CS434 Machine Learning and Data Mining – Homework 0

Exercising Mathematical Knowledge

Solutions - DO NOT DISTRIBUTE

Overview and Objectives. In this homework, we are going to practice some of the skills we'll be using in class. If you can solve most of these problems (even if you have to Google some identities or brush up on your knowledge) then you should be very well-equipped for the course. Otherwise, you'll need to spend extra time revising these topics. **This assignment will be graded on completion rather than correctness. This assignment is to help you (and us) gauge your familiarity with these concepts and might be a bit challenging depending on your background.**

How to Do This Assignment. We prefer solutions typeset in LaTeX but will accept scanned written work if it is legible. If a TA can't read your work, they can't give you credit. Submit your solutions to Canvas as a PDF.

Advice. Start early. Start early. Start early. You may be rusty on some of this material. Some of it might be new to you depending on your background – seek out resources to refresh yourself if so. Some helpful references:

- **Probability Refresher:** CS229 Probability Review from Stanford
- **Linear Algebra Refresher:** Zico Koltur's Linear Algebra Review
- **Differential Calculus Refresher:** Jackie Nicholas' booklet from University of Sydney
- **Integration Refresher:** Mary Barnes' booklet from University of Sydney

Also plenty of great videos online. Don't be scared. Let's begin!

Nearly all of the following questions rely on definitions of concepts in each area. The bolded question titles indicate which concepts are being tested. Googling these bolded concepts can also be a good place to start.

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.]

First we collect the probabilities given in the problem and associate them with our random variables:

$$P(R=1) = 73/365 = 0.2 \quad \text{[Prob. of Rain]}$$
$$P(R=0) = 292/365 = 0.8 \quad \text{[Prob. of No Rain]}$$
$$P(F=1|R=1) = 0.7 \quad \text{[Prob. of Forecasting Rain when it does actually rain]}$$
$$P(F=0|R=1) = 0.3 \quad \text{[Prob. of Forecasting No Rain when it does actually rain]}$$
$$P(F=1|R=0) = 0.3 \quad \text{[Prob. of Forecasting Rain when it doesn't actually rain]}$$
$$P(F=0|R=0) = 0.7 \quad \text{[Prob. of Forecasting No Rain when it doesn't actually rain]}$$

We know the weatherperson has forecasted rain, so we want to know $P(R=1|F=1)$.

$$\begin{aligned} P(R=1|F=1) &= \frac{P(F=1|R=1)P(R=1)}{P(F=1)} && \text{[Bayes Theorem]} \\ &= \frac{P(F=1|R=1)P(R=1)}{P(F=1,R=0)+P(F=1,R=1)} && \text{[Marginalization]} \\ &= \frac{P(F=1|R=1)P(R=1)}{P(F=1|R=0)P(R=0)+P(F=1|R=1)P(R=1)} && \text{[Chain Rule]} \\ &= \frac{0.7*0.2}{0.3*0.8+0.7*0.2} \approx 0.368 \end{aligned}$$

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:

$$\text{payout} = \begin{cases} \$1 & x = 1 \\ -\$1/4 & x \neq 1 \end{cases} \tag{2}$$

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?

Assuming a fair die, there is a 1/6 chance of landing on any number,

$$p(1) = \frac{1}{6}; \qquad p(\text{not } 1) = \frac{5}{6}$$

The expected outcome for a turn is

$$\$1\left(\frac{1}{6}\right) - \$\frac{1}{4}\left(\frac{5}{6}\right) = -\$\frac{1}{24}$$

So, we will lose money. Thus, it is not a good deal.

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):

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{3}$$

Using the properties of expectations, evaluate the following integral

$$\int_{-\infty}^{\infty} p(x)(ax^2+bx+c)dx \tag{4}$$

[*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).]

Note that equation (3) is an expectation of $ax^2 + bx + c$ with respect to the normal random variable x.

$$E[ax^2 + bx + c] = \int_{-\infty}^{\infty} p(x)(ax^2 + bx + c)dx$$

Applying linearity of expectation breaks this down to:

$$E[ax^2 + bx + c] = aE[x^2] + bE[x] + c$$

For standard normal distribution (0 mean, variance 1), we have,

$$E[x] = \int_{-\infty}^{\infty} p(x)x \; dx = 0 \qquad \text{[This is just the mean.]}$$

$$VAR[x] = E[x^2] - E[x]^2 = \left(\int_{-\infty}^{\infty} p(x)x^2 \; dx \right) - 0 = 1 \qquad \text{[This is just the variance.]}$$

$$\rightarrow E[x^2] = 1$$

Hence,

$$aE[x^2] + bE[x] + cE[1] = a + c$$

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.

$$p(x) = \begin{cases} 4x & 0 \leq x \leq 1/2 \\ -4x + 4 & 1/2 \leq x \leq 1 \end{cases} \tag{6}$$

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. (6).

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

This question comes down to integrating the piece-wise linear function $p(x)$ with respect to x. Remember that $C(x)$ is supposed to equal $\int_{-\infty}^{x} p(x) \; dx$ so we'll need to accumulate the $0 \leq x \leq 1/2$ piece of the integral when writing the expression for the $1/2 \leq x \leq 1$ portion. For similar reasons, we'll also need to define $C(x)$ for the regions outside the interval $[0, 1]$

$$C(x) = \begin{cases} 0 & x \leq 0 \\ \int_{0}^{x} 4x \; dx = 2x^2 & 0 \leq x \leq 1/2 \\ (\int_{0}^{1/2} 4x \; dx) + (\int_{1/2}^{x} -4x + 4 \; dx) = -2x^2 + 4x - 1 & 1/2 \leq x \leq 1 \\ 1 & x \geq 1 \end{cases}$$

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.]

First note that as x and b are both column vectors (d x 1 dimension), $x^T b$ results in a scalar value (product of matrices of dimensions 1 x d and d x 1 yields 1 x 1). Let's call this scalar value $a = x^T b$

$$
\begin{aligned}
x^T B x &= x^T b b^T x && \text{[Definition of B]} \\
&= x^T b (x^T b)^T && \text{[Property of Transpose]} \\
&= a(a)^T = a^2 && \text{[Our definition of a. And the transpose of a scalar is a scalar]} \\
\therefore \quad a^2 &\geq 0 \quad \blacksquare && [a^2 \text{is non-negative for real-valued } a]
\end{aligned}
$$

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

$$
\begin{aligned}
2x_1 + x_2 + x_3 &= 3, \\
4x_1 + 2x_3 &= 10, \\
2x_1 + 2x_2 &= -2.
\end{aligned}
$$

(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).

a) Let $A = \begin{bmatrix} 2 & 1 & 1 \\ 4 & 0 & 2 \\ 2 & 2 & 0 \end{bmatrix}$, $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$, and $\mathbf{b} = \begin{bmatrix} 3 \\ 10 \\ -2 \end{bmatrix}$. We can then write this system of equations as:

$$
Ax = b \longrightarrow \begin{bmatrix} 2 & 1 & 1 \\ 4 & 0 & 2 \\ 2 & 2 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 3 \\ 10 \\ -2 \end{bmatrix}
$$

b) Recall from the definition of the matrix inverse that $AA^{-1} = A^{-1}A = I$ for invertible matrix A, where I is the identity matrix. We can then left-multiply the equation $A\mathbf{x} = \mathbf{b}$ by A^{-1} to arrive at a expression for $\mathbf{x}$.

$$
\begin{aligned}
A^{-1}Ax &= A^{-1}b \\
Ix &= A^{-1}b \\
x &= A^{-1}b
\end{aligned}
$$

where $A^{-1} = \begin{bmatrix} -1 & -0.5 & 0.5 \\ 1 & -0.5 & 0 \\ 2 & -1/2 & -1 \end{bmatrix}$ such that

$$
\mathbf{x} = A^{-1}b = \begin{bmatrix} -1 & 0.5 & 0.5 \\ 1 & -0.5 & 0 \\ 2 & -1/2 & -1 \end{bmatrix} \begin{bmatrix} 3 \\ 10 \\ -2 \end{bmatrix} = \begin{bmatrix} 1 \\ -2 \\ 3 \end{bmatrix}
$$

You can plug these values back into the original system of equations if you want to reassure yourself that we've found the right answer.

3 Proving Things

1. **Finding Maxima of a Function [2pt]** Prove that $\ln x \leq x - 1, \quad \forall x > 0$ 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.]

Following the hint, let's consider the maximum of the function $f(x) = ln(x) - x + 1$. Taking the derivative:

$$f'(x) = \frac{1}{x} - 1$$

Setting this equal to zero reveals a critical point of $f(x)$ at $x = 1$. As the second derivative $f''(x) = -1/x^2$ is negative for all x, we know f is a concave function and thus the only critical point is the global maximum at $f(1) = 0$. $\therefore \quad ln(x) \leq x - 1 \quad \forall x > 0$ ■

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

$$\sum_{i=1}^{k} p_i = \sum_{i=1}^{k} q_i = 1 \tag{7a}$$

$$p_i > 0, q_i > 0, \quad \forall i \in \{1, \ldots, k\} \tag{7b}$$

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

$$KL(p||q) = \sum_{i=1}^{k} p_i \ln\left(\frac{p_i}{q_i}\right) \tag{8}$$

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. $KL(p||q) \geq 0$

[*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.]

Let $x = \frac{q_i}{p_i}$. We know that $\frac{q_i}{p_i} \geq 0$ as $q_i, p_i > 0 \quad \forall i$. So from the previous question, we know

$$\ln\left(\frac{q_i}{p_i}\right) \leq \frac{q_i}{p_i} - 1$$

or equivalently that

$$-\ln\left(\frac{q_i}{p_i}\right) \geq -\left(\frac{q_i}{p_i} - 1\right)$$

$$\begin{aligned}
KL(p||q) = \sum_{i=1}^{k} p_i \ln\left(\frac{p_i}{q_i}\right) &= -\sum_{i=1}^{k} p_i \ln\left(\frac{q_i}{p_i}\right) && \text{[Uses } \ln(a/b) = -\ln(b/a)\text{]} \\
&\geq -\sum_{i=1}^{k} p_i \left(\frac{q_i}{p_i} - 1\right) && \text{[Applies the inequality]} \\
&= -\sum_{i=1}^{k} (q_i - p_i) \\
&= -\sum_{i=1}^{k} q_i + \sum_{i=1}^{k} p_i \\
&= -1 + 1 = 0 && \text{[Discrete probability distributions sum to 1]}
\end{aligned}$$

$$\therefore \quad KL(p||q) \geq 0 \quad ■$$