ECE M146 Homework 5

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, $|u^T v|^2 \leq \|u\|^2\|v\|^2$.

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

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

   (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).

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 $\alpha_n \neq 0$.

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 $P(A, B|C = 0)$.

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

(c) Is A conditionally independent of B given C?

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

(e) Is A independent of 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 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 $\to G$.

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	?

(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 $P(X|G), X \in \{O, B, C, A\}$.

(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.

(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 $P(X|G), X \in \{O, B, C, A\}$.

(d) Repeat (b) with the class conditional distributions you get from (c).

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.

   (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.