ITU Computer and Informatics Faculty
BLG 454E Learning From Data, Spring 2018
Homework #1
Due March 07, 2018 10pm

HOMEWORK #1 SOLUTIONS

Important Note: If you see ANY mistakes please inform me via kivrakh@itu.edu.tr

1. (10 pts.) In general, the probability that it rains on Saturday is 25%.

   Weekend rain has the following relationships:

   - If it rains on Saturday, the probability that it rains on Sunday is 50%.
   - If it does not rain on Saturday, the probability that it rains on Sunday is 25%.

   Given that it rained on Sunday, what is the probability that it rained on Saturday?

   Correct answer: 40%

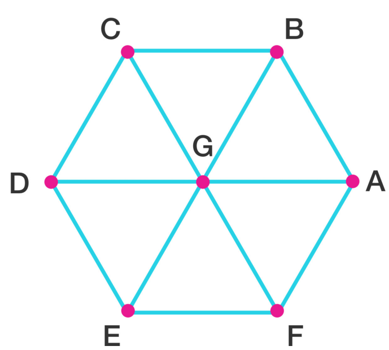
   Let T and U be the events of rain on Saturday and Sunday, respectively, and denote the event of no rain on Saturday as $\overline{T}$. The problem gives us:

   $P(T) = 25\%$ $P(U|T) = 50\%$ $P(U|\overline{T}) = 25\%$.

   Via Bayes' Theorem:

$$P(T|U) = \frac{P(T) \cdot P(U|T)}{P(U)} = \frac{P(T) \cdot P(U|T)}{P(T) \cdot P(U|T) + P(\overline{T}) \cdot P(U|\overline{T})} = \frac{25\% \cdot 50\%}{25\% \cdot 50\% + 75\% \cdot 25\%} = 40\%.$$

2. (20 pts.) A bug stands on a random point of the lattice below. Each point is equally likely to be the starting point.



   Every minute, the bug selects an adjacent point at random and moves to it. Each adjacent point is equally likely to be chosen. For example, if the bug is on point B, then each probability to move to the points A, C, or G is $\frac{1}{3}$.

   What is the probability that the bug reaches point A in 2 moves or less? *Each point is equally likely to be the bug's starting point. Also, assume starting at A will "reach" the point in 0 moves.*

   Correct answer:

$$\frac{22}{63}$$

Let A_2 be the event that bug reaches point A in two moves or less.

Let A, B, C, D, E, F, and G be the events that the bug starts on each point, respectively. The probability of each of these events is $\frac{1}{7}$.

If the bug starts on A, then it doesn't have to move:

$$P(A_2 \mid A) = 1.$$

If the bug starts on B, then it can get to A directly, or through G :

$$P(A_2 \mid B) = \frac{1}{3} + \left(\frac{1}{3} \times \frac{1}{6}\right) = \frac{7}{18}.$$

Due to the symmetry of the lattice, the probability to get to A is the same starting from point F :

$$P(A_2 \mid F) = P(A_2 \mid B) = \frac{7}{18}.$$

If the bug starts on C, then it can get to A through B or through G :

$$P(A_2 \mid C) = \left(\frac{1}{3} \times \frac{1}{3}\right) + \left(\frac{1}{3} \times \frac{1}{6}\right) = \frac{1}{6}.$$

Due to the symmetry of the lattice, the probability to get to A is the same starting from point E :

$$P(A_2 \mid E) = P(A_2 \mid C) = \frac{1}{6}.$$

If the bug starts at D, then it can only go through point G :

$$P(A_2 \mid D) = \frac{1}{3} \times \frac{1}{6} = \frac{1}{18}.$$

If the bug starts at G, it can go to A directly, or it can go through B or F :

$$P(A_2 \mid G) = \frac{1}{6} + 2\left(\frac{1}{6} \times \frac{1}{3}\right) = \frac{5}{18}.$$

$P(A_2)$ is the sum of all these probabilities multiplied by $\frac{1}{7}$:

$$P(A_2) = \frac{1}{7}\left(1 + 2\left(\frac{7}{18}\right) + 2\left(\frac{1}{6}\right) + \frac{1}{18} + \frac{5}{18}\right) \tag{1}$$

$$\tag{2}$$

$$= \boxed{\frac{22}{63}}. \tag{3}$$

3. (40 pts.) The idea for the **maximum likelihood estimate (MLE)** is to find the value of the parameter(s) for which the data has the highest probability. You are going to do this with the densities. Suppose the 1-dimension data points $x_1, x_2, ...x_n$ given in "*data.txt*" file are drawn from a normal(gauss) $N(\mu, \sigma^2)$ distribution, where μ and σ are unknown.

- (20 pts.) Formulate the likelihood function and derive the equation to find the maximum likelihood estimate for the pair (μ, σ^2).

- (20 pts.) Implement (write the code) MLE in **Matlab** or **Python** language and provide your plot that is similar to Figure 1. You are not allowed use any built-in functions except **histogram** functions to provide you a quick of the distribution of the data.

(a) Gaussian probability density function

$$P(x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x_i - \mu)^2\right], \qquad -\infty < x_i < \infty$$

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

Given N data samples, the likelihood is

$$\mathcal{L}(\mu, \sigma) = P(x_{1:N}|\mu, \sigma) = \prod_{i=1}^{N} P(x_i|\mu, \sigma) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x_i - \mu)^2\right]$$

and the log-likelihood is

$$\log \mathcal{L}(\mu, \sigma) = \log \frac{1}{\sqrt{2\pi\sigma^2}} + \left(-\frac{1}{2\sigma^2}\right) \sum_{i=1}^{N} (x_i - \mu)^2$$

In order to find estimate $\hat{\mu}$, we need to take derivative of $\log \mathcal{L}(\mu, \sigma)$ with respect to μ and equate it to zero.

$$\frac{\partial}{\partial \mu} \log \mathcal{L}(\mu, \sigma) = \frac{\partial}{\partial \mu} \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi\sigma^2}} + \frac{\partial}{\partial \mu} \left(-\frac{1}{2\sigma^2}\right) \sum_{i=1}^{N} (x_i - \mu)^2 = 0$$

$$= \frac{1}{\sigma^2} \left[\sum_{i=1}^{N} (x_i - \mu)\right] = 0$$

$$= \sum_{i=1}^{N} x_i - \sum_{i=1}^{N} \mu = 0$$

Therefore, the estimate of μ and the mean value of the given weights is

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

For the estimate $\hat{\sigma^2}$, we need to take derivative of $\log \mathcal{L}(\mu, \sigma)$ with respect to σ^2 and equate it to zero. In order to avoid possible mistakes, θ will be used instead of σ^2 throughout the equations.

$$\frac{\partial}{\partial \theta} \log \mathcal{L}(\mu, \theta) = \frac{\partial}{\partial \theta} \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi\theta}} + \frac{\partial}{\partial \theta} \left(-\frac{1}{2\theta}\right) \sum_{i=1}^{N} (x_i - \mu)^2 = 0$$

$$= -\frac{N}{2\theta} + \frac{1}{2\theta^2} \left[\sum_{i=1}^{N} (x_i - \mu)^2\right] = 0$$

$$= -N\theta + \sum_{i=1}^{N} (x_i - \mu)^2 = 0$$

Thus, the estimate of θ or σ^2 is

$$\hat{\theta} = \hat{\sigma^2} = \frac{\sum_{i=1}^{N} (x_i - \hat{\mu})^2}{N}$$

(b)

```
import numpy as np
import matplotlib.pyplot as plt
from numpy import loadtxt

data = loadtxt("data.txt")
mu = sum(data)/len(data)
variance = sum(np.power(data - mu,2))/len(data)

# Plot the histogram.
plt.hist(data, bins=25, normed=True, alpha=0.6, color='g', label="data")

# Plot the PDF.
xmin, xmax = plt.xlim()
x = np.linspace(xmin, xmax, 885)

num = np.exp(-(x-mu)**2/(2*variance))
denom = np.sqrt(2*np.pi*variance)
p= num/denom

plt.plot(x, p, 'k', linewidth=2, label="MLE fixed distribution")
title = "MLE results: mu = %.2f,  std = %.2f" % (mu, np.sqrt(variance))
plt.title(title)
plt.legend(bbox_to_anchor=(0.65, 0.8), loc=2, borderaxespad=0.)
plt.show()
```

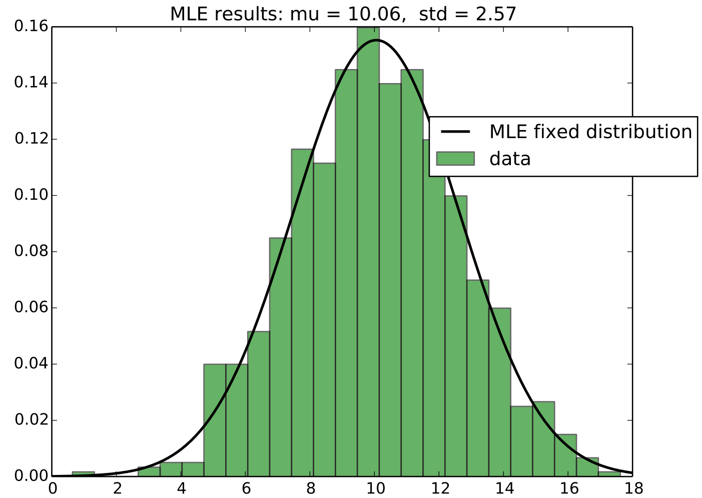


Figure 1: Data and fixed gaussian distribution with MLE

4. (30 pts.) In the Table 1 below, $\boldsymbol{x_1}$, $\boldsymbol{x_2}$, $\boldsymbol{x_3}$ and $\boldsymbol{x_i} \in \{0,1\}$, $i = 1,2,3$. $\boldsymbol{x_i}$ represent the i feature vector and $\boldsymbol{y} \in \{+,-\}$ represents the class label.

   (a) (15 pts.) Construct the Naive Bayes classifier for the given training dataset in Table 1. **Hint:** Estimate the class conditional prob. for each feature vector $\boldsymbol{x_1}$, $\boldsymbol{x_2}$, $\boldsymbol{x_3}$

   (b) (5 pts.) Predict the class label for $(x_1 = 1, x_2 = 1, x_3 = 1)$ data using trained Naive Bayes approach in part (a)

   (c) (10 pts.) Calculate the probabilities of $P(x_1 = 1)$, $P(x_2 = 1)$, and $P(x_1 = 1, x_2 = 1)$. Decide whether x_1 and x_2 are independent or not.

(a) $P(x_1 = 1|+) = 3/5 = 0.6, P(x_2 = 1|+) = 2/5 = 0.4, P(x_3 = 1|+) = 4/5 = 0.8,$
$P(x_1 = 0|+) = 2/5 = 0.4, P(x_2 = 0|+) = 3/5 = 0.6, P(x_3 = 0|+) = 1/5 = 0.2,$

$P(x_1 = 1|-) = 2/5 = 0.4, P(x_2 = 1|-) = 2/5 = 0.4, P(x_3 = 1|-) = 1/5 = 0.2.$
$P(x_1 = 0|-) = 3/5 = 0.6, P(x_2 = 0|-) = 3/5 = 0.6, P(x_3 = 0|-) = 4/5 = 0.8.$

(b) Let $R : (x_1 = 1, x_2 = 1, x_3 = 1)$ be the test instance. To determine its class, we need to compute $P(+|R)$ and $P(-|R)$. Using Bayes theorem:

$P(+|R) = \frac{P(R|+)P(+)}{P(R)}$ and $P(-|R) = \frac{P(R|-)P(-)}{P(R)}$.

Since $P(+) = P(-) = 5/10 = 0.5$ and $P(R)$ is the same for both class, R can be classified by comparing $P(R|+)$ and $P(R|-)$.

Naive bayes assumes features are independent, so we can write,

$P(R|+) = P(x_1, x_2, x_3|+) = P(x_1 = 1|+) \times P(x_2 = 1|+) \times P(x_3 = 1|+) = 0.192$
$P(R|-) = P(x_1 = 1|-) \times P(x_2 = 1|-) \times P(x_3 = 1|-) = 0.032$

Since $P(R|+)$ is larger, the record is assigned to (+) class.

(c) $P(x_1 = 1) = 5/10 = 0.5, P(x_2 = 1) = 4/10 = 0.4$ and $P(x_1 = 1, x_2 = 1) = P(x_1) \times P(x_2) = 0.2$. Therefore, x_1 and x_2 are independent.

Table 1: Training set for question 4

Instance	x_1	x_2	x_3	**y**
1	0	0	1	-
2	1	0	1	+
3	0	1	0	-
4	1	0	0	-
5	1	0	1	+
6	0	0	1	+
7	1	1	0	-
8	0	0	0	-
9	0	1	0	+
10	1	1	1	+