Probabilistic Machine Learning

Exercises W3

Ralf Herbrich, Jakob Walter

SS 2023

E-mail: {ralf.herbrich, jakob.walter}@hpi.de
Web: https://hpi.de

Lectures: Mo 9:15-10:45
Tutorial Location: L-E.03
Lecture Hall: L-E.03
Tutorials: Tu 13:30-15:00

1 Mathematical Exercises

The expected value of a function $g(X)$ of a random variable X with probability mass function p can be written as

$$E\left[g(X)\right] = \sum_{x} g(x)p(x). \tag{1}$$

where the sum is taken over all possible values x of X. If X is a continuous random variable with pdf p, then the same holds true:

$$E\left[g(X)\right] = \int_{-\infty}^{\infty} g(x)p(x)dx. \tag{2}$$

Lastly, a similar property also holds for joint distributions. Let g be a function of the discrete random variables X and Y with joint pmf $p(x, y)$. Then,

$$E\left[g(X, Y)\right] = \sum_{x}\sum_{y} g(x, y)p(x, y). \tag{3}$$

An equivalent property also holds for continuous random variables.

Exercise 1:

1. Use equation (3) to show that

$$E\left[a \cdot X + b \cdot Y\right] = a \cdot E\left[X\right] + b \cdot E\left[Y\right] \tag{4}$$

holds for discrete random variables X and Y and any $a \in \mathbb{R}$ and $b \in \mathbb{R}$. Using Fubini's theorem, it can be shown that the same also holds true for continuous random variables.

2. The variance of a random variable is defined to be

$$\mathrm{var}[X] = E\left[(X - E[X])^2\right] \overset{(a)}{=} E[X^2] - (E[X])^2 . \tag{5}$$

Using equation (4), show that (a) is indeed true.

3. Lastly, the covariance between two random variables is defined to be

$$\mathrm{cov}[X, Y] = E\left[(X - E[X]) \cdot (Y - E[Y])\right] \overset{(b)}{=} E[XY] - E[X] \cdot E[Y]. \tag{6}$$

Again, verify that (b) is true. You will need to use equation (4).

4. Verify that

$$\mathrm{var}[a \cdot X + b \cdot Y] = a^2 \cdot \mathrm{var}[X] + b^2 \cdot \mathrm{var}[Y] + 2ab \cdot \mathrm{cov}[X, Y]. \tag{7}$$

You can use equation (4), (5) and (6).

5. Verify that

$$\mathrm{cov}\left[a \cdot X + b \cdot Y, c \cdot W + d \cdot V\right] = ac \cdot \mathrm{cov}[X, W] + ad \cdot \mathrm{cov}[X, V] + bc \cdot \mathrm{cov}[Y, W] + bd \cdot \mathrm{cov}[Y, V]. \tag{8}$$

Solution 1:

1. Expanding the left-hand side of (4) using (3) we have

$$\begin{aligned}
E[a \cdot X + b \cdot Y] &= \sum_x \sum_y (a \cdot x + b \cdot y) \cdot p(x, y) \\
&= a \cdot \left(\sum_x \sum_y x \cdot p(x, y) \right) + b \cdot \left(\sum_x \sum_y y \cdot p(x, y) \right) \\
&= a \cdot \left(\sum_x x \cdot \underbrace{\sum_y p(x, y)}_{p(x)} \right) + b \cdot \left(\sum_y y \cdot \underbrace{\sum_x p(x, y)}_{p(y)} \right) \\
&= a \cdot \left(\sum_x x \cdot p(x) \right) + b \cdot \left(\sum_y y \cdot p(y) \right) \\
&= a \cdot E[X] + b \cdot E[Y] .
\end{aligned}$$

2. Expanding the left-hand side of (5) we have

$$\begin{aligned}
\mathrm{var}[X] = E\left[(X - E[X])^2\right] &= E\left[X^2 - 2 \cdot X \cdot E[X] + (E[X])^2\right] \\
&= E\left[X^2\right] - 2 \cdot (E[X])^2 + (E[X])^2 \\
&= E\left[X^2\right] - (E[X])^2 ,
\end{aligned}$$

where we used (4) and $E\left[E[X]\right] = E[X]$ in the second line.

3. Expanding the left-hand side of (6) we have

$$\begin{aligned}\mathrm{cov}[X] &= E\left[(X - E[X]) \cdot (Y - E[Y])\right] \\ &= E\left[XY - X \cdot E[Y] - Y \cdot E[X] + E[X] \cdot E[Y]\right] \\ &= E[XY] - 2 \cdot E[X] \cdot E[Y] + E[X] \cdot E[Y] \\ &= E[XY] - E[X] \cdot E[Y]\,,\end{aligned}$$

where we used (4) and $E\left[E[X]\right] = E[X]$ and $E\left[E[Y]\right] = E[Y]$ in the third line.

4. Using the result from (5) we have

$$\begin{aligned}\mathrm{var}\left[a \cdot X + b \cdot Y\right] &= E\left[(a \cdot X + b \cdot Y)^2\right] - \left(E[a \cdot X + b \cdot Y]\right)^2 \\ &= E\left[(a \cdot X)^2 + 2ab \cdot XY + (b \cdot Y)^2\right] - \left(a \cdot E[X] + b \cdot E[Y]\right)^2 \\ &= a^2 \cdot E\left[X^2\right] + 2ab \cdot E[XY] + b^2 \cdot E\left[Y^2\right] - a^2 \cdot (E[X])^2 - 2ab \cdot E[X] \cdot E[Y] - b^2 \cdot (E[Y])^2 \\ &= a^2 \cdot \left(E\left[X^2\right] - (E[X])^2\right) + b^2 \cdot \left(E\left[Y^2\right] - (E[Y])^2\right) + 2ab \cdot \left(E[XY] - E[X] \cdot E[Y]\right) \\ &= a^2 \cdot \mathrm{var}[X] + b^2 \cdot \mathrm{var}[Y] + 2ab \cdot \mathrm{cov}[X, Y]\,,\end{aligned}$$

where we used (4) in the second line and (5) and (6) in the final line.

5. Using the result from (6) we have

$$\begin{aligned}&\mathrm{cov}[a \cdot X + b \cdot Y, c \cdot W + d \cdot V] \\ &= E\left[(a \cdot X + b \cdot Y) \cdot (c \cdot W + d \cdot V)\right] - E\left[a \cdot X + b \cdot Y\right] \cdot E\left[c \cdot W + d \cdot V\right] \\ &= E[ac \cdot XW] + E[ad \cdot XV] + E[bc \cdot YW] + E[bd \cdot YV] - (a \cdot E[X] + b \cdot E[Y]) \cdot (c \cdot E[W] + d \cdot E[V]) \\ &= ac \cdot (E[XW] - E[X]E[W]) + ad \cdot (E[XV] - E[X]E[V]) + bc \cdot (E[YW] - E[Y]E[W]) + bd \cdot (E[YV] - E[Y]E[V]) \\ &= ac \cdot \mathrm{cov}[X, W] + ad \cdot \mathrm{cov}[X, V] + bc \cdot \mathrm{cov}[Y, W] + bd \cdot \mathrm{cov}[Y, V]\,,\end{aligned}$$

where we used (6) in the second line, (4) in the third line and (6) in the final line.

Exercise 2:

For this exercise, we will first need to define concavity and then introduce Jensen's inequality. A function $f : X \to \mathbb{R}$ is called *concave* if and only if it holds that for all $0 \leq t \leq 1$ and all $x_1, x_2 \in X$:

$$f(t \cdot x_1 + (1-t) \cdot x_2) \geq t \cdot f(x_1) + (1-t) \cdot cf(x_2). \tag{9}$$

Note that a concave function is thus simply the negative of a convex function. Checking this inequality can be cumbersome. For differentiable functions, checking concavity can be simplified by checking one of the following conditions:

1. A differentiable function f is concave on an interval if and only if its derivative function f' is monotonically decreasing on that interval.
2. If f is twice-differentiable, then f is concave if and only if f'' is non-positive. I.e. $f''(x) \leq 0$ for all $x \in X$.

Having defined concavity of functions, we can now define Jensen's inequality. In the context of probability theory, the inequality can be stated in the following form: If X is a random variable, and f is a concave function, then

$$E\left[f(x)\right] \leq f\left(E[X]\right). \tag{10}$$

You should now have all necessary knowledge to proof the following:

Let X be an M-state discrete random variable. That is, X takes on the values $x_1, x_2, \dots, x_M$. Use Jensen's inequality, to show that the entropy of X, satisfies

$$H[X] \leq \log(M). \tag{11}$$

Solution 2:

Using the definition of the entropy $H[X]$ we have

$$\begin{aligned} H[X] &= -E\left[\log(p(X))\right] \\ &= E\left[\log\left(\frac{1}{p(X)}\right)\right] \\ &\leq \log\left(E\left[\frac{1}{p(X)}\right]\right) \\ &= \log\left(\sum_{i=1}^{M} \frac{1}{p(x_i)} \cdot p(x_i)\right) \\ &= \log(M)\,, \end{aligned}$$

where the third line follows from the concavity of the logarithm.

2 Programming

In this exercise, you should build some intuition about the relationship between prior and posterior distributions and the likelihood. To do so, do the following:

1. Define a Beta distribution with parameters α and β.
2. Fix your true (unknown) parameter $\tilde{\pi}$ to 0.5.
3. Generate n data points $x_i|\tilde{p} \sim \mathrm{ber}(\tilde{\pi})$.
4. Compute the posterior distribution $p(\pi|\mathbf{x})$ of π given the data $\mathbf{x} = x_1, x_2, \dots, x_n$. We are using a conjugate prior for the likelihood function, thus our posterior is again a Beta-distribution. It has parameters

$$\alpha_{\text{post}} = \alpha + \sum_{i=1}^{n} x_i \tag{12}$$

$$\beta_{\text{post}} = \beta + n - \sum_{i=1}^{n} x_i. \tag{13}$$

You should now make a plot of the likelihood, the prior and the posterior. Normalize the likelihood so that it's value lies in $[0, 1]$. You can normalize a vector $\mathbf{l}$ using

$$\mathbf{l}_{\text{scaled}} = \frac{\mathbf{1} - l_{\min}}{l_{\max} - l_{\min}}. \tag{14}$$

where $l_{\min}$ and $l_{\max}$ are the minimum and maximum of $\mathbf{l}$ respectively.

You should now be able to play around with the code to answer the following questions:

1. For $\tilde{\pi} = 0.5$ compare the relationship between prior, posterior and likelihood for the following settings:

- $\alpha = \beta = 0.5$
- $\alpha = \beta = 1$
- $\alpha = \beta = 10$
- $\alpha = 1, \beta = 5$
- $\alpha = 5, \beta = 1$

It is said that the posterior is a compromise between prior and likelihood. Can you see why?

2. What is the relationship between likelihood and posterior when you choose a uniform prior ($\alpha = \beta = 1$)?

3. Set your true parameter to $\tilde{\pi} = 0.25$ and use the informative prior with $\alpha = \beta = 10$ What effect does the number of samples n have on the relationship between prior and posterior? Create a vector of n's using `n_list = Int.(unique(round.(exp10.(range(0, 5, length = 100)))))` and make two plots. The first one should show the relationship between n and the posterior mean. The second one should show the relationship between the posterior variance and n.