

Exercise 1: Bernoulli Distribution

Assume that N data points are independent, which means that the joint conditional density can be factorised into N separate terms, one for each data point:

$$\mathcal{L} = p(D | q) = \prod_{i=1}^N p(x_i | q) = \prod_{i=1}^N q^{x_i} (1 - q)^{1-x_i}.$$

This equation tells us how likely our dataset D is, given the current model parametrised by success probability q . Since the dataset is fixed, changing the model will result in different likelihood values. Maximum Likelihood method tries to find the model that maximises the likelihood \mathcal{L} . Normally, we will maximise the *natural logarithm* (i.e. logarithm with base e) of the likelihood because the estimated argument \hat{q} that maximises the log-likelihood will also maximise the likelihood.

$$\begin{aligned} \log \mathcal{L} &= \sum_{i=1}^N \log q^{x_i} (1 - q)^{1-x_i} = \sum_{i=1}^N (\log q^{x_i} + \log(1 - q)^{1-x_i}) \\ &= \sum_{i=1}^N (x_i \log q + (1 - x_i) \log(1 - q)) \\ &= \sum_{i=1}^N (x_i \log q) + \sum_{i=1}^N ((1 - x_i) \log(1 - q)). \end{aligned}$$

We now can find the optimal parameter by taking derivative, equating them to zero and solving for turning points. For q ,

$$\begin{aligned} \frac{\partial \log \mathcal{L}}{\partial q} &= \sum_{i=1}^N (x_i) \frac{\partial \log q}{\partial q} + \sum_{i=1}^N (1 - x_i) \frac{\partial \log(1 - q)}{\partial q} \\ &= \frac{1}{q} \sum_{i=1}^N (x_i) - \frac{1}{1 - q} \sum_{i=1}^N (1 - x_i) \end{aligned}$$

Equating $\partial \log \mathcal{L} / \partial q$ to zero yields

$$\frac{1}{\hat{q}} \sum_{i=1}^N (x_i) - \frac{1}{1 - \hat{q}} \sum_{i=1}^N (1 - x_i) = 0,$$

equivalently

$$(1 - \hat{q}) \sum_{i=1}^N (x_i) = \hat{q} \left(N - \sum_{i=1}^N (x_i) \right).$$

Simplifying it yields the maximum likelihood estimate of the Bernoulli distribution

$$\hat{q}_{\text{ML}} = \frac{1}{N} \sum_{i=1}^N x_i$$

Exercise 2: Univariate Gaussian Distribution

Assume that a dataset D of N values (i.e. $D = \{x_1, x_2, \dots, x_N\}$), was sampled from a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$. Assuming that the data points are independently and identically distributed. Let's get started by writing down the joint density function:

$$\mathcal{L} = p(D \mid \mu, \sigma^2) = \prod_{i=1}^N \mathcal{N}(x_i; \mu, \sigma^2) = \prod_{i=1}^N \frac{1}{\sigma\sqrt{2\pi}} \exp \left\{ -\frac{(x_i - \mu)^2}{2\sigma^2} \right\}$$

Follow the same approach as in Exercise 1 by taking the *natural logarithm* of \mathcal{L} as

$$\begin{aligned} \log \mathcal{L} &= \sum_{i=1}^N \log \frac{1}{\sigma\sqrt{2\pi}} \exp \left\{ -\frac{(x_i - \mu)^2}{2\sigma^2} \right\} \\ &= \sum_{i=1}^N \left(-\log \sigma - \frac{1}{2} \log(2\pi) - \frac{(x_i - \mu)^2}{2\sigma^2} \right) \end{aligned}$$

Unlike the Bernoulli distribution which has only one parameter q , the Gaussian distribution is characterised by two parameters: mean (μ) and variance (σ). Therefore, we have to calculate two partial derivatives with respect to μ and σ , i.e. $\partial \log \mathcal{L} / \partial \mu$ and $\partial \log \mathcal{L} / \partial \sigma$. By equating both derivatives to zero, we can find the maximum likelihood estimates $\hat{\mu}_{\text{ML}}$ and $\hat{\sigma}_{\text{ML}}^2$ for the Gaussian model. Calculating the two derivatives as follows:

$$\frac{\partial \log \mathcal{L}}{\partial \sigma} = \sum_{i=1}^N \left(-\frac{1}{\sigma} + \frac{(x_i - \mu)^2}{\sigma^3} \right) \quad \text{and} \quad \frac{\partial \log \mathcal{L}}{\partial \mu} = \sum_{i=1}^N \frac{x_i - \mu}{\sigma^2}$$

By equating both derivatives to zero, we obtain

$$\begin{aligned} \frac{\partial \log \mathcal{L}}{\partial \mu} = 0 &\Rightarrow \sum_{i=1}^N (x_i - \mu) = 0 \\ &\Rightarrow \mu = \frac{1}{N} \sum_{i=1}^N x_i, \end{aligned}$$

and

$$\begin{aligned} \frac{\partial \log \mathcal{L}}{\partial \sigma} = 0 &\Rightarrow N - \frac{1}{\sigma^2} \sum_{i=1}^N (x_i - \mu)^2 = 0 \\ &\Rightarrow \sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2 \end{aligned}$$

Hence, the ML estimate of the mean and variance of the Gaussian distribution is

$$\hat{\mu}_{\text{ML}} = \frac{1}{N} \sum_{i=1}^N x_i \quad \text{and} \quad \hat{\sigma}_{\text{ML}}^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \hat{\mu}_{\text{ML}})^2$$

Exercise 3a

Since the sensors are independent, the likelihood is

$$\mathcal{L}(x) = p(z_1, z_2 | x) = p(z_1|x)p(z_2|x)$$

and since the sensors are gaussian

$$\mathcal{L}(x) \propto e^{-\frac{(z_1-x)^2}{2\sigma^2}} \times e^{-\frac{(z_2-x)^2}{2\sigma^2}} = e^{-\frac{(z_1-x)^2 + (z_2-x)^2}{2\sigma^2}}$$

Here we ignored the irrelevant normalisation constants. Now the log-likelihood is given by

$$\log \mathcal{L}(x) = \frac{(z_1 - x)^2 + (z_2 - x)^2}{2\sigma^2} = \frac{(x - \bar{x})^2}{\sigma^2} + c(z_1, z_2),$$

where $\bar{x} = \frac{z_1 + z_2}{2}$, and $c(z_1, z_2)$ is a constant independent of x . The maximum likelihood estimate of x is defined as

$$\hat{x} = \arg \max_x \mathcal{L}(x) = \arg \min_x (-\log \mathcal{L}(x))$$

Now let's compute the min by differentiating $-\log \mathcal{L}(x)$ with respect to x

$$\frac{\partial \{-\log \mathcal{L}(x)\}}{\partial x} = \frac{2(x - \bar{x})}{\sigma^2} = 0$$

Therefore, $\hat{x}_{\text{ML}} = \bar{x} = (z_1 + z_2)/2$.

Exercise 3b

The sensors are independent

$$\mathcal{L}(x) = p(z_1, z_2|x) = p(z_1|x)p(z_2|x) \propto e^{-\frac{(z_1-x)^2}{2\sigma_1^2}} \times e^{-\frac{(z_2-x)^2}{2\sigma_2^2}}$$

The negative log-likelihood is then

$$\begin{aligned} -\log \mathcal{L}(x) &= \frac{1}{2} \left[\frac{(z_1 - x)^2}{\sigma_1^2} + \frac{(z_2 - x)^2}{\sigma_2^2} \right] + \text{const} \\ &= \frac{1}{2} \left[\left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} \right) x^2 - 2 \left(\frac{z_1}{\sigma_1^2} + \frac{z_2}{\sigma_2^2} \right) x \right] + \text{const} \\ &= \frac{1}{2} \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} \right) \left[x - \frac{\sigma_1^{-2} z_1 + \sigma_2^{-2} z_2}{\sigma_1^{-2} + \sigma_2^{-2}} \right]^2 + \text{const} \end{aligned}$$

which is maximised with respect to x when

$$\hat{x}_{\text{ML}} = \frac{\sigma_1^{-2} z_1 + \sigma_2^{-2} z_2}{\sigma_1^{-2} + \sigma_2^{-2}}.$$

For example, if the sensors are $p(z_1|x) \sim \mathcal{N}(x, 10^2)$ and $p(z_2|x) \sim \mathcal{N}(x, 20^2)$. Suppose we obtain sensor readings of $z_1 = 130$ and $z_2 = 170$, then

$$\hat{x}_{\text{ML}} = \frac{130/10^2 + 170/20^2}{1/10^2 + 1/20^2} = 138.0$$

It shows that the ML estimate is closer to the more confident measurement (i.e. smaller variance).