# Practice

This document has $9$ questions.

# Question-1

## Statement

Consider a dataset that has only $100$ points, out of which $20$ points have the value $1$, $50$ have value $2$ and $30$ have value $3$. We use a categorical distribution to model this data. The parameters of the distribution are:

$$p = P(x = 1), \quad q = P(x = 2), \quad r = P(x = 3)$$

If the distribution seems unfamiliar to you, think about an imaginary dice with three faces. What is the likelihood function for this data under this distribution?

## Options

### (a)

$$p^{30} \cdot q^{50} \cdot r^{20}$$

### (b)

$$(p + q + r)^{100}$$

### (c)

$$p^{20} \cdot q^{50} \cdot r^{30}$$

## Answer

(c)

## Solution

Let us use $a, b, c$ to refer to the number of ones, twos and threes respectively. By the i.i.d assumption, we have:

$$L(p, q, r; D) = p^a \cdot q^b \cdot r^c$$

# Question-2

## Statement

What is the value of $p + q + r$? Enter your answer correct to two decimal places.

## Answer

$1.0$

Range: $[0.99, 1.01]$

## Solution

Since the sample space is $\{1, 2, 3\}$, the probabilities should sum to $1$.

# Question-3

## Statement

What is the maximum likelihood estimate of $p$? Enter your answer correct to two decimal places.

## Answer

$0.2$

Range: $[0.19, 0.21]$

## Solution

The log-likelihood is:

$$l(a, b, c; D) = a \log p + b \log q + c \log r$$

If we want to maximize this likelihood, we can't just differentiate the function and set it to zero, as there is a constraint of $p + q + r = 1$ involved. One way to get around this is to substitute $r = 1 - p - q$ to get an unconstrained problem in two variables $(p, q)$:

$$l(a, b, c; D) = a \log p + b \log q + c \log(1 - p - q)$$

We can now compute the partial derivatives with respect to $p$ and $q$ and set them to zero. A fair amount of algebra will convince us that:

$$\hat{p} = \frac{a}{a+b+c}, \quad \hat{q} = \frac{b}{a+b+c}, \quad \hat{r} = \frac{c}{a+b+c}$$

An interesting insight that this equation gives us is the case when $c = 0$. This reduces to the MLE for a Bernoulli random variable. We can also see how this equation could be extended to the case of a categorical distribution that has a support whose cardinality is $k$. This is left as an exercise to the learners.

# Question-4

## Statement

Consider a dataset of $n$ data-points, $D = \{x_1, \cdots, x_n\}$. If we assume these points to have been generated from a Gaussian distribution, $\mathcal{N}(\mu, \sigma^2)$, what is the expression for the log-likelihood after removing constant terms?

(1) Constant terms are those that don't depend on either $\mu$ or $\sigma^2$

(2) $\log$ always means $\log_e$ unless otherwise specified.

## Options

### (a)

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

### (b)

$$-\log \sigma - \frac{1}{2\sigma^2} \cdot \sum_{i=1}^{n} (x_i - \mu)^2$$

### (c)

$$-n \log \sigma - \frac{1}{2\sigma^2} \cdot \sum_{i=1}^{n} (x_i - \mu)^2$$

### (d)

$$n \log \sigma + \frac{1}{2\sigma^2} \cdot \sum_{i=1}^{n} (x_i - \mu)^2$$

## Answer

(c)

## Solution

First, we compute the likelihood. Using the i.i.d assumption:

$$L(\mu, \sigma^2; D) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[\frac{-(x_i - \mu)^2}{2\sigma^2}\right]$$

Next, the log-likelihood:

$$l(\mu, \sigma^2; D) = \sum_{i=1}^{n} \left[-\log(\sqrt{2\pi}\sigma) - \frac{(x_i - \mu)^2}{2\sigma^2}\right]$$

The first term inside the summation is independent of the $x_i$s, hence it can be taken out after scaling it by a factor of $n$:

$$l(\mu, \sigma^2; D) = -n \log(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2} \cdot \sum_{i=1}^{n} (x_i - \mu)^2$$

After removing the constants from the first term, we get:

$$l(\mu, \sigma^2; D) = -n \log \sigma - \frac{1}{2\sigma^2} \cdot \sum_{i=1}^{n} (x_i - \mu)^2$$

# Question-5

## Statement

Consider a dataset of heights of $300$ individuals. The first $100$ are drawn from the active pool of basketball players in the NBA. The next $100$ are drawn from the list of chess grand masters. The last $100$ are drawn randomly from the city of Chennai. All $300$ individuals are in the age-group of $20$ to $25$. If we use a GMM to understand this data, what is a good choice of $K$, the number of mixtures?

## Answer

2

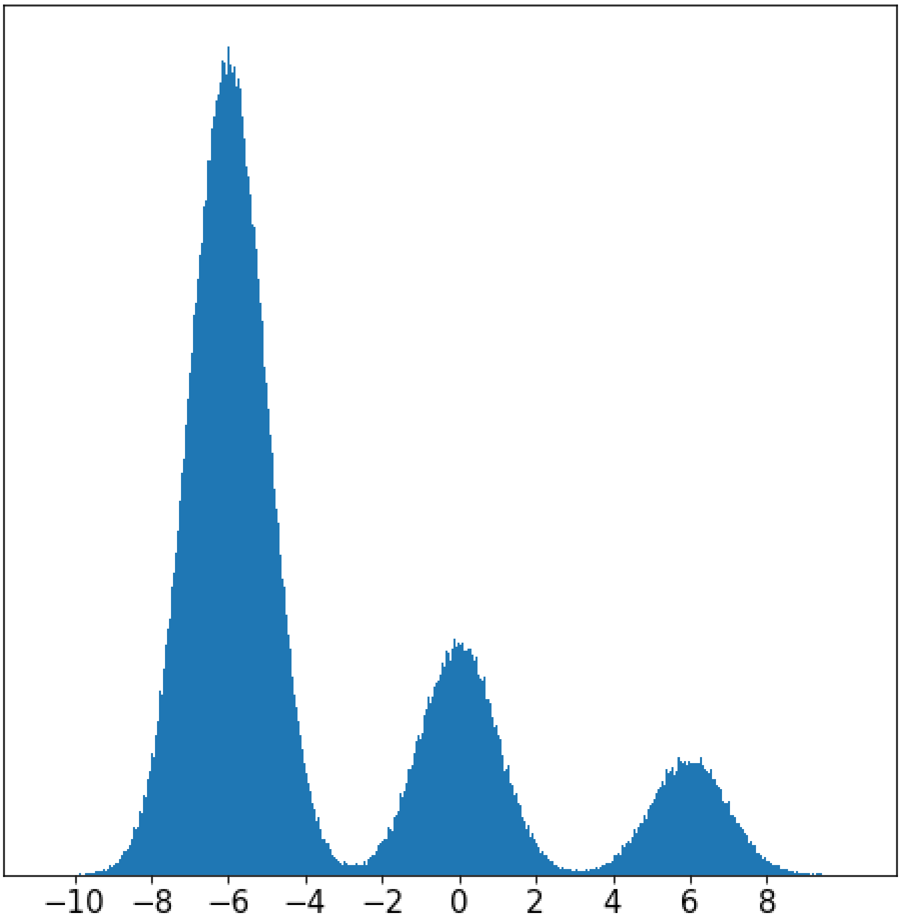
## Solution

Though there are three different classes of people, as far as heights are concerned, basketball players certainly are in a different zone. Height is not correlated with chess. Hence we can reason that the average height of a chess player will not be too different from the average height of someone from Chennai. So, $2$ mixtures would be sufficient.

# Common Data for questions (6) to (8)

## Statement

Consider the histogram of one million points sampled from a GMM with three mixtures as shown in the figure below. The mixtures are labeled from left to right as $1$, $2$ and $3$. The mean for each mixture is one of the ticks displayed on the x-axis. All the mixtures have unit variance:



# Question-6

## Statement

What is the mean of mixture-3? Note that the mean is an integer here.

## Answer

$6$

## Solution

Visual inspection

# Question-7

## Statement

Which of the following could be the values of $\pi_1, \pi_2$ and $\pi_3$?

## Options

### (a)

$\pi_1 = \pi_2 = \pi_3 = \frac{1}{3}$

### (b)

$\pi_1 = 0.4, \pi_2 = 0.1, \pi_3 = 0.4$

### (c)

$\pi_1 = 0.7, \pi_2 = 0.2, \pi_1 = 0.1$

### (d)

$\pi_1 = 0.4, \pi_2 = 0.4, \pi_3 = 0.1$

## Answer

(c)

## Solution

The heights of the mixtures give us an idea of their importance.

# Question-8

## Statement

If the point $-3$ is observed, what is the probability that it has come from mixture-2? Use the values of $\pi_1, \pi_2, \pi_3$ obtained from the previous question. Enter your answer correct to two decimal places.

## Answer

$0.22$

## Solution

We need to compute $P(z = 2 \mid x = -3)$. Using Bayes' rule:

$$P(z = 2 \mid x = -3) = \frac{P(z = 2) \cdot f(x = -3 \mid z = 2)}{f(x = -3)}$$

We have:

- $\mu_1 = -6, \mu_2 = 0, \mu_3 = 6.$
- $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 1$
- $\pi_1 = 0.7, \pi_2 = 0.2, \pi_3 = 0.1$

$$f(x \mid z = k) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left[\frac{-(x - \mu_k)^2}{2\sigma_k^2}\right]$$

We now have to compute the each of these quantities.

# Question-9

## Statement

Assume that you are given a set of one $10000$ data-points in $\mathbb{R}$. You fit a GMM with $K = 2$ for this dataset using the EM algorithm to estimate the parameters. The EM algorithm was initialized as follows:

(1) $\mu_1 = -1, \mu_2 = 1$

(2) $\pi_1 = \pi_2 = 0.5$

(3) $\sigma_1^2 = \sigma_2^2 = 1$

The estimated means are $\hat{\mu}_1$ and $\hat{\mu}_2$ for the the two mixtures. A little while later, a domain expert comes and tells you that the dataset given to you was actually sampled from a Gaussian with mean $0$ and variance $1$. Which of the following options is true? Code the EM algorithm and observe what happens.

## Options

### (a)

$\hat{\mu}_1$ is very close to $\hat{\mu}_2$ but both are not close to $0$

### (b)

$\hat{\mu}_1$ is not close to $\hat{\mu}_2$ and neither of them is close to $0$

### (c)

$\hat{\mu}_1$ is very close to $\hat{\mu}_2$ and both are close to $0$

## Answer

(b)

## Solution

The points that are in the interval around the mean, somewhere around $[-0.7, 0.7]$, form some sort of a barrier. The mixture on the left is unable to advance beyond a certain point to the right. Likewise, the mixture on the right is unable to advance beyond a certain point to the left. This is observed for initializations of the means that are significantly far away from the true mean. This problem will be discussed during the programming session.