ECE M146 — Homework 4 Solution

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.

Solution:

(a) See figure below.

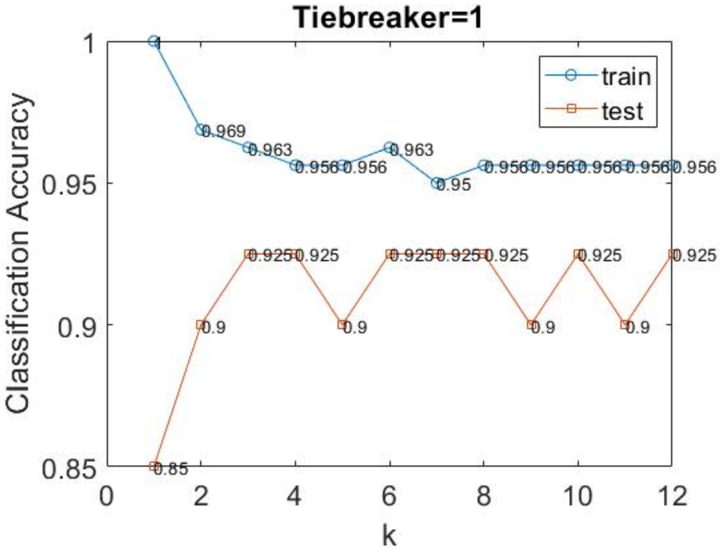


Figure 1: $y_{tie} = 1$

(b) See figure below.

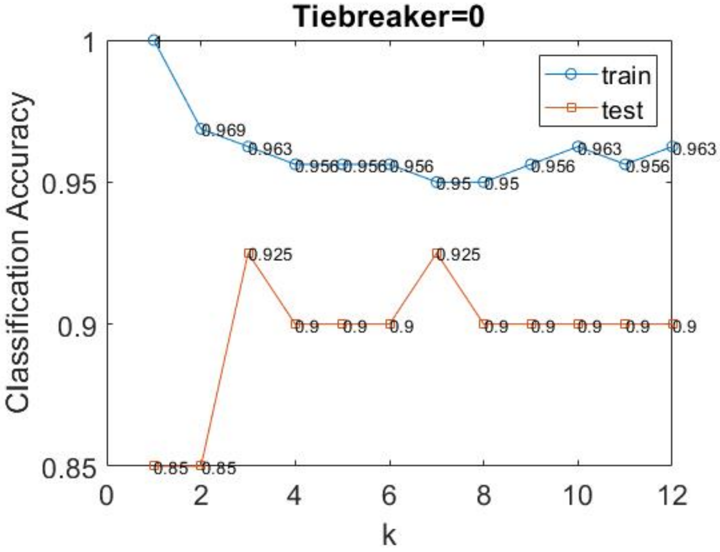


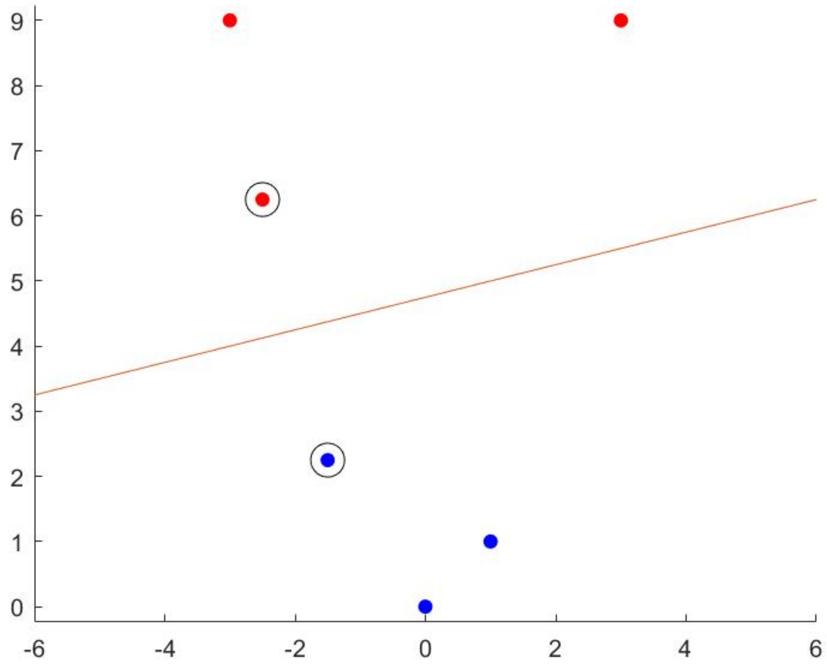
Figure 2: $y_{tie} = 0$

(c) With larger k, the training accuracy gets smaller and the testing accuracy does not change much. This shows that we are over-fitting with small k. In (a), testing accuracy increases with the increase in even k. For (b), the testing accuracy decreases with even k. They are not contradictory to each other. This behavior is a result of tie breaking rule (b) as we deterministically choose the label when there is a tie for even k. Predicting 1 in case of a tie gives us a better result than predicting 0 in case of a tie is more robust to over-fitting.

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?
Solution: Yes.



(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.

Solution: The two points that are closest are $(-1.5, 2.25)$ with negative label and $(-2.5, 6.25)$ with positive label. They are the support vectors and the maximum margin separating hyperplane is given by $-x_1 + 4x_2 - 19 = 0$ by finding a line that has the normal vector $[-1, 4]^T$ and also passes through the mid-point of the support vectors, i.e., $(-2, 4.25)$. The line is drawn on the above figure.

(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)?

Solution: Since we only have two support vectors, only the Lagrange multiplier corresponding to the support vectors are non-zero. Let α_2 denote the Lagrange multiplier for the data $x^{(2)}$ and similarly α_4 for the data $x^{(4)}$. From the condition $\sum_{i=1}^{6} \alpha_i y_i = 0$, we get $\alpha_2 = \alpha_4 = \alpha_0$. Write down the objective of the dual problem of SVM:

$$\begin{aligned} W(\boldsymbol{\alpha}) &= \sum_{i=1}^{6} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{6} y_i y_j \alpha_i \alpha_j x^{(i)T} x^{(j)} \\ &= 2\alpha_0 - \frac{1}{2}\alpha_0^2 x^{(2)T} x^{(2)} + \alpha_0^2 x^{(2)T} x^{(4)} - \frac{1}{2}\alpha_0^2 x^{(4)T} x^{(4)} \\ &= 2\alpha_0 - 8.5\alpha_0^2. \end{aligned}$$

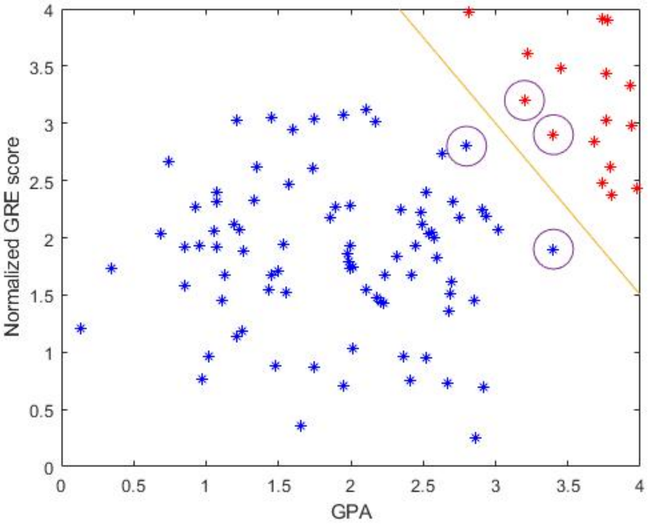
Maximizing $W(\alpha)$ over α_0, we get $\alpha_2 = \alpha_4 = \alpha_0 = \frac{2}{17}$. Using $w = \sum_{m \in \mathcal{S}} \alpha_m y_m x_m$, we get $w = [-\frac{2}{17}, \frac{8}{17}]^T$. To find b, recall that

$$y_i \left(w^T x^{(i)} + b \right) = 1$$

for any support vectors x_i. Use any support vector, we can get $b = -\frac{38}{17}$. The w and b we find by solving the dual problem is a scaled version of $[w_1, w_2]^T$ and w_0 in part (d). These solutions therefore give the same separating hyperplane.

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?
   Solution: Yes, the data is 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.
   Solution: We get $w = [3, 2]^T$ and $b = -15$. The hyperplane $w^T x + b = 0$,i.e., $3x_{GPA} + 2x_{GRE} - 15 = 0$, is shown in the figure above.

   (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., $a^T Pa$. 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.

Solution: There are 4 support vectors.

The corresponding $a = [5.2667, 5.4111, 1.0889, 1.2332]^T$ in MATLAB and $a = [3.9169, 5.5214, 0.7079, 2.3124]^T$ in python. The support vectors are circled in the above figure. In order to use CVX (cvxpy) to solve this problem, first find a matrix P where $P_{ij} = y^{(i)}y^{(j)}\langle x^{(i)}, x^{(j)}\rangle$. Then let CVX (cvxpy) maximizes $\sum_{i=1}^{m} a_i - \frac{1}{2}a^T Pa$.

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, $|u^T v|^2 \leq \|u\|^2\|v\|^2$.
Solution 1: From the definition of kernel, we have

$$\begin{aligned} K(x_1, x_2)^2 &= (\phi(x_1)^T \phi(x_2))^2 \\ &\leq (\phi(x_1)^T \phi(x_1))(\phi(x_2)^T \phi(x_2)) \\ &= K(x_1, x_1)K(x_2, x_2). \end{aligned}$$

The inequality comes from the Cauchy-Schwartz inequality.

Solution 2: For an alternative solution, we consider the 2×2 Gram matrix

$$\boldsymbol{K} = \begin{bmatrix} K(x_1, x_1) & K(x_1, x_2) \\ K(x_2, x_1) & K(x_2, x_2) \end{bmatrix}$$

Since $K(x_1, x_2)$ is a valid kernel, $\boldsymbol{K}$ is positive definite with $|\boldsymbol{K}| \geq 0$. This shows that $K(x_1, x_2)^2 \leq K(x_1, x_1)K(x_2, x_2)$.

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

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

   Solution: Suppose $K_1(x,x')$ has positive semi-definite Kernel matrix $\boldsymbol{K}_1$ and $K_2(x,x')$ has positive semi-definite Kernel matrix $\boldsymbol{K}_2$ with the same dimension. Then it is easy to show that $K(x,x')$ has Kernel matrix $\boldsymbol{K} = \boldsymbol{K}_1 + \boldsymbol{K}_2$ which is also positive semi-definite. In other words, if $z^T\boldsymbol{K}_1 z \geq 0, \forall z$ and $z^T\boldsymbol{K}_2 z \geq 0, \forall z$, then $z^T\boldsymbol{K}z \geq 0, \forall z$.

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

   Solution: We assume the mapping function for $K_1(x,x')$ is $\phi^{(1)}(x)$ and similarly $\phi^{(2)}(x)$ for $K_2(x,x')$. Moreover, we further assume the dimension of $\phi^{(1)}(x)$ is M and the dimension of $\phi^{(2)}(x)$ is N. We can then expand $K(x,x')$.

$$\begin{aligned}
K(x,x') &= K_1(x,x')K_2(x,x') \\
&= \phi^{(1)}(x)^T\phi^{(1)}(x')\phi^{(2)}(x)^T\phi^{(2)}(x') \\
&= \sum_{i=1}^{M}\phi_i^{(1)}(x)\phi_i^{(1)}(x')\sum_{j=1}^{N}\phi_j^{(2)}(x)\phi_j^{(2)}(x') \\
&= \sum_{i=1}^{M}\sum_{j=1}^{N}\left[\phi_i^{(1)}(x)\phi_j^{(2)}(x)\right]\left[\phi_i^{(1)}(x')\phi_j^{(2)}(x')\right] \\
&= \sum_{k=1}^{MN}\phi_k(x)\phi_k(x') = \phi(x)^T\phi(x').
\end{aligned}$$

   In the above equation, $\phi(x)$ is a $MN \times 1$ column vector with the k-th element given by $\phi_i^{(1)}(x) \times \phi_j^{(2)}(x)$. For a given k, the corresponding i and j are calculated as follows: $i = \lfloor (k-1)/N \rfloor + 1$, and $j = (k-1) \mod N + 1$.

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

   Solution: Consider the Taylor series expansion for the exponential function:

$$K(x,x') = \sum_{n=0}^{\infty}\frac{K_1(x,x')^n}{n!}.$$

   Using results from (a) and (b) repeatedly across terms; each term respectively shows that $K(x,x')$ is a valid kernel.

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 $\alpha_n \neq 0$.

Solution: From the KKT condition on complementary slackness, we find that for each data points with $0 < \alpha_n < C$, i.e., $n \in \mathcal{M}$, we have

$$y^{(n)}(w^T x^{(n)} + b) = 1.$$

Multiplying by $y^{(n)}$ on both sides and then summing over $\mathcal{M}$ (note that the square of $y^{(n)}$ is always 1), we have:

$$b = \frac{1}{N_{\mathcal{M}}}\sum_{n\in\mathcal{M}}\left(y^{(n)} - w^T x^{(n)}\right).$$

Rewrite w in terms of α by using $w = \sum_{m\in\mathcal{S}}\alpha_m y^{(m)}x^{(m)}$. We find

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