) = \text{sign}(w_1 x + w_0)$. Write down the optimal value of w and its classification accuracy on the above 6 points. There might be more than one optimal solution, writing down one of them is enough.

   (b) Given two samples x and z in $\mathbb{R}$, define the kernel $K : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ as

   $$K(x, z) = xz(1 + xz).$$

   Find the corresponding feature map $\phi(x)$.

   (c) Apply $\phi(x)$ to the data and plot the points in the induced feature space $\mathbb{R}^2$. Are these points linearly separable now?

   (d) Look at the points in the induced feature space. Construct a maximum margin separating hyperplane in $\mathbb{R}^2$ which can be parameterized by $w_1\phi_1(x) + w_2\phi_2(x) + w_0 = 0$. Draw this hyperplane on your plot and circle the support vectors.

   (e) Draw the decision boundary of the separating hyperplane you found in (d) in the original $\mathbb{R}$ feature space.

   (f) Find the α_i, w and b in

   $$h(x) = \text{sign}\left(\sum_{n \in \mathcal{S}} \alpha_n y_n K(x_n, x) + b\right) = \text{sign}\left(w^T \phi(x) + b\right).$$

   Do this by solving the dual form of the quadratic program. How is w and b related to your solution in part (d)?

6. In this exercise, we will use MATLAB to solve both the primal and the dual problem of SVM. In *Data.csv*, the first two columns contain feature vectors $x^{(i)} \in \mathbb{R}^2$ and the last column contains the label $y^{(i)} \in \{-1, 1\}$. We will use CVX as the optimization solver in this problem. For help with CVX, refer to the CVX Users' Guide. Attach your code for submission. For Python user, feel free to use the following libraries: math, csv, numpy, matplotlib and cvxpy.

   (a) **Visulization** Use different color to plot data with different labels in the 2-D feature space. Is the data linearly separable?

   (b) **The Primal Problem** Use CVX to solve the primal problem of this form:

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2}\|w\|^2 \\ s.t. \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1, \quad i = 1, \cdots, m \end{aligned}$$

   Report w and b. Plot the hyperplane defined by w and b.

   (c) **The Dual Problem** Use CVX to solve the dual problem of this form:

$$\begin{aligned} \max_{\alpha} \quad & W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} a_i a_j \langle x^{(i)}, x^{(j)} \rangle \\ s.t. \quad & 0 \leq \alpha_i, i = 1, \cdots, m \\ & \sum_{i=1}^{m} \alpha_i y^{(i)} = 0. \end{aligned}$$

   Use the resulting a to identify the support vectors on the plot. Report you non-zero $a_i's$. How many support vectors do you have? Circle those support vectors.

   Note: The latter part of $W(a)$ is in quadratic form, i.e., $a^T P a$. To use CVX, first find P and then use *quad_form(a,P)*. For Python user, you will need to add a small number to the diagonal of P matrix to make cvxpy work. i.e. Run the following code before using cvxpy: "P += 1e-13 * numpy.eye(29)", where 29 is the total number of data. Also, assume it is 0 if a number is less than 1e-9.