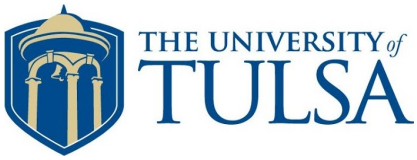


Deep Learning
CS-6863

Homework 1

Instructor:
Mahdi Khodayar

Author:
Selim Karaoglu

February 16, 2022

Section 1: Probability review:

Consider a sample of data S = 1,1,0,1,0 created by flipping a coin x five times, where 0 denotes that the coin turned up heads and 1 denotes that it turned up tails.

1. what is the sample mean for this data?

   solution

   The **sample mean** is calculated by simply taking the average of all the measurements in the sample. Given the features x for each variable on dataset, sample mean of the dataset can be calculated with:

   $$\bar{x} = \frac{\Sigma x_i}{n} = \frac{x_1 + x_2 + ... + x_n}{n}$$

   Applying sample mean formula to the given dataset S results with;

   $$\bar{x} = \frac{1+1+0+1+0}{5} = \frac{3}{5} = 0.6$$

2. what is the sample variance for this data?

   solution

   Sample variance measures how spread out the data in a dataset is. Two datasets may have the same mean but could have been distributed very differently. Variance is employed in statistics to quantify these differences. To find the sample variance:

   - Calculate the sample mean $\bar{x}$.
   - Subtract the $\bar{x}$ value from the value of each measurement (x_i).
   - Square the resulting values.
   - Add the results together to get the sum of squared deviations from the mean.
   - Finally, divide this by the total number of measurements minus one (n -1)

   In equation form, this looks like:

   $$s^2 = \left(\frac{1}{n-1}\right)\sum_{i=1}^{n}(x_i - \bar{x})^2$$

   Sample variation of given data can be calculated with this formula.

   - $\bar{x}$ is already calculated in previous question. $\bar{x} = 0.6$
   - $(x_i - \bar{x})^2$ for each data point in the data results with; $(0.16, 0.16, 0.36, 0.16, 0.36)$
   - The sum of squares is all the squared differences added together.
     $SS = \sum_{i=1}^{n}(x_i - \bar{x})^2 = (0.16 + 0.16 + 0.36 + 0.16 + 0.36).0.25 = 0.3$

3. what is the probability of observing this data, assuming it was generated by flipping a coin with an equal probability of heads and tails (i.e., the probability distribution is P(x=1)=0.5, P(x=0)=0.5)?

   solution

   The sample presented in this examples have 5 features. In a data with n features, the event space is 2^n. This sample has 5 features, so the event space is 32. This sample is just one of the 32 possible outcome. Therefore; probability of observing this data is $\frac{1}{32} = 0.03125$

4. Note that the probability of this data sample would be greater if the value of P(x=1) was not 0.5, but instead some other value. What is the value that maximizes the probability of the sample S. Please justify your answer. Plugging in our values for $x_1, x_2, ..., X_5$ into the above formula, we find that the best $p = \frac{3}{5}$.

   solution

   The goal is to find the value $p(x = 1)$ that maximizes the probability of sample S. The probability of S can be interpreted as:

   $$\prod_{i=1}^{5} p^{x_i}(1-p)^{(1-x_i)} = p^{\sum_{i=1}^{5} x_i}(1-p)^{n-\sum_{i=1}^{5} x_i}$$

As a function of p, the given formula should be maximized. *log* of this formula, $l(p)$, can be written as:

$$l(p) = \left(\sum_{i=1}^{5} x_i\right) log(p) + \left(n - \sum_{i=1}^{5} x_i\right) log(1-p)$$

The p value that maximizes the $p(x = 1)$ can be obtained by calculating the p that maximizes the probability of $l(p)$. To achieve this, derivative of $l(p)$ is taken with respect to p, set to zero and solve for p.

$$\frac{dl(p)}{dp} = \frac{1}{p}\sum_{i=1}^{5} x_i - \frac{1}{(1-p)}\left(n - \sum_{i=1}^{5} x_i\right) = 0$$

$$\implies \frac{\sum_{i=1}^{5} x_i - pn}{p(1-p)} = 0$$

$$\implies pn = \sum_{i=1}^{5} x_i$$

$$\implies p = \frac{1}{n}\sum_{i=1}^{5} x_i$$

Now if the values of x are inserted to the formula, the best p value appers to be $\frac{3}{5}$.

5. Consider the following joint probability table over variables y and z, where y takes a value from the set a, b, c and z takes a value from the set T, F.

	y		
2*z	a	b	c
T	0.2	0.1	0.2
F	0.05	0.15	0.3

What is P(z=T and y=b)?

solution

As can be observed from above table, P(z=T and y=b) is equal to 0.1.

What is P(z=T | y=b)?

solution

To solve this problem, the conditional probability can be employed.

$p(z = T|y = b) = \frac{p(z=T \bigwedge y=b)}{p(y=b)} = \frac{0.1}{0.1+0.15} = 0.4$

6. Match the distribution name to its probability density function (PDF):

a) Multivariate Gaussian
b) Exponential
c) Uniform
d) Binomial
e) Bernoulli

f) $p^x(1-p)^{(1-x)}$
g) $\frac{1}{b-a}$ when $a \leq x \leq b$; or 0
h) $\binom{n}{x} p^x(1-p)^{n-x}$
i) $\lambda e^{-\lambda x}$ when $x \geq 0$; or 0.
j) $\frac{1}{\sqrt{(2\pi)^d|\sum|}} exp(-\frac{1}{2} - (x-\mu)^\top \sum^{-1}(x-\mu))$

solution

The matching names and functions are: "a with j", "b with i", "c with g", "d with f" and "e with h".

7. What is the mean, variance, and entropy of Bernoulli(p) random variable?

solution

The mean is p, the variance is $p(1-p)$, and the entropy is $-(1-p)log(1-p) - plog(p)$.

8. If the variance of a zero-mean random variable x is σ^2 what is the variance of $2x$? What about the variance of $x + 2$?

   solution

   The variance σ^2 of a random valuable X with expected value $E(X) = \mu$ can be interpreted as $var(X) = E((X - \mu)^2)$. So by definition, σ^2 is not affected by added constant. $var(X + C) = var(X)$ for every constant C because $(X + C) - E(X + C) = X - E(X)$ since the constants are cancelling. However, multiplying the random variable X with constant C will result differently; $var(CX) = C^2 var(X)$ because $(CX - E(CX))^2 = C^2(X - EX)^2$. So for each constant a and b:

$$var(a + bX) = b^2 var(X)$$

9. If X and Y are independent random variables, show that E[XY] = E[X] E[Y]

   solution

   Two variables are independent if the knowledge of the first variable does not influence the results of the second variable and vice versa. This can be notated mathematically with given X takes a certain value α, does not change Y takes a value, say β. Therefore;

   X and Y are indepented if $P(Y = \beta \cap X = \alpha) = P(X = \alpha)P(Y = \beta)$ for all α, β. Suppose X and Y are independent variables, then;

$$\begin{aligned} E[XY] &= \sum_{\alpha,\beta} \alpha\beta P(X = \alpha, Y = \beta) \\ &= \sum_{\alpha,\beta} \alpha\beta P(X = \alpha)P(Y = \beta) \\ &= \sum_{\alpha} \alpha P(X = \alpha) \sum_{\beta} (Y = \beta) \\ &= E[X]E[Y] \end{aligned}$$

10. Alice rolls a die and calls up Bob and Chad to tell them the outcome A. Due to disturbance in the phones, Bob and Chad think the roll was B and C, respectively. Is B independent of C ? Is B independent of C given A ?

    solution

    B is not independent of C. B is independent of C given A.

11. For events A and B, prove $P(A|B) = (P(B|A)P(A))/P(B)$

    solution

    The probability of the events A and B happening, $P(A \cap B)$, is equal to the probability of A, $P(A)$ multiplied by probability of B given that A has occured, $P(B|A)$. When formulated; $P(A \cap B) = P(A)P(B|A)$. At the same time, the probability of A and B can also be calculated as probability of B times the probabilty of A given B. Which can be formulated as; $P(A \cap B) = P(B)P(A|B)$. By equating the right side of these two formulas;

$$P(A)P(B|A) = P(B)P(A|B)$$

    and this leads to the famous Bayes' Theorem,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

12. For events A, B, and C, rewrite $P(A, B, C)$ as a product of several conditional probabilities and one unconditional probability involving a single event. Your conditional probabilities can use only one event on the left side of the conditional bar. For example $P(A|C)$ and $P(A)$ would be okay but $P(A, B|C)$ is not.

    solution

    By using chain rule;

    $P(A, B, C) = P(A|B, C)P(BC)$

13. Let A be any event, and let X be a random variable defined by:

$$X = \begin{cases} 1 & if event A occurs, \\ 0 & otherwise. \end{cases}$$

X is sometimes called the indicator random variable for the event A. Show that E[X]=P(A), where E[X] is the expected value of random variable X.

solution

The expectation, $E(X)$, of a random variable X on sample space S is defined as;

$$E(X) = \sum_{s \in S} X(s).P(s)$$

this also can be defined as;

$$E(X) = \sum_{t \in range(X)} t.P(X = t)$$

since the range of random valuable X is 0 and 1, there is no need for summation from previous formula.

$$E(X) = 1.P(A = 1) + 0.P(A = 0)$$

$$E(X) = 1.P(A = 1)$$

$$E(X) = P(A)$$

14. Let X, Y, and Z be random variables taking Boolean values. The following table lists the probability of each possible assignment of 0 and 1 to the variables X, Y, and Z:

	z=0		z=1	
	x=0	x=1	x=0	x=1
y=0	1/15	1/15	4/15	2/15
y=1	1/10	1/10	8/45	4/45

For example, $P(X = 0, Y = 1, Z = 0) = 1/10$ and $P(X = 1, Y = 1, Z = 1) = 4/45$.

(a) is X independent of Y ?

(b) is X conditionally independent of Y given Z ? why or why not?

(c) Calculate $P(X = 0|X + Y > 0)$

solution

(a) Two variables are independent if; $\forall x, y : P(x, y) = P(x).P(y)$. Therefore X is not independent of Y.

(b) X is conditionally independent of Y given Z only if $\forall x, y, z : P(x, y|z) = P(x|z).P(y|z)$. Applying this formula to the given table suggests that X is conditionally independent of Y given Z.

(c) Here, Conditional probability formula can be applied to calculate $P(X = 0|X + Y > 0)$. As conditional probability formula suggests; $P(X = 0|X + Y > 0) = P(X + Y > 0 \cap X = 0)/P(X + Y > 0)$. This means to calculate the required probability, nominator is equal to probability of X being 0 and probability of X + Y is greater than zero and the denominator is equal to probability of X + Y is greater than zero. Nominator probability is sum of probabilities where x and y both equal to 0: $\frac{1}{10} + \frac{8}{45} = \frac{5}{18}$. For the denominator, probability is equal to sum of probabilities where x + y is greater than 0: $\frac{1}{10} + \frac{1}{10} + \frac{1}{15} + \frac{2}{15} + \frac{4}{45} + \frac{8}{45} = \frac{2}{3}$. Now placing calculated probabilities on formula:

$$P(X = 0|X + Y > 0) = P(X + Y > 0 \cap X = 0)/P(X + Y > 0)$$

$$\Rightarrow P(X = 0|X + Y > 0) = \frac{\frac{5}{28}}{\frac{2}{3}} = \frac{5}{12}$$

Section 2: Maximum Likelihood Estimation Maximum a Posteriori Estimation:

This problem explores two different techniques for estimating an unknown parameter of a probability distribution: the maximum likelihood estimate (MLE) and the maximum a posteriori probability (MAP) estimate. Suppose we observe the values of n iid random variables $X_1, X_2, ..., X_n$. Each X_i can take three values a, b, or c (that is, the domain of each random variable is a,b,c). Also, each X_i is drawn from the following probability distribution function:

$P(X_i = a) = \theta_1$ and $P(X_i = B) = \theta_2$

Where θ_1 and θ_2 are the parameters of our probability distribution function. Our goal is to estimate the parameter vector $\theta =< \theta_1, \theta_2 >$ from these observed values of X_1 through X_n.

Maximum Likelihood Estimation:

The first estimator of θ that we consider is the maximum likelihood estimator. For any hypothetical value $\theta =< \hat{\theta_1}, \hat{\theta_2}$, we compute the probability of observing the data/outcome $X_1, X_2, ..., X_n$ if the parameter value was $\theta \leq \hat{\theta_1}, \hat{\theta_2}$. Probability of observed data is called likelihood, and the function $L(\hat{\theta})$ that maps $\hat{\theta}$ to the corresponding likelihood is called the likelihood function. A natural way to estimate the optimal parameter vector is to compute: $\hat{\theta_{MLE}} = \arg\max_{\hat{\theta}} L(\hat{\theta})$.

1. Write a formula for the likelihood function $L(\hat{\theta})$ Your function should depend on the random variables $X_1, X_2, ..., X_n$ and the hypothetical parameter θ. Does the likelihood function depend on the order of $X_1, X_2, ..., X_n$? why?

   solution

   To formally introduce the maximum likelihood estimation method, the likelihood function should be defined first. Assume $X_1, X_2, ..., X_n$ is a random distribution with a parameter θ (Here this θ might be a vector of θ's; $\theta = (\theta_1, \theta_2, ..., \theta_k)$). Suppose that, for each discrete variable X_i in $X_1, X_2, ..., X_n$, the observed values are $x_1, x_2, ..., x_n$. The likelihood function as the probability of the observed sample as a function of θ can be interpreted as:

   $$L = (X_1, X_2, ..., X_n; \theta) = P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n; \theta)$$

   $$P_{X_1 X_2 ... X_n}(x_1, x_2, ..., x_n; \theta)$$

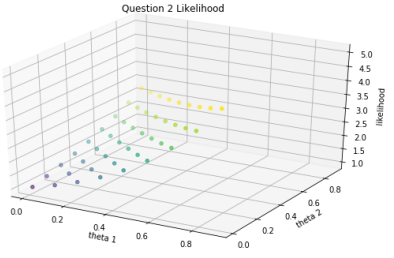
   The vector notation of $X = (X_1, X_2, ..., X_n)$ can be employed to get the formula bit more compact;

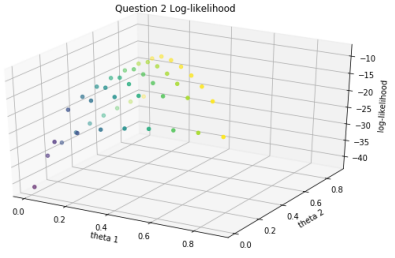
   $$L(x; \theta) = P_X(x; \theta)$$

   As can be observed from likelihood function formula, it does not depend on the order of feature vectors X_is.

2. Suppose that $n = 10$ and the dataset contains 6 number of 'a's, 3 'b's, and 1 'c'. Write a short computer program that plots the likelihood function of this data for each values of $\hat{\theta}$ where the domain of $\hat{\theta_1}$ is $p = 0, 0.01, 0.02, \ldots, 1$ and the domain of $\hat{\theta_2}$ is also p. For the plot, one X-axis should be $\hat{\theta_1}$, one Y-axis should be $\hat{\theta_2}$ and a Z axis should be the likelihood function value $L(\hat{\theta})$. Please submit the code with your homework.

   solution





3. Estimate $\hat{\theta_{MLE}}$ by marking the point for which the likelihood is largest. Can you find a closed-form solution for $\hat{\theta_{MLE}}$? Does the close-form solution agree with the plot?

solution

On the first question it was given that $\hat{\theta_{MLE}} = \arg\max_{\hat{\theta}} L(\hat{\theta})$. Taking the log of this $L(\theta)$ and maximizing it will yield the same result for the Maximum Likelihood Estimation, therefore; $\hat{\theta_{MLE}} = \arg\max_{\hat{\theta}} \ln L(\hat{\theta})$. Taking the derivative of the $L(\theta)$ with respect to each θ value then set it to zero, first;

$$\frac{\partial ln(L(\theta))}{\partial \theta_1} = \frac{\alpha_1}{\theta_1} + \frac{\alpha_3}{1-\theta_1-\theta_2} = 0$$

Solving this equation yields the result;

$$\theta_1 = \frac{\alpha_1(1-\theta_2)}{\alpha_1+\alpha_3}$$

Further, applying the derivative for θ_2 results as;

$$\frac{\partial ln(L(\theta))}{\partial \theta_2} = \frac{\alpha_2}{\theta_2} + \frac{\alpha_3}{1-\theta_1-\theta_2} = 0$$

Taking these two formulas and replacing the θ_1 value with the resulting calculation on the second equation;

$$\frac{\alpha_2}{\theta_2} = (1-\theta_2-\frac{\alpha_1(1-\theta_2)}{\alpha_1+\alpha_3})^{-1}$$

$$\theta_2 = \frac{\alpha_2}{\alpha_1+\alpha_2+\alpha_3}$$

Using the above formula for θ_2 on the previous formula for θ_1 results with;

$$\theta_1 = \frac{\alpha_1(1-\frac{\alpha_1}{\alpha_1+\alpha_2+\alpha_3})}{\alpha_1+\alpha_3}$$

$$\theta_1 = \frac{\alpha_1}{\alpha_1+\alpha_2+\alpha_3}$$

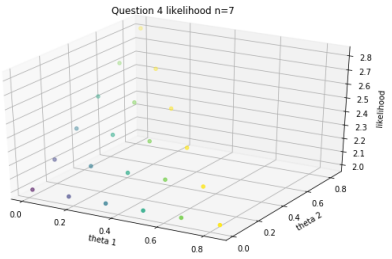
Since both θ_1 and θ_2 are calculated, last observation can be interpreted as;

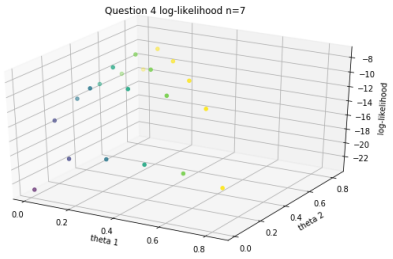
$$1-\theta_1-\theta_2 = 1-\frac{\alpha_1}{\alpha_1+\alpha_2+\alpha_3}-\frac{\alpha_2}{\alpha_1+\alpha_2+\alpha_3} = \frac{\alpha_3}{\alpha_1+\alpha_2+\alpha_3}$$

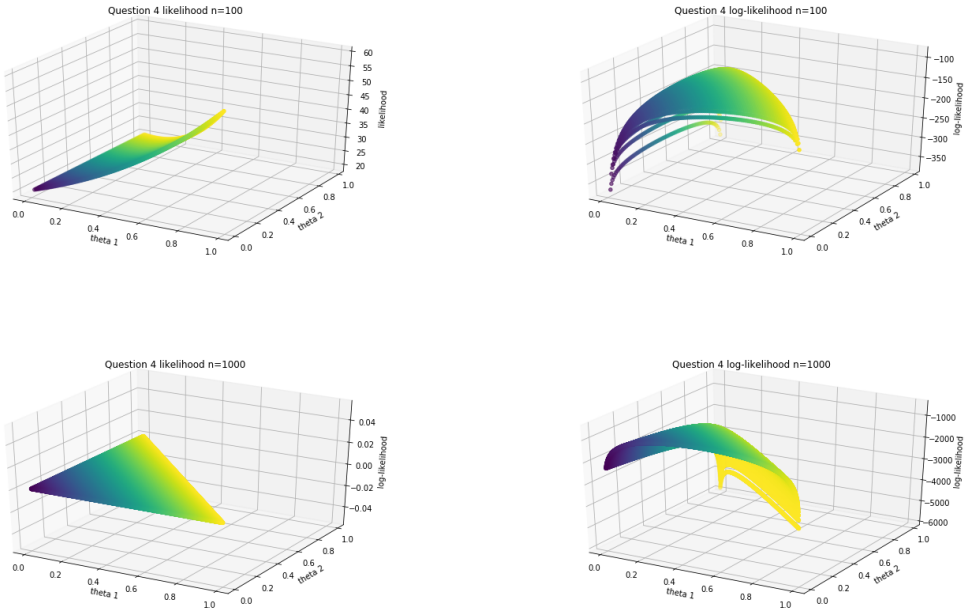
4. Create 3 more likelihood plots:

- one where we have n=7 and we have 2 'a', 3 'b', 2 'c' values.
- one where we have n=100 and we have 60 'a', 20 'b', 20 'c' values.
- one where we have n=1000 and we have 500 'a', 0 'b', 500 'c' values.

solution







5. Can you describe how the likelihood function and maximum likelihood parameter change in the three datasets of previous question ?

   solution

   In the first example of Question 4, the likelihood function has different values throughout the distribution of different θ values. In the second example the shape becomes a little more narrow but shows a significant increase on likelihood values. Last example shows that when θ_2 value is equal to 0, there would be no change on likelihood, it's only the value of θ_1. The second example shows the easiest likelihood calculation, owing to the fact that the third example is insignificant and the first example shows a lot closer likelihood values.

6. In maximum likelihood estimation, we had $\hat{\theta_{MLE}} = \arg\max_{\hat{\theta}} L(\hat{\theta})$. Can you write the formula for maximum a posteriori estimation? That is, what is $\hat{\theta_{MAP}}$?

7. As we discussed in the class, a conjugate prior helps us have a simple formula for the maximum a posteriori estimation. For example, in Coin Flip example we saw that the Bernoulli prior distribution was very simple. Now, in this problem (our homework), can you propose a prior distribution formula (i.e., $P(\theta)$)?

8. Please compute the optimal $\hat{\theta_{MAP}}$

   solution for 6, 7 and 8. questions

   For each $\hat{\theta}$ the maximum a posteriori estimation can be calculated with the respected α and β values:

$$\hat{\theta}^1_{MAP} = \frac{\alpha_1 + \beta_1 - 1}{\sum_{i=1}^{3}(\alpha_j + \beta_j - 1)}$$

$$\hat{\theta}^2_{MAP} = \frac{\alpha_2 + \beta_2 - 1}{\sum_{i=1}^{3}(\alpha_j + \beta_j - 1)}$$

   and therefore θ_3 can be calculated as;

$$\hat{\theta}^3_{MAP} = 1 - \hat{\theta}^1_{MAP} - \hat{\theta}^2_{MAP}$$

   We can calculate the $\hat{\theta_{MAP}}$ by using a distribution "D", for this homework the Drichlet distribution can be utilized. The maximum a posteriori;

$$\hat{\theta_{MAP}} = \arg\max_{\theta} P(\theta|D)$$

We know that we can interpret the $P(\theta|D)$ as;

$$\hat{\theta_{MAP}} = \arg\max_{\theta} P(D|\theta)P(\theta)$$

here the $P(D|\theta)$ is the likelihood and $P(\theta)$ is the prior. This prior is defined as;

$$P(\theta) = \frac{\theta_1^{\beta_1-1}\theta_2^{\beta_2-1}\theta_3^{\beta_3-1}}{\beta(\theta_1,\theta_2,\theta_3)} = Drichlet(\beta_1,\beta_2,\beta_3)$$

9. Please create the 3 plots you had in question 4 for $\hat{\theta_{MAP}}$ using your proposed prior distribution.
solution