ECE M146 Homework 5 Solution
Introduction to Machine Learning
Instructor: Lara Dolecek
TA: Zehui (Alex) Chen, Ruiyi (John) Wu

Please upload your homework to Gradescope by May 14, 11:59 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. 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 $K(x_1, x_2)^2 \leq K(x_1, x_1)K(x_2, x_2)$.

2. 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 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 another word, 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 on across terms and with each term respectively shows that $K(x, x')$ is a valid kernel.

3. 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 (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

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

4. Consider 3 random variables A,B and C with joint probabilities $P(A, B, C)$ listed in the following table.

	C=0		C=1	
	B=0	B=1	B=0	B=1
A=0	0.096	0.024	0.27	0.03
A=1	0.224	0.056	0.27	0.03

(a) Calculate $P(A|C = 0)$, $P(B|C = 0)$, and $P(A, B|C = 0)$.
Solution:

$$P(A|C=0) = \begin{cases} 0.3, A = 0 \\ 0.7, A = 1 \end{cases} \qquad P(B|C=0) = \begin{cases} 0.8, B = 0 \\ 0.2, B = 1 \end{cases}$$

$$P(A, B|C=0) = \begin{cases} 0.24, A = 0, B = 0 \\ 0.06, A = 0, B = 1 \\ 0.56, A = 1, B = 0 \\ 0.14, A = 1, B = 1 \end{cases}$$

(b) Calculate $P(A|C = 1)$, $P(B|C = 1)$, and $P(A, B|C = 1)$.
Solution:

$$P(A|C=1) = \begin{cases} 0.5, A = 0 \\ 0.5, A = 1 \end{cases} \qquad P(B|C=1) = \begin{cases} 0.9, B = 0 \\ 0.1, B = 1 \end{cases}$$

$$P(A, B|C=1) = \begin{cases} 0.45, A = 0, B = 0 \\ 0.05, A = 0, B = 1 \\ 0.45, A = 1, B = 0 \\ 0.05, A = 1, B = 1 \end{cases}$$

(c) Is A conditionally independent of B given C?
Solution: Yes. From the above, we can verify $P(A|C = 1)P(B|C = 1) = P(A, B|C = 1)$ and $P(A|C = 0)P(B|C = 0) = P(A, B|C = 0)$.

(d) Calculate $P(A)$, $P(B)$, and $P(A, B)$.
Solution:

$$P(A) = \begin{cases} 0.42, A = 0 \\ 0.58, A = 1 \end{cases} \qquad P(B) = \begin{cases} 0.86, B = 0 \\ 0.14, B = 1 \end{cases}$$

$$P(A, B) = \begin{cases} 0.366, A = 0, B = 0 \\ 0.034, A = 0, B = 1 \\ 0.494, A = 1, B = 0 \\ 0.086, A = 1, B = 1 \end{cases}$$

(e) Is A independent of B?
Solution: No. It is easy to verify that $P(A)P(B) \neq P(A, B)$.

5. Let us revisit the restaurant selection problem in HW3. You are trying to choose between two restaurants (sample 9 and sample 10) to eat at. To do this, you will train a classifier based on your past experiences (sample 1-8). The features for each restaurants and your judgment on the goodness of sample 1-8 are summarized by the following chart. In this exercise, instead of a decision tree, you will use the Naïve

Sample #	HasOutdoorSeating	HasBar	IsClean	HasGoodAtmosphere	IsGoodRestaurant
1	0	0	0	1	1
2	1	1	0	0	0
3	0	1	1	1	1
4	1	0	0	1	1
5	1	1	1	0	0
6	1	0	1	0	1
7	1	1	0	1	1
8	0	0	1	1	1
9	0	1	0	1	?
10	1	1	1	1	?

Bayes classifier to decide whether restaurant 9 and 10 are good or not. For clarity, we abbreviate the names of the features and label as follows: HasOutdoorSeating $\to O$, HasBar $\to B$, IsClean $\to C$, HasGoodAtmosphere $\to A$, and IsGoodRestaurant $\to G$.

(a) Train the Naïve Bayes classifier by calculating the maximum likelihood estimate of class priors and class conditional distributions. Namely, calculate the maximum likelihood estimate of the following: $P(G)$, and $P(X|G), X \in \{O, B, C, A\}$.
Solution: The maximum likelihood of class priors are just the relative frequency of each class. We therefore have:

$$P(G=0) = \frac{2}{8} = \frac{1}{4}, P(G=1) = \frac{6}{8} = \frac{3}{4}.$$

The class conditional distribution can be estimated similarly by calculating the relative frequency of the features conditional on the class. We get:

$$\begin{aligned}
&P(O=0|G=0) = 0, P(O=0|G=1) = \frac{3}{6} = \frac{1}{2};\\
&P(B=0|G=0) = 0, P(B=0|G=1) = \frac{4}{6} = \frac{2}{3};\\
&P(C=0|G=0) = \frac{1}{2}, P(C=0|G=1) = \frac{3}{6} = \frac{1}{2};\\
&P(A=0|G=0) = 1, P(A=0|G=1) = \frac{1}{6}.
\end{aligned}$$

(b) For Sample #9 and #10, make the decision using

$$\hat{G}_i = \underset{G_i \in \{0,1\}}{\operatorname{argmax}} \quad P(G_i)P(O_i, B_i, C_i, A_i|G_i),$$

where O_i, B_i, C_i, and A_i are the feature values for the i-th sample.
Solution: Using previous results, for $i = 9$:

$$P(G_i = 0)P(O_i, B_i, C_i, A_i|G_i = 0) = \frac{1}{4} \times 0 \times 1 \times \frac{1}{2} \times 0 = 0,$$

and

$$P(G_i = 1)P(O_i, B_i, C_i, A_i|G_i = 1) = \frac{3}{4}\times\frac{1}{2}\times\frac{1}{3}\times\frac{1}{2}\times\frac{5}{6} > P(G_i = 0)P(O_i, B_i, C_i, A_i|G_i = 0).$$

We then decide $\hat{G}_9 = 1$.
For $i = 10$:

$$P(G_i = 0)P(O_i, B_i, C_i, A_i|G_i = 0) = \frac{1}{4} \times 1 \times 1 \times \frac{1}{2} \times 0 = 0,$$

and

$$P(G_i = 1)P(O_i, B_i, C_i, A_i|G_i = 1) = \frac{3}{4}\times\frac{1}{2}\times\frac{1}{3}\times\frac{1}{2}\times\frac{5}{6} > P(G_i = 0)P(O_i, B_i, C_i, A_i|G_i = 0).$$

We then decide $\hat{G}_{10} = 1$.

(c) We use Laplace smoothing to avoid having class conditional probabilities that are strictly 0. To use Laplace smoothing for a binary classifier, add 1 to the numerator and add 2 to the denominator when calculating the class conditional distributions. Let us re-calculate the class conditional distributions with Laplace smoothing. Namely, calculate the maximum likelihood estimate of $P(X|G), X \in \{O, B, C, A\}$.
Solution: The class conditional distribution are:

$$\begin{aligned}
&P(O = 0|G = 0) = \frac{1}{4}, P(O = 0|G = 1) = \frac{4}{8} = \frac{1}{2};\\
&P(B = 0|G = 0) = \frac{1}{4}, P(B = 0|G = 1) = \frac{5}{8};\\
&P(C = 0|G = 0) = \frac{2}{4} = \frac{1}{2}, P(C = 0|G = 1) = \frac{4}{8} = \frac{1}{2};\\
&P(A = 0|G = 0) = \frac{3}{4}, P(A = 0|G = 1) = \frac{2}{8} = \frac{1}{4}.
\end{aligned}$$

(d) Repeat (b) with the class conditional distributions you get from (c).
Solution: Using previous results, for $i = 9$:

$$P(G_i = 0)P(O_i, B_i, C_i, A_i|G_i = 0) = \frac{1}{4} \times \frac{1}{4} \times \frac{3}{4} \times \frac{1}{2} \times \frac{1}{4} = 0.0059,$$

and

$$P(G_i = 1)P(O_i, B_i, C_i, A_i|G_i = 1) = \frac{3}{4} \times \frac{1}{2} \times \frac{3}{8} \times \frac{1}{2} \times \frac{3}{4} = 0.0527.$$

We then decide $\hat{G}_9 = 1$.
For $i = 10$:

$$P(G_i = 0)P(O_i, B_i, C_i, A_i|G_i = 0) = \frac{1}{4} \times \frac{3}{4} \times \frac{3}{4} \times \frac{1}{2} \times \frac{1}{4} = 0.0176,$$

and

$$P(G_i = 1)P(O_i, B_i, C_i, A_i|G_i = 1) = \frac{3}{4} \times \frac{1}{2} \times \frac{3}{8} \times \frac{1}{2} \times \frac{3}{4} = 0.0527$$

We then decide $\hat{G}_{10} = 1$.

6. In class, we learned a Naïve Bayes classifier for binary feature values, i.e., $x_j \in \{0, 1\}$ where we model the class conditional distribution to be Bernoulli. In this exercise, you are going to extend the result to the case where features that are non-binary.

   We are given a training set $\{(x^{(i)}, y^{(i)}); i = \{1, \cdots, m\}\}$, where $x^{(i)} \in \{1, 2, \cdots, s\}^n$ and $y^{(i)} \in \{0, 1\}$. Again, we model the label as a biased coin with $\theta_0 = P(y^{(i)} = 0)$ and $1 - \theta_0 = P(y^{(i)} = 1)$. We model each non-binary feature value $x_j^{(i)}$ (an element of $x^{(i)}$) as a biased dice for each class. This is parameterized by:

$$
\begin{aligned}
&P(x_j = k|y = 0) = \theta_{j,k|y=0},\ k = 1, \cdots, s-1;\\
&P(x_j = s|y = 0) = \theta_{j,s|y=0} = 1 - \sum_{k=1}^{s-1} \theta_{j,k|y=0};\\
&P(x_j = k|y = 1) = \theta_{j,k|y=1},\ k = 1, \cdots, s-1;\\
&P(x_j = s|y = 1) = \theta_{j,s|y=1} = 1 - \sum_{k=1}^{s-1} \theta_{j,k|y=1};
\end{aligned}
$$

   Notice that we do not model $P(x_j = s|y = 0)$ and $P(x_j = s|y = 1)$ directly. Instead we use the above equations to guarantee all probabilities for each class sum to 1.

   (a) Using the **Naïve Bayes (NB) assumption**, write down the joint probability of the data:

$$P(x^{(1)}, \cdots, x^{(m)}, y^{(1)}, \cdots, y^{(m)})$$

   in terms of the parameters θ_0, $\theta_{j,k|y=0}$ and $\theta_{j,k|y=1}$. You may find the indicator function $\mathbf{1}(\cdot)$ useful.
   Solution:

$$
\begin{aligned}
&P(x^{(i)}, \cdots, x^{(m)}, y^{(i)}, \cdots, y^{(m)})\\
=&\prod_{i=1}^{m} P(x^{(i)}, y^{(i)})\\
=&\prod_{i=1}^{m} \theta_0^{\mathbf{1}(y^{(i)}=0)} (1-\theta_0)^{\mathbf{1}(y^{(i)}=1)} \prod_{j'=1}^{n} \prod_{k'=1}^{s} \theta_{j',k'|y=0}^{\mathbf{1}(x_{j'}^{(i)}=k' \wedge y^{(i)}=0)} \theta_{j',k'|y=1}^{\mathbf{1}(x_{j'}^{(i)}=k' \wedge y^{(i)}=1)}.
\end{aligned}
\tag{2}
$$

   (b) Now, maximize the joint probability you get in (a) with respect to each of θ_0, $\theta_{j,k|y=0}$, and $\theta_{j,k|y=1}$. Write down your resulting θ_0, $\theta_{j,k|y=0}$ and $\theta_{j,k|y=1}$ and show intermediate steps. Explain in words the meaning of your results.
   Solution: Take the negative log of Equation (1) and we get:

$$
\begin{aligned}
J(\theta_0, \theta_{j,k|y=0}, \theta_{j,k|y=1}) = -\sum_{i=1}^{m} \Bigg\{ &\mathbf{1}(y^{(i)} = 0) \log(\theta_0) + \mathbf{1}(y^{(i)} = 1) \log(1 - \theta_0)\\
&+ \sum_{j'=1}^{n} \sum_{k'=1}^{s} \left[\mathbf{1}(x_{j'}^{(i)} = k' \wedge y^{(i)} = 0) \log(\theta_{j',k'|y=0}) + \mathbf{1}(x_{j'}^{(i)} = k' \wedge y^{(i)} = 1) \log(\theta_{j',k'|y=1}) \right] \Bigg\}.
\end{aligned}
$$

We first find θ_0 that minimize J.

$$\frac{\partial J}{\partial \theta_0} = -\frac{\sum_{i=1}^m \mathbf{1}(y^{(i)} = 0)}{\theta_0} + \frac{\sum_{i=1}^m \mathbf{1}(y^{(i)} = 1)}{1 - \theta_0}.$$

Setting the derivative to 0 we get

$$\theta_0 = \frac{\sum_{i=1}^m \mathbf{1}(y^{(i)} = 0)}{m}.$$

Next we find $\theta_{j,k|y=0}$ for a particular j and $k \neq s$. We first take the derivative with respect to $\theta_{j,k|y=0}$. Notice that in J, we also have $\theta_{j,s|y=0}$ that also depends on $\theta_{j,k|y=0}$.

$$\frac{\partial J}{\partial \theta_{j,k|y=0}} = -\frac{\sum_{i=1}^m \mathbf{1}(x_j^{(i)} = k \wedge y^{(i)} = 0)}{\theta_{j,k|y=0}} + \frac{\sum_{i=1}^m \mathbf{1}(x_j^{(i)} = s \wedge y^{(i)} = 0)}{\theta_{j,s|y=0}}.$$

Setting the derivative to 0 we get

$$\theta_{j,k|y=0} = \frac{\sum_{i=1}^m \mathbf{1}(x_j^{(i)} = k \wedge y^{(i)} = 0)}{\sum_{i=1}^m \mathbf{1}(x_j^{(i)} = s \wedge y^{(i)} = 0)} \theta_{j,s|y=0}.$$

Using the above equation for all $k \neq s$ and $\theta_{j,s|y=1} = 1 - \sum_{k=1}^{s-1} \theta_{j,k|y=1}$ we get:

$$\theta_{j,k|y=0} = \frac{\sum_{i=1}^m \mathbf{1}(x_j^{(i)} = k \wedge y^{(i)} = 0)}{\sum_{i=1}^m \mathbf{1}(y^{(i)} = 0)}.$$

Similarly, we have:

$$\theta_{j,k|y=1} = \frac{\sum_{i=1}^m \mathbf{1}(x_j^{(i)} = k \wedge y^{(i)} = 1)}{\sum_{i=1}^m \mathbf{1}(y^{(i)} = 1)}.$$

These result shows that the maximum likelihood estimate of the class conditional probability is the fraction of data in each class that belongs to class k.