Introduction to Machine Learning: Assignment #1

Shadi Albarqouni
shadi.albarqouni@ukbonn.de

University of Bonn — October 24, 2022

1 Bayes' Theorem

Bayes's theorem is to the theory of probability what Pythagoras's theorem is to geometry. — Sir Harold Jeffreys, 1973

$$p(H = h|Y = y) = \frac{p(H = h)p(Y = y|H = h)}{p(Y = y)} \tag{1}$$

where

- the term $p(H)$ represents what we know about possible values of hypotheses H before we see any data/observations; this is called the prior distribution
- The term $p(Y = y|H = h)$ represent the probability at a point corresponding to the actual observations, y which is called the likelihood.
- The term $p(Y = y)$ is known as the marginal likelihood and computed as $\sum_{h' \in \mathcal{H}} p(H = h')p(Y = y|H = h')$
- The term $p(H = h|Y = y)$ represent the posterior distribution

Info: We have learned this in Lecture 02 and further details on Bayes' Theorem and Decision Theory can be found in Ch.03 and Ch.05, respectively.

1.1 Email Spam Checker

In most of our email clients, spam filters are included to determine whether an incoming email is a spam (fake) or non-spam (not fake). It is estimated that 60% of all incoming emails are spam. While 90% of spam emails have forged headers, only 20% of non-spam does.

Question 1

What is the probability that an email is spam given that it has a forged header? (Provide an explanation, please)

(a) 90%

(b) 54%

(c) 62%

(d) 87%

Question 2

If you learned that around 80% of spam emails have file attachments while only 10% of the non-spam emails do, what is the probability that an email is spam given that it has an attachment? ((Provide an explanation, please)

(a) 80%

(b) 48%

(c) 92%

(d) 4%

Question 3

What is the probability that an email has forged header or an attachment? (Provide an explanation, please)

(a) 3%

(b) 62%

(c) 82%

(d) 52%

1.2 COVID-19 Tests

There is a possibility of purchasing a test “A” to detect a COVID-19 disease that is present in 10% of the population. The test "A" has a probability of 99% giving a positive result if the person is infected (aka *true positive rate* or *sensitivity*). However, the test gives a false positive result (aka *fall-out*) in 5% of the cases.

Question 4

Suppose Eva tests positive for the COVID-19 disease, what is the chance that Eva actually has it? (Provide an explanation, please)

(a) 68%

(b) 16%

(c) 75%

(d) 10%

Question 5

After Eva panicked, she repeated the test, which was also positive; what is the likelihood that Eva might actually have it? (Provide an explanation, please)

(a) 57%

(b) 99%

(c) 97%

(d) 16%

Question 6

A recently developed test “B” has a sensitivity of 90% and fall-out of 1%. As you are an ML expert, the procurement officer has asked you which one he is supposed to order. To make this decision, you repeated the Question 4, however, when Eva used the test “B”, and then compared the chance that Eva actually has it. Which test do you think is more effective in identifying the infected persons? (Provide an explanation, please)

(a) Test "A"

(b) Test "B"

1.3 Implementation

Write a short script `predictCOVID.py` which takes the probability of having COVID, True Positive Rate, and False Positive Rate of a given test, as arguments, and output the probability that the person is infected given the test is positive.

Command Line

```
$ python predictCOVID.py 0.10 0.99 0.05

The probability of having COVID: 0.10
True Positive Rate (TPR): 0.99
False Positive Rate (FPR): 0.05
The probability the patient is infected given the test is +ve: ....
```

The script `predictCOVID` should have

predictCOVID.py

```
#! /usr/bin/python
import sys
print('The probability of having COVID:', sys.argv[1])
...
```

2 Univariate Models

One of the common discrete distributions is the binomial distribution where the probability mass function (pmf) and its cumulative distribution function (cdf) are given as

$$P(Y = s) = \text{Bin}(s|N, \theta) \triangleq \binom{N}{s} \theta^s (1 - \theta)^{N-s}, \tag{2}$$

$$P(Y \leq s) = \sum_{i=0}^{s} \binom{N}{i} \theta^i (1 - \theta)^{N-i}, \tag{3}$$

where $\binom{N}{k} = \frac{N!}{(N-k)!k!}$, θ is the probability of event $y = 1$, N is the number of trials, and $s \triangleq \sum_{n=1}^{N} \mathbb{I}(y_n = 1)$ is the total number of an event $y = 1$.

Info: We have learned this in Lecture 02 and further details on Bernoulli and Binomial distributions can be found in Ch.02.

2.1 Quality Control

One of the major players in Semiconductor industry heavily produces 14-nm Fin Field-Effect Transistor (FinFET)[1], which are carefully packaged in boxes of 20 for shipment. However, internal inspections have shown that 5% of their FinFET are defective.

Question 7

What is the chance that a box, ready for shipment, contains exactly 1 defective FinFET chip? (Provide an explanation, please)

(a) 26%

(b) 73%

(c) 37%

(d) 64%

Question 8

What is the chance that the same box contains at least 1 defective FinFeT chip? (Provide an explanation, please)

(a) 26%

(b) 73%

(c) 37%

(d) 64%

[1] https://en.wikipedia.org/wiki/Fin_field-effect_transistor

Question 9

On average, how many FinFEt chips in a box, ready for shipment, will most probably be defective? (Provide an explanation, please)

(a) 3 chips

(b) 5 chips

(c) 10 chips

(d) 1 chip

2.2 Gaussian Distribution

One of the common continious distributions is the Gaussian distribution where the probability density function (pdf) is given as

$$p(y) = \mathcal{N}(y|\mu, \sigma^2) \triangleq \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \tag{4}$$

where $\sqrt{2\pi\sigma^2}$ is the normalization constant.

Question 10

What are the mean and the standard deviation for the following gaussian distribution; $\frac{1}{\sqrt{50\pi}} e^{-\frac{(y^2-6y+9)}{50}}$ (Provide an explanation, please)

(a) 3, 5

(b) 9, 25

(c) 9, 5

(d) 3, 25

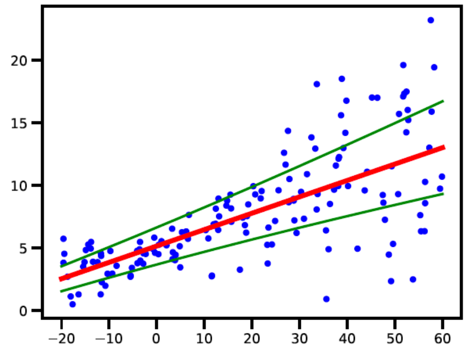
Question 11

The Gaussian distribution is widely used because (Check all that apply)

- ☐ two parameters to control and intepret
- ☐ easy to implement
- ☐ has zero entropy
- ☐ it has been found and extensively used by de Moivre

Question 12

The Gaussian distribution is typically used as a conditional probability distribution for regression problems as $p(y|x,\theta) = \mathcal{N}(y|f_\mu(x;\theta), f_\sigma(x;\theta)^2)$, which one of these formulas reflect the heteroscedastic regression (cf. the figure on the right hand) – (Provide an explanation, please)



(a) $\mathcal{N}(y|\boldsymbol{w}_\mu^T\boldsymbol{x} + b, \sigma_+(\boldsymbol{w}_\sigma^T\boldsymbol{x}))$

(b) $\mathcal{N}(y|\boldsymbol{w}^T\boldsymbol{x} + b, \sigma^2)$

(c) $\mathcal{N}(y|\boldsymbol{w}_\mu^T\boldsymbol{x} + b, \sigma(\boldsymbol{w}_\sigma^T\boldsymbol{x}))$

(d) $\mathcal{N}(y|\boldsymbol{w}^T\boldsymbol{x} + b, \sigma(\boldsymbol{w}^T\boldsymbol{x}))$

2.3 Implementation

2.3.1 Binomial

Write a short script `binomial.py` which takes the probability of an event $y = 1$, the number of trials N, and the total number s of an event $y = 1$ as arguments, and output the following: $P(Y = s)$, $P(Y \leq s)$, and $P(Y \leq s)$.

Command Line

```
$ python binomial.py 0.05 20 3

The probability of of an event: 0.05
The number of trials: 20
The total number of an event is happening: 3
....
The probability that the event is happening 3 times: ....
The probability that the event is happening at most 3 times: ....
....
```

2.3.2 Gaussian

If you were asked to implement a function for Question 12, i.e., for a given model $f(x;\theta)$, you are supposed to draw the green lines. How would you implement this function?

3 Multivariate Models

The bivariate Gaussian distribution is given as

$$\mathcal{N}(\boldsymbol{y}|\boldsymbol{\mu}, \Sigma) = \frac{1}{2\pi|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right) \tag{5}$$

where

- $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$,
- $\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12}^2 \\ \sigma_{12}^2 & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$ with
- $\rho = \frac{\text{Cov}(Y_1,Y_2)}{\sqrt{\mathbb{V}[Y_1]\mathbb{V}[Y_2]}} = \frac{\sigma_{12}^2}{\sigma_2\sigma_2}$ as a correlation coefficient, and
- $p(y_1|y_2) = \mathcal{N}\left(y_1|\mu_1 + \frac{\rho\sigma_1\sigma_2}{\sigma_2^2}(y_2-\mu_2), \sigma_1^2 - \frac{(\rho\sigma_1\sigma_2)^2}{\sigma_2^2}\right)$ as the conditional distribution.

Info: We have learned this in Lecture 02 and further details on Multivariate models can be found in Ch.03.

Question 13

Given a set of 2d points centered around zero mean with a unit standard deviation for σ_1 and σ_2 and a correlation coffiecient of 0.7, could you write the bivariate gaussian distribution?

Question 14

What would be the expected value of y_1 given $y_2 = 1$?

Question 15

What happens if you take the limit $\rho \to 0$ (aka uncorrelated random variables) of the conditional distribution?

Question 16

Same as before, but with the limit $\rho \to 1$ (aka perfectly correlated random variables) of the conditional distribution?

Question 17

Could you tell whether the covariance matris is full, diagonal, or spherical?

(a) full

(b) diagonal

(c) spherical