DSCI 552 – MACHINE LEARNING FOR DATA SCIENCE
HOMEWORK – 3

1) Given the following statistics, what is the probability that a person has a particular disease in a town if he/she has tested positive from a home testing kit

- 2% percent of the population in the town has the disease
- 80% of those who have the disease test positive on the home kit
- 10% of those who use the kit will have false positives.

ANS:

Aim:

To find the probability that a person has a particular disease given that they have tested positive using a home testing kit,

$$P(Disease \mid Positive)$$

$P(Disease \mid Positive)$ is the probability of having the disease given a positive test result.

Given:

$P(Disease)$ is the probability of having the disease

$$P(Disease) = 2\% = 0.02$$

$P(Positive \mid Disease)$ is the probability of testing positive given that the person has the disease

$$P(Positive \mid Disease) = 80\% = 0.80$$

$P(Positive \mid No\ Disease)$ is the probability of testing positive but the person doesn't actually have the disease (False Positive).

$$P(Positive \mid No\ Disease) = 10\% = 0.10$$

$P(No\ Disease)$ is the probability of person not having the disease. As this metric is not available directly, we can subtract $1 -$ probability of person having a disease i.e., $P(Disease)$

$$P(No\ Disease) = 1 - P(Disease) = 1 - \ 0.02 = 0.98\ (98\%)$$

$P(Positive)$ is the probability of testing positive, which is the sum of true positives and false positives.

$$P(Positive) =$$

$$P(Positive \mid Disease) \times P(Disease)\ +\ P(Positive \mid No\ Disease) \times P(No\ Disease)$$

(True Positives + False Positives = Total Probability of testing positive)

=> $P(Positive) = 0.80 \times 0.02 + 0.10 \times 0.98 = 0.114$

we can use Bayes' Theorem to find if the person has the diseases while he/she is tested positive using a home kit. As per the Bayes Theorem

$$P(Disease \mid Positive) = \frac{P(Positive \mid Disease) \times P(Disease)}{P(Positive)}$$

=> $P(Disease \mid Positive) = \frac{0.80\ \times 0.02}{0.114}$

=> $P(Disease \mid Positive) = \ 0.1404 \approx 14.04\%$

Hence, the probability that a person has a particular disease given that they have tested positive using a home testing kit is 14.04%.

2) In this problem we will perform Maximum Likelihood Estimation to find the parameters of a Gaussian Distribution. Consider the data distribution of n one dimensional points. Let them be denoted by the variable X. Then, if we assume they come from a Gaussian Distribution with mean μ and Variance V, X comes from the probability distribution:

$$\mathrm{P}(\mathrm{x} \mid \mu\ ,\ \mathrm{V}) = \frac{1}{\sqrt{2\pi V}} e^{-\frac{(x-\mu)^2}{2V}}$$

Apply MLE on the above equation by using the following hints.

a) The probability values of the Gaussian Distribution over X is given by

$$P(X \mid \mu, V) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi V}} e^{-\frac{(xi-\mu)^2}{2V}}$$

We need to maximize this to find the values of μ and **V**. That is done by take the partial derivative of this equation with respect to μ and **V** separately, setting it to 0 and solving for the values

b) Minimizing the log of a function is the same as maximizing the function itself. Take the log of the equation to minimize it.

<u>HINTS:</u>

b) Derivative of log(x) is 1/x

c) Derivative of f(g(x)) is f'(g(x)).g'(x)

d) log (ab) = log a + log b

e) log (e^x) = x

f) log (a^b) = b log a

<u>ANS:</u>

$$P(X \mid \mu, V) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi V}} e^{-\frac{(x_i - \mu)^2}{2V}}$$

Applying Log on both sides

$$=> log\ (P(X \mid \mu, V)) = log(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi V}} e^{-\frac{(x_i - \mu)^2}{2V}})$$

by using hint (d)

$$=> log\ (P(X \mid \mu, V)) = n\ log(\frac{1}{\sqrt{2\pi V}}) + \sum_{i=1}^{n} (log\ (e^{-\frac{(x_i - \mu)^2}{2V}})$$

by using hint (e)

$$=> log(P(X \mid \mu, V)) = \; n \, log(2\pi V)^{-\frac{1}{2}}) - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2V}$$

by using hint (f)

$$=> log\;(P(X \mid \mu, V)) = -\frac{n}{2}\; log\;(2\pi V) - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2V}$$

$$\boxed{log\;(P(X \mid \mu, V)) = -\frac{n}{2} log\;(2\pi V) - \frac{1}{2V}\sum_{i=1}^{n}(x_i - \mu)^2}$$

Derivation for μ:

$$\frac{\partial}{\partial \mu} log\;(P(X \mid \mu, V)) = \frac{\partial}{\partial \mu}(-\frac{n}{2} log\;(2\pi V) - \frac{1}{2V}\sum_{i=1}^{n}(x_i - \mu)^2)$$

Equating $\frac{\partial}{\partial \mu} log\;(P(X \mid \mu, V))$ to 0

$$=> 0 = \sum_{i=1}^{n}(x_i - \mu)$$

$$=> n\mu = \sum_{i=1}^{n} x_i$$

$$\boxed{=> \; \mu = \frac{1}{n}\sum_{i=1}^{n} x_i}$$

Derivation for V:

$$\frac{\partial}{\partial V} log\;(P(X \mid \mu, V)) = \frac{\partial}{\partial V}(-\frac{n}{2} log\;(2\pi V) - \frac{1}{2V}\sum_{i=1}^{n}(x_i - \mu)^2)$$

Equating $\frac{\partial}{\partial V} log\;(P(X \mid \mu, V))$ to 0

by using hints (b & c)

$$=> 0 = -\frac{n}{2V} + \frac{1}{2V^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

$$\boxed{=> \; \frac{n}{2V} = \frac{1}{2V^2}\sum_{i=1}^{n}(x_i - \mu)^2}$$

$$=> nV = \sum_{i=1}^{n} (x_i - \mu)^2$$

$$=> V = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2$$

********************************* **THE END** ***********************************