CSCI 5521: Introduction to Machine Learning (Spring 2020)[1]

Homework 1

Questions

1. (**30 points**) Consider the class of K-interval classifier I_K in $\mathbb{R}$ which is specified by K intervals $[a_1, b_1]$, $[a_2, b_2]$,...,$[a_K, b_K]$ and labels any example positive iff it lies inside any of the K intervals.

   (a) What is the VC dimension of I_1 denoted by VC(I_1)? Prove your answer. You need to show why the classifiers can shatter VC(I_1) data points but not VC(I_1)+1 data points. **Answer:** $VC(I_1) = 2$. A set of two points can be shattered, since there is only a single block of positive examples that could lie within the interval. But no set of 3 points can be shattered, because it can not be labeled in alternating +,-,+ order.

   (b) What is the VC dimension of I_2 denoted by VC(I_2)? Prove your answer. You need to show why the classifiers can shatter VC(I_2) data points but not VC(I_2)+1 data points. **Answer:** Similarly, $VC(I_2) = 4$. A set of four points can be shattered, since there are at most two blocks of positive examples that could lie within the interval. But no set of 5 points can be shattered, because it can not be labeled in alternating +,-,+,-,+ order.

   (c) What is the VC dimension of I_K denoted by VC(I_K)? Prove your answer. You need to show why the classifiers can shatter VC(I_K) data points but not VC(I_K)+1 data points. **Answer:** Similarly, $VC(I_k) = 2k$. A set of $2k$ points can be shattered, since there are at most k blocks of positive examples that could lie within the interval. But no set of $2k + 1$ points can be shattered, because it can not be labeled in alternating +,-,+, ..., +,-,+ order.

---

2. (**30 points**) Find the Maximum Likelihood Estimation (MLE) for the following pdf. In each case, consider a random sample of size n. Show your calculation:

   (a) $f(x|\theta) = \frac{1}{\theta} e^{-\frac{x}{\theta}}$, $x > 0, \theta > 0$

   (b) $f(x|\theta) = 2\theta x^{2\theta-1}$, $0 < x \leq 1$, $0 < \theta < \infty$

[1]Instructor: Rui Kuang (kuang@cs.umn.edu). TAs: Tianci Song (song0309@umn.edu) and Ruyuan Wan (wanxx199@umn.edu).

(c) $f(x|\theta) = \frac{1}{2\theta}$, $0 \leq x \leq 2\theta$ (Hint: You can draw the likelihood function and pick a θ based on all the data points.)

Answer:

(a)

$$\mathcal{L}(\theta|\mathbf{x}) = \frac{1}{(\theta)^n}\exp(-\frac{1}{\theta}\sum_i x_i) \tag{1}$$

$$\Rightarrow \log \mathcal{L}(\theta|\mathbf{x}) = -n\log(\theta) - \frac{1}{\theta}\sum_i x_i \tag{2}$$

$$\Rightarrow \frac{\partial \log \mathcal{L}(\theta|\mathbf{x})}{\partial\theta} = -\frac{n}{\theta} + \frac{\sum_i x_i}{(\theta)^2} \tag{3}$$

Set (3) to zero we obtain

$$\hat{\theta} = \frac{\sum_i x_i}{n}$$

(b)

$$\mathcal{L}(\theta|\mathbf{x}) = (2\theta)^n(\prod_i x_i)^{2\theta-1} \tag{4}$$

$$\Rightarrow \log \mathcal{L}(\theta|\mathbf{x}) = n\log(2\theta) + (2\theta-1)(\sum_i \log x_i) \tag{5}$$

$$\Rightarrow \frac{\partial \log \mathcal{L}(\theta|\mathbf{x})}{\partial\theta} = \frac{n}{\theta} + 2\sum_i \log x_i \tag{6}$$

Set eq:2 to zero we obtain

$$\hat{\theta} = -\frac{n}{2\sum_i \log x_i}$$

(c) The likelihood function $\mathcal{L}(\theta|\mathbf{x}) = \frac{1}{2\theta^n}$ is monotonically decreasing in the domain $\theta > 0$, which means the smallest value of θ maximizes the likelihood function. Thus, $\hat{\theta}$ is the largest value of the set $\{\frac{x_i}{2}|i = 1, \ldots, n\}$ since $\theta \geq \frac{x}{2}$.

3. (**30 points**) Let $P(x|C)$ denote a Bernoulli density function for a class $C \in \{C_1, C_2\}$ and $P(C)$ denote the prior,

(a) Given the priors $P(C_1)$ and $P(C_2)$, and the Bernoulli densities specified by $p_1 \equiv p(x = 0|C_1)$ and $p_2 \equiv p(x = 0|C_2)$, derive the classification rules for classifying a sample x into C_1 and C_2 based on the posteriors $P(C_1|x)$ and $P(C_2|x)$. (Hint: give rules for classifying $x = 0$ and $x = 1$.)

(b) Consider D-dimensional independent Bernoulli densities specified by $p_{ij} \equiv p(x_j = 0|C_i)$ for $i = 1, 2$ and $j = 1, 2, \ldots, D$. Derive the classification rules for classifying a sample x into C_1 and C_2. It is sufficient to give your rule as a function of x.

(c) Follow the definition in 3(b) and assume $D = 2$, $p_{11} = 0.6$, $p_{12} = 0.1$, $p_{21} = 0.6$, and $p_{22} = 0.9$. For three different priors ($P(C_1) = 0.2, 0.6, 0.8$ and $P(C_2) = 1 - P(C_1)$), calculate the posterior probabilities $P(C_1|x)$ and $P(C_2|x)$. (Hint: Calculate the probabilities for all possible samples $(x_1, x_2) \in \{(0, 0), (0, 1), (1, 0), (1, 1)\}$).

Answer:

(a) We have the following 4 equations, derived using the Bayes Rule:

$$P(C_1|x = 0) = \frac{p(x = 0|C_1)P(C_1)}{p(x = 0)} = \frac{p_1 P(C_1)}{p_1 P(C_1) + p_2 P(C_2)}$$
$$P(C_2|x = 0) = \frac{p(x = 0|C_2)P(C_2)}{p(x = 0)} = \frac{p_2 P(C_2)}{p_1 P(C_1) + p_2 P(C_2)}$$
$$P(C_1|x = 1) = \frac{p(x = 1|C_1)P(C_1)}{p(x = 1)} = \frac{(1 - p_1)P(C_1)}{(1 - p_1)P(C_1) + (1 - p_2)P(C_2)}$$
$$P(C_2|x = 1) = \frac{p(x = 1|C_2)P(C_2)}{p(x = 1)} = \frac{(1 - p_2)P(C_2)}{(1 - p_1)P(C_1) + (1 - p_2)P(C_2)}$$

Combining the equations we have:

$$P(C_1|x) = \frac{P(C_1)p_1^{(1-x)}(1 - p_1)^x}{P(C_1)p_1^{(1-x)}(1 - p_1)^x + P(C_2)p_2^{(1-x)}(1 - p_2)^x}$$
$$P(C_2|x) = \frac{P(C_2)p_2^{(1-x)}(1 - p_2)^x}{P(C_1)p_1^{(1-x)}(1 - p_1)^x + P(C_2)p_2^{(1-x)}(1 - p_2)^x}$$

As the denominators are the same, we can ignore them and say:

$$\hat{P}(C_1|x) = P(C_1)p_1^{(1-x)}(1 - p_1)^x$$
$$\hat{P}(C_2|x) = P(C_2)p_2^{(1-x)}(1 - p_2)^x$$

Therefore, if $\hat{P}(C_1|x) > \hat{P}(C_2|x)$, we classify x as C_1. Otherwise, we classify it as C_2.

(b) By the Bayes Rule, we have that

$$P(C_i|x) = \frac{p(x|C_i)P(C_i)}{p(x)} = \frac{p(x|C_i)P(C_i)}{p(x|C_1)P(C_1) + p(x|C_2)P(C_2)}$$

then

$$P(C_1|x) = \frac{p(x|C_1)P(C_1)}{p(x|C_1)P(C_1) + p(x|C_2)P(C_2)}$$
$$P(C_2|x) = \frac{p(x|C_2)P(C_2)}{p(x|C_1)P(C_1) + p(x|C_2)P(C_2)}$$

Since the D dimensions are independent, we can say that

$$P(C_1|x) = \frac{P(C_1)\prod_{j=1}^{D} p(x_j|C_1)}{P(C_1)\prod_{j=1}^{D} p(x_j|C_1) + P(C_2)\prod_{j=1}^{D} p(x_j|C_2)}$$
$$P(C_2|x) = \frac{P(C_2)\prod_{j=1}^{D} p(x_j|C_2)}{P(C_1)\prod_{j=1}^{D} p(x_j|C_1) + P(C_2)\prod_{j=1}^{D} p(x_j|C_2)}$$

where

$$p(x_j|C_i) = p_{ij}^{(1-x_j)}(1-p_{ij})^{x_j}$$

As the denominators are the same, we can ignore then and write:

$$\hat{P}(C_1|x) = P(C_1)\prod_{j=1}^{D} p(x_j|C_1) = P(C_1)\prod_{j=1}^{D} p_{1j}^{(1-x_j)}(1-p_{1j})^{x_j}$$
$$\hat{P}(C_2|x) = P(C_2)\prod_{j=1}^{D} p(x_j|C_2) = P(C_2)\prod_{j=1}^{D} p_{2j}^{(1-x_j)}(1-p_{2j})^{x_j}$$

Therefore, if $\hat{P}(C_1|x) > \hat{P}(C_2|x)$, we classify x as C_1. Otherwise, we classify it as C_2.

(c) Results are the following:

	$P(C_1) = 0.2$	$P(C_1) = 0.6$	$P(C_1) = 0.8$
$P(C_1\|x_1 = 0, x_2 = 0)$	0.0270	0.1429	0.3077
$P(C_2\|x_1 = 0, x_2 = 0)$	0.9730	0.8571	0.6923
$P(C_1\|x_1 = 0, x_2 = 1)$	0.6923	0.9310	0.9730
$P(C_2\|x_1 = 0, x_2 = 1)$	0.3077	0.0690	0.0270
$P(C_1\|x_1 = 1, x_2 = 0)$	0.0270	0.1429	0.3077
$P(C_2\|x_1 = 1, x_2 = 0)$	0.9730	0.8571	0.6923
$P(C_1\|x_1 = 1, x_2 = 1)$	0.6923	0.9310	0.9730
$P(C_2\|x_1 = 1, x_2 = 1)$	0.3077	0.0690	0.0270

4. (**30 points**) Using the provided training, validation, and test datasets, write a Python script to calculate the maximum likelihood estimation on the training set. Consider a prior function defined with respect to sigma as

$$P(C_1|\sigma) = \frac{1}{1+e^{-\sigma}}, \tag{7}$$

and $P(C_2) = 1 - P(C_1)$. Using the learned Bernoulli distributions and the given prior function, classify the samples in the validation set using your classification rules for $\sigma = -5, -4. - 3, -2, -1, 0, 1, 2, 3, 4, 5$. Finally, choose the best prior (the one that gives the lowest error rate on the validation set) and use it to classify the samples in the test set. Print to the Python console (either in terminal or PyCharm) a table of error rate of each prior on the validation set and the error rate using the best prior on the test set. (Hint: if some Bernoulli probabilities are 0, you can replace them with a small probability such as 10^{-10} to avoid the numerical problem.)

Answer:

σ	Validation Error
-5	0.54
-4	0.54
-3	0.54
-2	0.51
-1	0.49
0	0.52
1	0.45
2	0.46
3	0.46
4	0.46
5	0.46

The best prior is obtained for $\sigma = 1$. The test error is 0.475.