ECE M146 | Homework 4
Introduction to Machine Learning | Monday, April 26, 2021
Instructor: Lara Dolecek | Due: Monday, May 3, 2021
TA: Zehui (Alex) Chen | chen1046@ucla.edu

Please upload your homework to Gradescope by May 3, 4:00 pm.
Please submit a single PDF directly on Gradescope
You may type your homework or scan your handwritten version. Make sure all the work is discernible.

1. In this section, you will use the k-NN classifier to predict whether or not a person is admitted into the graduate program at UCLA. We will use both *UCLA_EE_grad_2030.csv* and *UCLA_EE_grad_2031.csv*. Each dataset has 100 samples. Use the first 40 samples in *UCLA_EE_grad_2030.csv* as the testing dataset and the remaining 160 samples as the training dataset.

   We are going to build the k-NN classifier from first principles. You may **not** use **fitcknn** (**sklearn.neighbors.KNeighborsClassifier** for python) in this problem as you will get an incorrect answer by using those built-in functions.

   The k-NN classifier classifies a data point with feature x_{test} based on a training set by performing the following procedures:

   - Compute the distance from x_{test} to the features of all training points. We will use the Euclidean distance in this problem.
   - Find the k nearest neighbors of this point.
   - Classify this points based on the majority class of its k nearest neighbors.

   We use the following two rules to handle ties:

   (a) Let d_k be the distance of the k-th nearest neighbor of x_{test}. If there are multiple training points that have the same distance d_k from x_{test}, choose those points with the smallest index to be included in the k nearest neighbors.

   (b) For even k, among all k nearest neighbors of a data point, if the number of points from class 1 is the same as the number of points from class 0, classify this data point as y_{tie} deterministically.

   Answer the following questions:

   (a) With $y_{tie} = 1$, implement the k-NN algorithm. Find and plot the training and testing accuracy for $k = 1, 2, \cdots, 12$.

   (b) With $y_{tie} = 0$, implement the k-NN algorithm. Find and plot the training and testing accuracy for $k = 1, 2, \cdots, 12$.

   (c) Comment on the performance of the k-NN classifiers in (a) and (b). How does larger k affect the training and testing error? How does the parity of k affect the performance of the k-NN classifier in (a) and (b), respectively? Are they contradictory to each other? Explain why.

2. You are given the following dataset which consists of $x^{(i)} \in \mathbb{R}^2$ and $y^{(i)} \in \{-1, 1\}$:

i	$x_1^{(i)}$	$x_2^{(i)}$	$y^{(i)}$
1	-3	9	1
2	-2.5	6.25	1
3	3	9	1
4	-1.5	2.25	-1
5	0	0	-1
6	1	1	-1

   (a) Plot the data. Is the data linearly separable?

   (b) Find and circle the support vectors by inspection. Find and plot the maximum margin separating hyperplane using basic geometry. Hint: there are only two support vectors.

   (c) Find the α_i, w and b in

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

   where $\mathcal{S}$ is the index set of all support vectors. Do this by solving the dual problem as a quadratic problem. How are w and b related to your solution in part (b)?

3. In this exercise, we will use MATLAB (or python) to solve both the primal and the dual problem of SVM using the dataset *UCLA_EE_grad_2031.csv*. In *UCLA_EE_grad_2031.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 (cvxpy) as the optimization solver in this problem. For help with CVX (cvxpy), refer to the CVX Users' Guide or cvxpy Users' Guide. Attach your code for submission.

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

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

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

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 for numerical stability. i.e., run the following code before using cvxpy: "P += 1e-2 * numpy.eye(100)", where 100 is the total number of data points. Also, assume a number is effectively 0 if it is very small.

4. Show that a kernel function $K(x_1, x_2)$ 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, $|u^T v|^2 \leq \|u\|^2 \|v\|^2$.

5. 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).

6. In class, we learned that the soft margin SVM has 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:

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

Note that $\langle z, s \rangle$ is an alternative expression for the inner product $z^T s$. As usual, $y^{(i)} \in \{+1, -1\}$.

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 indices of data points having $0 < \alpha_n < C$, parameter $N_{\mathcal{M}}$ denotes the size of the set $\mathcal{M}$, and $\mathcal{S}$ denotes the set of indices of data points having $\alpha_n \neq 0$.