

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,

$$P(\text{Disease} \mid \text{Positive})$$

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

Given:

$P(\text{Disease})$ is the probability of having the disease

$$P(\text{Disease}) = 2\% = 0.02$$

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

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

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

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

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

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

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

$$P(\text{Positive}) = P(\text{Positive} | \text{Disease}) \times P(\text{Disease}) + P(\text{Positive} | \text{No Disease}) \times P(\text{No Disease})$$

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

$$\Rightarrow P(\text{Positive}) = 0.80 \times 0.02 + 0.10 \times 0.98 = 0.114$$

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

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

$$\Rightarrow P(\text{Disease} | \text{Positive}) = \frac{0.80 \times 0.02}{0.114}$$

$$\Rightarrow P(\text{Disease} | \text{Positive}) = 0.1404 \approx 14.04\%$$

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:

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

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

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

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

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.

HINTS:

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

c) Derivative of $f(g(x))$ is $f'(g(x)) \cdot g'(x)$

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

e) $\log(e^x) = x$

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

ANS:

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

Applying Log on both sides

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

by using hint (d)

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

by using hint (e)

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

by using hint (f)

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

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

Derivation for μ :

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

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

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

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

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

Derivation for V :

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

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

by using hints (b & c)

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

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

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

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

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