Machine Learning: Homework 2

Skye Chen zc969@nyu.edu

Due 11:59 p.m. Friday, March 6, 2020

Instructions

- **Collaboration policy:** Homeworks must be done individually, except where otherwise noted in the assignments. "Individually" means each student must hand in their own answers, and each student must write and use their own code in the programming parts of the assignment. It is acceptable for students to collaborate in figuring out answers and to help each other solve the problems, though you must in the end write up your own solutions individually, and you must list the names of students you discussed this with. We will be assuming that, as participants in an undergraduate course, you will be taking the responsibility to make sure you personally understand the solution to any work arising from such collaboration.
- **Online submission:** You must submit your solutions online on NYU Classes. We recommend that you use LaTeX, but we will accept scanned / pictured solutions as well.

Problem 1: More Probability Review

(a) **[5 Points]** For events A and B, prove

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

Solution:

Proof.

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

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

□

(b) **[5 Points]** 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 conditioning bar. For example, $P(A|C)$ and $P(A)$ would be okay, but $P(A, B|C)$ is not.
Solution:

Proof.

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

□

(c) **[5 Points]** Let A be any event, and let X be a random variable defined by

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

X is sometimes called the indicator random variable for the event A. Show that $\mathbb{E}[X] = P(A)$, where $\mathbb{E}[X]$ denotes the *expected value* of X.
Solution:

Proof.

$$\mathbb{E}[X] = 1 \times P(A) + 0 \times (1 - P(A)) = P(A).$$

□

(d) Let X, Y, and Z be random variables taking values in $\{0, 1\}$. 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 $P(X = 1, Y = 1, Z = 1) = 4/45$.

(i) **[5 Points]** Is X independent of Y? Why or why not?
Solution: X is not independent of Y, since $P(X = 0) = \frac{11}{18}$, $P(Y = 1) = \frac{7}{15}$, and $P(X = 0, Y = 1) = \frac{5}{18}$, which indicates that $P(X = 0, Y = 1) \neq P(X = 0)P(Y = 1)$. Therefore, they are not independent.

(ii) **[5 Points]** Is X conditionally independent of Y given Z? Why or why not?
Solution: X is conditionally independent of Y given Z, i.e. $P(X = i|Y = j, Z = k) = P(X = i|Z = k)$.
From table, $P(X = 1|Y = 0, Z = 1) = P(X = 1|Y = 1, Z = 1) = P(X = 1|Z = 1) = \frac{1}{3}$; $P(X = 1|Y = 0, Z = 0) = P(X = 1|Y = 1, Z = 0) = P(X = 1|Z = 0) = \frac{1}{2}$; $P(X = 0|Y = 0, Z = 1) = P(X = 0|Y = 1, Z = 1) = P(X = 0|Z = 1) = \frac{2}{3}$; $P(X = 0|Y = 0, Z = 0) = P(X = 0|Y = 1, Z = 0) = P(X = 0|Z = 0) = \frac{1}{2}$. Thus, for every $i, j, k \in \{0, 1\}$, we have $P(X = i|Y = j, Z = k) = P(X = i|Z = k)$.We can conclude that X is conditionally independent of Y given Z.

(iii) **[5 Points]** Calculate $P(X = 0|X + Y > 0)$.
Solution:

$$P(X = 0|X+Y > 0) = \frac{P(X = 0, X + Y > 0)}{P(X + Y > 0)} = \frac{P(X = 0, Y = 1)}{P(X = 0, Y = 1) + P(X = 1, Y = 0) + P(X = 1, Y = 1)} = \frac{5}{12}.$$

Problem 2: Maximum Likelihood and 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[1] random variables $X_1, \ldots, X_n$ drawn from a single Bernoulli distribution with parameter θ. In other words, for each X_i, we know that

$$P(X_i = 1) = \theta \quad \text{and} \quad P(X_i = 0) = 1 - \theta.$$

Our goal is to estimate the value of θ from these observed values of X_1 through X_n.

[1] iid means Independent, Identically Distributed.

Maximum Likelihood Estimation

The first estimator of θ that we consider is the maximum likelihood estimator. For any hypothetical value $\hat{\theta}$, we can compute the probability of observing the outcome $X_1, \ldots, X_n$ if the true parameter value θ were equal to $\hat{\theta}$. This probability of the observed data is often called the *data likelihood*, and the function $L(\hat{\theta})$ that maps each $\hat{\theta}$ to the corresponding likelihood is called the *likelihood function*. A natural way to estimate the unknown parameter θ is to choose the $\hat{\theta}$ that maximizes the likelihood function. Formally,

$$\hat{\theta}^{\text{MLE}} = \underset{\hat{\theta}}{\text{argmax}}\, L(\hat{\theta}).$$

(a) **[5 Points]** Write a formula for the likelihood function, $L(\hat{\theta})$. Your function should depend on the random variables $X_1, \ldots, X_n$ and the hypothetical parameter $\hat{\theta}$. Does the likelihood function depend on the order of the random variables?
Solution:
Since $X_i \sim \text{Bernoulli}(\theta)$, the likelihood function can be represented as:

$$L(\hat{\theta}) = \prod_{i=1}^{n} \theta^{X_i}(1-\theta)^{1-X_i}.$$

The likelihood function does not depend on the order by the commutative property of product. Alternate way of representing likelihood function: Assume that among n tests, there are α_1 tests with result 1 and α_0 tests with result 0, then the likelihood function can be rewrite as:

$$L(\hat{\theta}) = \theta^{\alpha_1}(1-\theta)^{\alpha_0}.$$

(b) **[5 Points]** Suppose that $n = 10$ and the data set contains six 1s and four 0s. Write a short computer program that plots the likelihood function of this data for each value of $\hat{\theta}$ in $\{0, 0.01, 0.02, \ldots, 1.0\}$. For the plot, the x-axis should be $\hat{\theta}$ and the y-axis $L(\hat{\theta})$. Scale your y-axis so that you can see some variation in its value. Please submit both the plot and the code that made it. Please include all plots for this question in the `problem2.pdf` file, as well as the source code for producing them. That is, do not submit the source code and plots as separate files.
Solution:

```
#First create a random list with 6 1s and 4 0s.
n = 10
X = [0] * 4 + [1] * 6
shuffle(X)
# create list theta
theta = [0.01*i for i in range(100)]
#calculate likelihood function for each theta
likelihood_list = []
for i in range(len(theta)):
    likelihood = 1
    for j in range(n):
        likelihood = likelihood * theta[i]**(int(X[j]))*(1-theta[i])**(1-int(X[j]))
    likelihood_list.append(likelihood)
#plot
plt.plot(theta, likelihood_list)
plt.xlabel("$\hat{ \\theta} $")
plt.ylabel("Value of likelihood function")
```

Question2-(b)

The plot is presented as Figure 1.

(c) **[5 Points]** Estimate $\hat{\theta}^{\text{MLE}}$ by marking on the x-axis the value of $\hat{\theta}$ that maximizes the likelihood. Find a closed-form formula for the MLE. Does the closed form agree with the plot?
Solution: Using code below, get $\hat{\theta} = 0.6$.

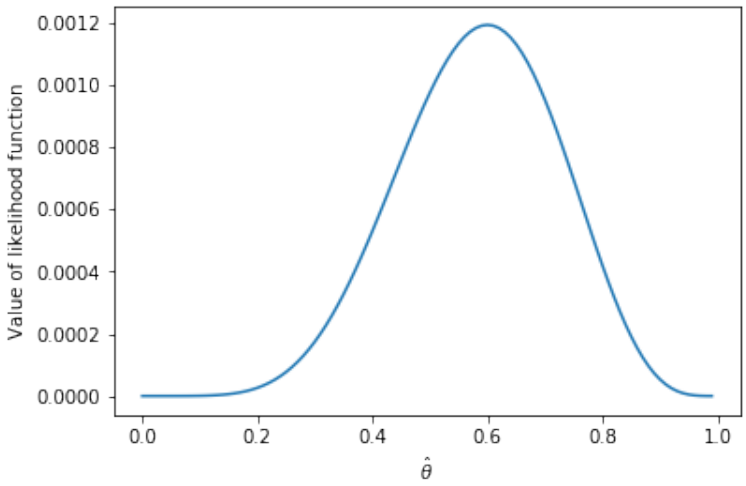

Figure 1: Plot for θ and value of likelihood function for question 2-(b)

```
likelihood_array = np.asarray(likelihood_list)
theta_hat_index = (-likelihood_array).argsort()[:1][0]
theta_hat = theta[theta_hat_index]
print('The value of \\theta_hat that maximizes the likelihood is', theta_hat)
#The value of \theta_hat that maximizes the likelihood is 0.6
```

Question2-(c)

By plugging in $\hat{\theta} = 0.6$, the formula of likelihood function is $\prod_{i=1}^{n} 0.6^{X_i} 0.4^{1-X_i}$. If we want to derive t=$\hat{\theta}$ mathematically, we first take the log of the likelihood function, then we set the derivative of the log-likelihood being 0. Assume that among n tests, there are α_1 tests with result 1 and α_0 tests with result 0, then the likelihood function can be rewrite as:

$$L(\hat{\theta}) = \theta^{\alpha_1}(1-\theta)^{\alpha_0}$$

$$\hat{\theta}^{\text{MLE}} = \underset{\hat{\theta}}{\text{argmax}} \ln L(\hat{\theta}) = \underset{\hat{\theta}}{\text{argmax}} \ln \theta^{\alpha_1}(1-\theta)^{\alpha_0}.$$

$$\text{Set derivative to 0: } \frac{d}{d\theta} \ln \theta^{\alpha_1}(1-\theta)^{\alpha_0} = 0$$

$$0 = \alpha_1 \frac{1}{\hat{\theta}} - \frac{\alpha_0}{1-\hat{\theta}} \rightarrow \hat{\theta} = \frac{\alpha_1}{\alpha_1 + \alpha 0}.$$

In our test, $\alpha_1 = 6$ and $\alpha_0 = 4$, which gives $\hat{\theta} = \frac{6}{6+4} = 0.6$. Thus proved that it agrees with the plot.

Maximum a Posteriori Probability Estimation

In the maximum likelihood estimate, we treated the true parameter value θ as a fixed (non-random) number. In cases where we have some prior knowledge about θ, it is useful to treat θ itself as a random variable, and express our prior knowledge in the form of a prior probability distribution over θ. For example, suppose that the $X_1, \ldots, X_n$ are generated in the following way:

- First, the value of θ is drawn from a given prior probability distribution
- Second, $X_1, \ldots, X_n$ are drawn independently from a Bernoulli distribution using this value for θ.

Since both θ and the sequence $X_1, \ldots, X_n$ are random, they have a joint probability distribution. In this setting, a natural way to estimate the value of θ is to simply choose its most probable value given its prior distribution plus the observed data $X_1, \ldots, X_n$.

$$\hat{\theta}^{\text{MAP}} = \underset{\hat{\theta}}{\text{argmax}}\, P(\theta = \hat{\theta} | X_1, \ldots, X_n).$$

This is called the maximum a posteriori probability (MAP) estimate of θ. Using Bayes rule, we can rewrite the posterior probability as follows:

$$P(\theta = \hat{\theta} | X_1, \ldots, X_n) = \frac{P(X_1, \ldots, X_n | \theta = \hat{\theta}) P(\theta = \hat{\theta})}{P(X_1, \ldots, X_n)}.$$

Since the probability in the denominator does not depend on $\hat{\theta}$, the MAP estimate is given by

$$\begin{aligned}\hat{\theta}^{\text{MAP}} &= \underset{\hat{\theta}}{\text{argmax}}\, P(X_1, \ldots, X_n | \theta = \hat{\theta}) P(\theta = \hat{\theta}) \\ &= \underset{\hat{\theta}}{\text{argmax}}\, L(\hat{\theta}) P(\theta = \hat{\theta}).\end{aligned}$$

In words, the MAP estimate for θ is the value $\hat{\theta}$ that maximizes the likelihood function multiplied by the prior distribution on θ. When the prior on θ is a continuous distribution with density function p, then the MAP estimate for θ is given by

$$\hat{\theta}^{\text{MAP}} = \underset{\hat{\theta}}{\text{argmax}}\, L(\hat{\theta}) p(\hat{\theta}).$$

For this problem, we will use a Beta(3,3) prior distribution for θ, which has density function given by

$$p(\hat{\theta}) = \frac{\hat{\theta}^2 (1 - \hat{\theta})^2}{B(3, 3)},$$

where $B(\alpha, \beta)$ is the beta function and $B(3, 3) \approx 0.0333$.

(d) **[5 Points]** Suppose, as in part (c), that $n = 10$ and we observed six 1s and four 0s. Write a short computer program that plots the function $\hat{\theta} \mapsto L(\hat{\theta}) p(\hat{\theta})$ for the same values of $\hat{\theta}$ as in part (c).
Solution:
Using the code below:

```
#function to calculate p(\theta)
def p_theta(theta):
    return (theta**2)*((1-theta)**2)/0.033333
#calculate likelihood function for each theta
likelihood_list = []
alpha_1 = sum(X)
alpha_0 = n-sum(X)
for i in range(len(theta)):
    likelihood = theta[i]**alpha_1
    likelihood = likelihood * (1-theta[i])**alpha_0
    likelihood = likelihood *  p_theta(theta[i])
    likelihood_list.append(likelihood)
#plot
plt.plot(theta, likelihood_list)
plt.xlabel("$\hat{ \\theta} $")
plt.ylabel("Value of MAP likelihood function")
```

Question2-(b)

The plot is presented as Figure 2.

(e) **[5 Points]** Estimate $\hat{\theta}^{\text{MAP}}$ by marking on the x-axis the value of $\hat{\theta}$ that maximizes the function. Find a closed form formula for the MAP estimate. Does the closed form agree with the plot? **Solution:**
Using code below, get $\hat{\theta} = 0.57$.

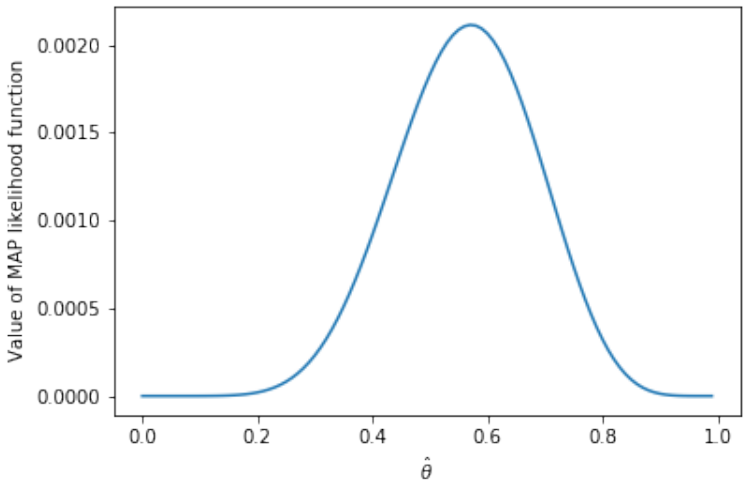

Figure 2: Plot for θ and value of MAP likelihood function for question 2-(d)

```
likelihood_array = np.asarray(likelihood_list)
theta_hat_index = (-likelihood_array).argsort()[:1][0]
theta_hat = theta[theta_hat_index]
print('The value of \\theta_hat that maximizes the likelihood is', theta_hat)
# The value of \theta_hat that maximizes the likelihood is 0.5700000000000001
```

Question2-(d)

If we want to derive mathematically, using similar approach, we set 0 to the derivative of its log-likelihood function:

$$\hat{\theta}^{\text{MAP}} = \underset{\hat{\theta}}{\text{argmax}} \ln L(\hat{\theta}) p(\hat{\theta}) = \underset{\hat{\theta}}{\text{argmax}} \ln \theta^{\alpha_1} (1-\theta)^{\alpha_0} p(\hat{\theta}).$$

$$\text{Set derivative to 0: } \frac{d}{d\theta} \ln \theta^{\alpha_1+2} (1-\theta)^{\alpha_0+2} - \ln \frac{1}{B(3,3)} = 0$$

$$0 = (\alpha_1+2)\frac{1}{\hat{\theta}} - \frac{(\alpha_0+2)}{1-\hat{\theta}} \rightarrow \hat{\theta} = \frac{\alpha_1+2}{\alpha_1+\alpha_0+4}.$$

In our test, $\alpha_1 = 6$ and $\alpha_0 = 4$, which gives $\hat{\theta} = \frac{6+2}{6+4+4} = 0.57$. Thus proved that it agrees with the plot.

(f) **[5 Points]** Compare the MAP estimate to the MLE computed from the same data in part (c). Briefly explain any significant difference.
Solution:
For MLE, θ is assumed to be a deterministic number(not-random); however for MAP, we assume that θ is a random variable formed by a prior probability distribution. Therefore, the probability distribution of θ, i.e. $p(\theta)$, will play a role in the likelihood function and influence the result when we maximize the likelihood function. The difference can also be seen after we solved the MLE and MAP mathematically in question 2c and 2e. In this example, MLE gives $\hat{\theta} = 0.6$, whereas in MAP, $\hat{\theta} = 0.57$. When sample number is small, hallucinating terms is significant.

(g) **[5 Points]** Comment on the relationship between the MAP and MLE estimates as n goes to infinity.
Solution:
As n goes to infinity, MAP will converge to MLE since the constant(hallucinating terms) in the formula of MAP will not contribute much as n being larger.