CS434 Machine Learning and Data Mining – Homework 0

Exercising Mathematical Knowledge

Student: Ya Zou
Date: 30 Sep,2023

1 Probability

1. **Bayes Theorem and Marginalization [1 pt]** The weatherperson has predicted rain tomorrow, but we don't trust her. Plus we have heard of this new thing called probability and we want to test it out. In recent years, it has rained only 73 days each year (assume there are no leap years in our world such that a year is 365 days). When it actually rains, the weatherperson correctly forecasts rain 70% of the time. When it doesn't rain, she incorrectly forecasts rain 30% of the time. What is the probability that it will rain tomorrow?

   [*Hint*: It is useful to consider two binary random variables – whether it rains or not ($R \in \{0, 1\}$) and whether or not the weatherperson forecasts rain ($F \in \{0, 1\}$). Start by listing all the probabilities the problem provides – e.g. P(F=0|R=0), P(F=1|R=0),P(F=0|R=1), P(F=1|R=1), P(R=1), P(R=0). Then consider how to get P(R=1 | F=1) – the probability that it will rain given that the weatherperson forecasted rain.]

R = 1 : It rains, and otherwise R = 0

F = 1 : Weatherperson forecastes rain, and otherwise F = 0

$P(R=1) = 73/365 = 0.2$ $\quad$ $P(R=0) = 1 - 0.2 = 0.8$

$P(F=1 \mid R=1) = 0.7$ $\quad$ $P(F=1 \mid R=0) = 0.3$

The law of total prob:

$$P(F=1) = P(F=1|R=1) * P(R=1) + P(F=1|R=0) * P(R=0)$$

Bayes Theorem:
$$P(R=1 \mid F=1) = \frac{P(F=1|R=1) \cdot P(R=1)}{P(F=1)}$$

$$= \frac{0.7 \cdot 0.2}{0.7 \cdot 0.2 + 0.3 \cdot 0.8}$$

$$= \frac{0.14}{0.38}$$

$$= 0.3684$$

Thus, It will Rain tomorrow given that the weatherperson forcasted rain is 36.84%

2. **Computing Expected Values from Discrete Distributions [1 pt]** We are machine learners with a slight gambling problem (very different from gamblers with a machine learning problem!). Our friend, Diane, is proposing the following payout on the roll of a fair, 6-sided die:

$$\text{payout} = \begin{cases} \$1 & x = 1 \\ -\$1/4 & x \neq 1 \end{cases} \tag{1}$$

where $x \in \{1, 2, 3, 4, 5, 6\}$ is the outcome of the roll, $(+)$ means payout to us and $(-)$ means payout to Diane. Is this a good bet? That is to say, are we expected to make money if we play?

x	1	2	3	4	5	6
$P(X=x)$	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$

$$E(x) = \sum_{\text{all } x} x\, P(x)$$

$$= 1/6 \cdot (\$1) + \frac{5}{6}(-\$1/4)$$

$$= \frac{1}{6} + \left(-\frac{5}{24}\right)$$

$$= -\frac{1}{24} = -0.0417$$

the expected value is -0.0417, since the number is negative, therefore, it is not a good bet, Because you can expect to lose \$0.0417 on average, if you gamble.

3. **Linearity of Expectation [1 pt]** A random variable x distributed according to a standard normal distribution (mean zero and unit variance) has the following probability density function (pdf):

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{2}$$

Using the properties of expectations, evaluate the following integral

$$\int_{-\infty}^{\infty} p(x)(ax^2 + bx + c)dx \tag{3}$$

[*Hint:* We are not sadistic (okay, we're a little sadistic, but not for this question). This is *not* a calculus question. The simple solution relies on linearity of expectation and the provided mean/variance of p(x).]

$$\int_{-\infty}^{\infty} P(x)(ax^2+bx+c)\,dx$$

$$\Rightarrow \int_{-\infty}^{\infty} P(x)(ax^2)dx + \int_{-\infty}^{\infty} P(x)(bx)dx + \int_{-\infty}^{\infty} P(x)\,dx$$

$$\Rightarrow a\int_{-\infty}^{\infty} P(x)\,x^2dx + b\int_{-\infty}^{\infty} P(x)\,x\,dx + c\int_{-\infty}^{\infty} P(x)\,dx$$

Note: $\int_{-\infty}^{\infty} P(x)\,dx = 1$

$$\Rightarrow a\int_{-\infty}^{\infty} P(x)\,x^2dx + b\int_{-\infty}^{\infty} P(x)\,x\,dx + c(1)$$

Note: $\mu_x = E(x) = \int_{-\infty}^{\infty} x f(x)\,dx = 0$

$$\Rightarrow a\int_{-\infty}^{\infty} P(x)\,x^2dx + b(0) + c(1)$$

Note:

$$Var = \int_{-\infty}^{\infty}(x-\mu_x)^2 f(x)\,dx$$

$$= \int_{-\infty}^{\infty}(x-0)^2 f(x)\,dx$$

$$= \int_{-\infty}^{\infty} x^2 f(x)\,dx$$

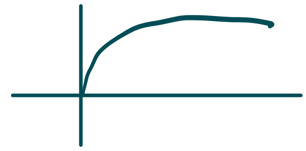
$$\Rightarrow a(1) + b(0) + c(1)$$

$$\Rightarrow a + c$$

thus: the result is $a+c$

4. **Cumulative Density Functions / Calculus [1 pt]** X is a continuous random variable over the interval [0,1] with the probability density function (PDF) shown below.

$$p(x) = \begin{cases} 4x & 0 \leq x \leq 1/2 \\ -4x + 4 & 1/2 \leq x \leq 1 \end{cases} \tag{4}$$



Recall that a cumulative density function (CDF) is defined as $C(x) = P(X \leq x)$ or the probability that a sample from p is less than x – which can be computed as $C(x) = \int_{-\infty}^{x} p(x)\, dx$. Derive the equation for the CDF $C(x)$ corresponding to the PDF in Eq. (4).

[*Hint:* Okay. This one *is* a calculus question. But it is a piece-wise linear function so still not all that sadistic.]

$$\begin{cases} P(x) = 4x, \; 0 \leq x \leq \frac{1}{2} \\ P(x) = -4x + 4, \; \frac{1}{2} \leq x \leq 1 \end{cases}$$

$$C(x) = \int_0^x 4t\, dt \Rightarrow C(x) = -2t^2 + 4t \Big|_0^{\frac{1}{2}}$$

$$\Rightarrow C(x) = -2t^2 + 4t - (\tfrac{1}{2} + 2) \Rightarrow C(x) = 2t^2, \; t \in [0, \tfrac{1}{2}]$$

$$C(x) = \int_{\frac{1}{2}}^{x} -4t + 4\, dt \Rightarrow C(x) = -2t^2 + 4t - (-2(\tfrac{1}{2})^2 + 4(\tfrac{1}{2}))$$

$$\Rightarrow C(x) = -2t^2 + 4t - \tfrac{5}{2}, \; t \in [\tfrac{1}{2}, 1]$$

Thus:

$$\begin{cases} C(x) = 2t^2, \; t \in [0, \frac{1}{2}] \\ C(x) = -2t^2 + 4t - \frac{5}{2}, \; t \in [0, 1] \end{cases}$$

2 Linear Algebra

1. **Transpose and Associative Property [1pt]** Define matrix $B = bb^T$, where $b \in \mathbb{R}^{d\times 1}$ is a column vector that is not all-zero. Show that for any vector $x \in \mathbb{R}^{d\times 1}$, $x^T Bx \geq 0$.

   [*Hint*: Try to get $x^T Bx$ to look like the product of two identical scalars. Note that $b^T x = (x^T b)^T$, that $a^T = a$ for scalar value a, and that matrix multiplication is associative.]

2. **Solving Systems of Linear Equations with Matrix Inverse [1pt]** Consider the following system of equations:

$$2x_1 + x_2 + x_3 = 3,$$
$$4x_1 + 2x_3 = 10,$$
$$2x_1 + 2x_2 = -2.$$

   (a) Write the system as a matrix equation of the form $Ax = b$.

   (b) Solve for $Ax = b$ by using the matrix inverse of A (You can definitely use software to get the inverse).

1. Any vector $b = \begin{bmatrix} 1 \\ 4 \\ 3 \end{bmatrix}$ $b^T = \begin{bmatrix} 1 & 4 & 3 \end{bmatrix}$ where $b \in R^{d\times 1}$ and b is not all-zero,

$$B = bb^T = \begin{bmatrix} 1 \\ 4 \\ 3 \end{bmatrix} \cdot \begin{bmatrix} 1 & 4 & 3 \end{bmatrix} = \begin{bmatrix} 1 & 4 & 3 \\ 4 & 16 & 12 \\ 3 & 12 & 9 \end{bmatrix}$$

$$b^T B b = \begin{bmatrix} 1 & 4 & 3 \end{bmatrix} \cdot \begin{bmatrix} 1 & 4 & 3 \\ 4 & 16 & 12 \\ 3 & 12 & 9 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 4 \\ 3 \end{bmatrix}$$

$$= \begin{bmatrix} 26 & 104 & 78 \end{bmatrix} \begin{bmatrix} 1 \\ 4 \\ 3 \end{bmatrix} = 676$$

thus $b^T B b = 676$ which is greater than 0

Therefore, any vector $x \in R^{d\times 1}$, $x^T Bx \geq 0$

2. (a)

$$\underbrace{\begin{bmatrix} 2 & 1 & 1 \\ 4 & 0 & 2 \\ 2 & 2 & 0 \end{bmatrix}}_{A} \cdot \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}}_{x} = \underbrace{\begin{bmatrix} 3 \\ 10 \\ -2 \end{bmatrix}}_{B}$$

$$x = B$$

(b)

$$X = A^{-1} \cdot B$$

$$X = \underbrace{\begin{bmatrix} -1 & 1/2 & 1/2 \\ 1 & -1/2 & 0 \\ 2 & -1/2 & -1 \end{bmatrix}}_{A^{-1}} \cdot \underbrace{\begin{bmatrix} 3 \\ 10 \\ -2 \end{bmatrix}}_{B}$$

$$X = \begin{bmatrix} 1 \\ -2 \\ 3 \end{bmatrix} \Rightarrow \begin{cases} x_1 = 1 \\ x_2 = -2 \\ x_3 = 3 \end{cases}$$

3 Proving Things

1. **Finding Maxima of a Function [2pt]** Prove that $\ln x \le x - 1, \quad \forall x > 0$ with equality if and only if $x = 1$.
[*Hint:* Consider the function $f(x) = \ln(x) - (x - 1)$ and show its maximum value for x > 0 is f(1)=0. This will be obvious from plots, but you'll need some calculus to actually prove it.]

$f(x) = \ln(x) - (x-1)$

$f'(x) = \frac{1}{x} - 1$ → To find critical points, we can set $f'(x) = 0$ thus $\frac{1}{x} - 1 = 0 \rightarrow x = 1$

$f''(1) = \left(-1/1^2\right) = -1$ since $f''(1)$ is negative,

It means that $x=1$ corresponds to a local maximum

To comfir that equality holds if and only if $x=1$, we have $\ln(1) - (1-1) = 0$, so for $x=1$, both side equal to 0, which satisfy the inequality.

Thus, we used function $f(x) = \ln x - (x-1)$, $\forall x > 0$, we've proven that $\ln(x) \le x - 1$, $\forall x > 0$ with equality if and ony if $x=1$.

2. **Proving Abstract Concepts [2pt]** Consider two discrete probability distributions p and q over k outcomes:

$$\sum_{i=1}^{k} p_i = \sum_{i=1}^{k} q_i = 1 \tag{5a}$$

$$p_i > 0, q_i > 0, \quad \forall i \in \{1, \dots, k\} \tag{5b}$$

The Kullback-Leibler (KL) divergence between p and q is given by:

$$KL(p||q) = \sum_{i=1}^{k} p_i \ln\left(\frac{p_i}{q_i}\right) \tag{6}$$

It is common to refer to $KL(p||q)$ as a measure of distance (even though it is not a proper metric). Many algorithms in machine learning are based on minimizing KL divergence between two probability distributions.

Using the results from part 1, show that $KL(p||q)$ is non-negative – i.e. $KL(p||q) \ge 0$

[*Hint:* This question can be solved using the definition of $KL(p||q)$, the inequality from 3.1, recalling that $log(a/b) = -log(b/a)$, and the constraint in 5a/b.]

We need to prove $KL(P||q) \ge 0$, we are goiy to prove it by contradiction that $KL(P||q) \le 0$.

$KL(P||q) = \sum_{i=1}^{k} P_i(\ln(P_i|q_i))$, we know $\log(\frac{a}{b}) = -\log(\frac{b}{a})$,

thus $KL(P||q) = \sum_{i=1}^{k} P_i(-\ln(q_i|P_i))$, from 3.1 we know that any positive numbers a and b, $\ln(a/b) \leq (a-b)/b$ with equality if only if $a=b$:

$$-\ln(q_i/p_i) \leq (q_i/p_i - 1)$$, we can subsitute

$$\Rightarrow KL(P\|q) = \sum_{i=1}^{k} p_i * (-\ln(q_i/p_i)) \leq \sum_{i=1}^{k} p_i * (q_i/p_i - 1)$$

$\Rightarrow$ simplify . $KL(P\|q) \leq 1 * \sum_{i=1}^{k} (q_i/p_i - 1)$

$$\Rightarrow KL(P\|q) \leq \sum_{i=1}^{k} (q_i - p_i)$$

$$\Rightarrow KL(P\|q) \leq \sum_{i=1}^{k} q_i - \sum_{i=1} p_i$$

$\Rightarrow$ From (5a) we know $\sum_{i=1}^{k} p_i = \sum_{i=1}^{k} q_i = 1$, thus we get:

$$\Rightarrow KL(P\|q) \leq 1-1$$

$$\Rightarrow KL(P\|q) \leq 0$$

Therefore, we have shown above that $KL(P\|q) \leq 0$ since it log can not be negative, it proved that $KL(P\|q)$ is non-negative:

$$KL(P\|q) \geq 0$$

4 Debriefing (required in your report)

1. Approximately how many hours did you spend on this assignment?

   I approximately spend 3 hours on this assignment.

2. Would you rate it as easy, moderate, or difficult?

   I would rate in higher than moderate but low than difficult.

3. Did you work on it mostly alone or did you discuss the problems with others?

   I worked mostly myself, I discussed the prove part and probability with my friends.

4. How deeply do you feel you understand the material it covers (0%–100%)?

   I feel I understand 86% the material covers

5. Any other comments?

   I don't have other comments.