(To obtain support for these exercises, please post your questions on the Microsoft Team, or talk to one of lecturers over their office hours.)

Exercise 1: Bernoulli Distribution

We have discussed continuous random variables and shown how to specify them via their probability density functions. A *discrete* random variable can only take on a countable number of values. The distribution of a discrete random variable X with parameter θ is given by its probability mass function (pmf),

$$p_\theta(x) = \Pr(X = x).$$

If $X^{(1)}, \ldots, X^{(n)}$ are n independent discrete random variables with the same probability mass function p_θ, then their joint probability mass function is given by

$$p_\theta(x^{(1)}, \ldots, x^{(n)}) = \Pr\left(X^{(1)} = x^{(1)} \wedge \cdots \wedge X^{(n)} = x^{(n)}\right) = \prod_{i=1}^{n} p_\theta(x^{(i)}).$$

The likelihood function for the parameter θ given n independent observations is defined a

$$\mathcal{L}(\theta \mid x^{(1)}, \ldots, x^{(n)}) := \prod_{i=1}^{n} p_\theta(x^{(i)}).$$

A Bernoulli random variable X with parameter $q \in [0, 1]$ is a *discrete* random variable that can only take two values, 0 or 1, and where $\Pr(X = 1) = q$ and $\Pr(X = 0) = 1 - q$. The parameter q is often called the *success probability*. For example, the outcome of flipping an unbiased coin can be modelled as a Bernoulli random variable with success probability $1/2$, where the event $X = 0$ indicates that the coin came up with head, and the event $X = 1$ indicates that the coin came up with tail.
Assume that we have a dataset $D = \{x^{(1)}, \ldots, x^{(n)}\}$ of n observations. Let us further assume that each observation $x^{(i)}$ is an independent sample from a Bernoulli distribution, all with the same success probability q.

a) Explain why the joint probability mass function of the observations D can be written as

$$p_q\left(x^{(1)}, \ldots, x^{(n)}\right) = \prod_{i=1}^{n} q^{x^{(i)}}(1-q)^{1-x^{(i)}}.$$

b) Compute the log-likelihood function for the parameter q, i.e., $\log \mathcal{L}(q \mid x^{(1)}, \ldots, x^{(n)})$.

c) Compute the maximum likelihood estimate for the success probability q given the observations D, i.e., compute

$$\begin{aligned}\hat{q}_{\mathrm{ML}} &= \arg\max_q \ \mathcal{L}\left(q \mid x^{(1)}, \ldots, x^{(n)}\right) \\ &= \arg\min_q \ -\log \mathcal{L}\left(q \mid x^{(1)}, \ldots, x^{(n)}\right).\end{aligned}$$

Hint: Solve the equation

$$\frac{d(\log \mathcal{L})}{dq} = 0.$$

Exercise 2: Univariate Gaussian Distribution

Gaussian (also called normal) random variables are used in many continuous applications. The Gaussian distribution is defined over the set of real numbers (i.e. $(-\infty, +\infty)$), and has a probability density function (pdf)

$$p\left(y \mid \mu, \sigma^2\right) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}.$$

This distribution is characterised by two parameters, the mean μ and the variance σ^2. The Gaussian distribution is often denoted $\mathcal{N}(\mu, \sigma^2)$.
Assume that we have a dataset $D = \{x^{(1)}, \ldots, x^{(n)}\}$ of n observations. Let us further assume that each observation $x^{(i)}$ is an independent sample from the normal distribution $\mathcal{N}(\mu, \sigma^2)$.
Unlike the Bernoulli distribution which has only one parameter q, the Gaussian distribution is characterised by two parameters: mean (μ) and variance (σ).

a) Recall that the joint density function for two independent continuous random variables X and Y with the same density function q is

$$p(x, y) = q(x)q(y).$$

Write the joint density function for the observations D.

b) Compute the log-likelihood function for the parameters μ and σ^2, i.e., $\log \mathcal{L}(\mu, \sigma^2 \mid x^{(1)}, \ldots, x^{(n)})$.

c) Compute the maximum likelihood estimate for the mean μ and the variance σ^2 given the observations D, i.e., find

$$\begin{aligned}
\widehat{(\mu, \sigma^2)}_{\mathrm{ML}} &= \underset{\mu, \sigma^2}{\arg\max}\ \mathcal{L}\left(\mu, \sigma^2 \mid x^{(1)}, \ldots, x^{(n)}\right) \\
&= \underset{\mu, \sigma^2}{\arg\min}\ -\log \mathcal{L}\left(\mu, \sigma^2 \mid x^{(1)}, \ldots, x^{(n)}\right).
\end{aligned}$$

Hints: Note that the log-likelihood is in this case is a function of two variables. Hence, you need to compute the partial derivates and solve

$$\frac{\partial(\log \mathcal{L})}{\partial \mu} = \frac{\partial(\log \mathcal{L})}{\partial \sigma} = 0.$$

Furthermore, you may find the following derivatives useful.

$$\begin{aligned}
\frac{d}{dx}(\log x) &= \frac{1}{x\ln(2)} \\
\frac{d}{dx}\left(\frac{1}{x^2}\right) &= -\frac{2}{x^3} \\
\frac{d}{dx}(x-y)^2 &= 2(y-x)
\end{aligned}$$

Exercise 3

Suppose we have two independent measurements $(z^{(1)}, z^{(2)})$ of a length $x \in \mathbb{R}$. Consider the following two probabilistic models of the measurements.

a) $z^{(1)} \sim \mathcal{N}(x, \sigma^2)$ and $z^{(2)} \sim \mathcal{N}(x, \sigma^2)$ (i.e., equal variance).

b) $z^{(1)} \sim \mathcal{N}(x, \sigma_1^2)$ and $z^{(2)} \sim \mathcal{N}(x, \sigma_2^2)$ (i.e., unequal variances).

Determine the maximum likelihood estimate of x for each model.