P. Koumoutsakos
ETH Zentrum, CLT E 13
CH-8092 Zürich

Spring semester 2020

HW 1 - Probability & Bayesian Inference

Issued: February 24, 2020
Due Date: March 09, 2020, 08:00am

1-Week Milestone: Solve tasks 1 and 2

Task 1: Probability Theory Reminders

In this exercise we fix the notation we will use during this course and refresh our memory on basic properties of random variables. Present your answers *in detail*.

a) [10pts] A random variable with normal (or Gaussian) distribution $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$ has probability density function (pdf) given by

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{1}$$

Show that the mean and the variance of X are given by $\mathbb{E}\left[X\right] = \mu$ and $\mathbb{E}\left[(X-\mu)^2\right] = \sigma^2$, respectively.

The mean of X is given by,

$$\begin{aligned}
\mathbb{E}\left[X\right] &= \int_{-\infty}^{+\infty} x f_X(x) dx \\
&= \int_{-\infty}^{+\infty} (x-\mu) f_X(x) dx + \mu \int_{-\infty}^{+\infty} f_X(x) dx \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{+\infty} y e^{-\frac{y^2}{2\sigma^2}} dy + \mu \\
&= \mu,
\end{aligned}$$

where we have used that $\int_{-\infty}^{+\infty} f_X(x) dx = 1$ because f_X is a density and that $\int_{-\infty}^{+\infty} y e^{-\frac{y^2}{2\sigma^2}} dy = 0$ due to symmetry.

The variance of X is given by,

$$\mathbb{E}\left[(X-\mu)^2\right] = \int_{-\infty}^{+\infty} (x-\mu)^2 f_X(x) dx.$$

By setting $y = x - \mu$ and $a = \frac{1}{2\sigma^2}$, we obtain

$$\begin{aligned}
\mathbb{E}\left[(X-\mu)^2\right] &= \sqrt{\frac{a}{\pi}} \int\limits_{-\infty}^{+\infty} y^2 e^{-ay^2} dy \\
&= -\sqrt{\frac{a}{\pi}} \frac{\partial}{\partial a} \int\limits_{-\infty}^{+\infty} e^{-ay^2} dy \\
&= -\sqrt{\frac{a}{\pi}} \frac{\partial}{\partial a} \sqrt{\frac{\pi}{a}} \\
&= \frac{1}{2a} \\
&= \sigma^2 .
\end{aligned}$$

b) [10pts] The probability that a random variable X with pdf f_X is less or equal than any $x \in \mathbb{R}$ is given by,

$$P(X \leq x) = F_X(x) = \int_{-\infty}^{x} f_X(z) \, \mathrm{d}z \, . \tag{2}$$

The function F_X is called the cumulative distribution function (cdf).
The Laplace distribution with parameters μ and β has pdf,

$$f(x) = \frac{1}{2\beta} \exp\left(-\frac{|x-\mu|}{\beta}\right) . \tag{3}$$

i) Find the cdf of the Laplace distribution.
Combining eq. (2) and eq. (3), the cdf of the Laplace distribution is

$$F_X(x) = \frac{1}{2\beta} \int_{-\infty}^{x} \exp\left(-\frac{|z-\mu|}{\beta}\right) \mathrm{d}z \, .$$

Apply $s = z - \mu$,

$$F_X(x) = \frac{1}{2\beta} \int_{-\infty}^{x-\mu} \exp\left(-\frac{|s|}{\beta}\right) \mathrm{d}s \, .$$

Due to the absolute value, we have two cases. For $x - \mu < 0$,

$$F_X(x) = \frac{1}{2\beta} \int_{-\infty}^{x-\mu} \exp\left(\frac{s}{\beta}\right) \mathrm{d}s = \frac{1}{2} \exp\left(\frac{x-\mu}{\beta}\right) ,$$

and for $x - \mu \geq 0$,

$$\begin{aligned}
F_X(x) &= \frac{1}{2\beta} \int_{-\infty}^{0} \exp\left(\frac{s}{\beta}\right) \mathrm{d}s + \frac{1}{2b} \int_{0}^{x-\mu} \exp\left(-\frac{s}{\beta}\right) \mathrm{d}s \\
&= 1 - \frac{1}{2} \exp\left(-\frac{x-\mu}{\beta}\right) .
\end{aligned}$$

Finally,

$$F_X(x) = \begin{cases} \frac{1}{2}\exp\left(\frac{x-\mu}{\beta}\right), & x < \mu \\ 1 - \frac{1}{2}\exp\left(-\frac{x-\mu}{\beta}\right), & x \geq \mu\,. \end{cases}$$

ii) Use the cdf to find the median of the Laplace distribution.

The median is defined as the point m that satisfies $F_X(m) = \frac{1}{2}$. For both cases, $x < \mu$ or $x \geq \mu$, it is easy to verify that $m = \mu$.

c) [10pts] The pdf of the quotient $Q = X/Y$ of two random variables X, Y is given by,

$$f_Q(q) = \int_{-\infty}^{\infty} |x|\, f_{X,Y}(qx, x)\, \mathrm{d}x\,, \tag{4}$$

where $f_{X,Y}$ is the joint pdf of X and Y.

Assume that X and Y are independent random variables with pdfs $f_X(x) = \mathcal{N}\left(x|0, \sigma_X^2\right)$ and $f_Y(y) = \mathcal{N}\left(y|0, \sigma_Y^2\right)$.

i) Find the joint pdf of X and Y.

Since X and Y are independent it holds that $f_{X,Y} = f_X f_Y$,

$$\begin{aligned} f_{X,Y}(x,y) &= \mathcal{N}\left(x|0,\sigma_X^2\right)\mathcal{N}\left(y|0,\sigma_Y^2\right) \\ &= \frac{1}{2\pi\sigma_x\sigma_y}\exp\left(-\frac{x^2}{2\sigma_x^2} - \frac{y^2}{2\sigma_y^2}\right)\,. \end{aligned}$$

ii) Show that $Q = X/Y$ follows a Cauchy distribution with zero location parameter and scale $\gamma = \sigma_X/\sigma_Y$. The pdf of a Cauchy distribution with location parameter x_0 and scale γ is given by,

$$f(x) = \frac{1}{\pi}\frac{\gamma}{(x-x_0)^2 + \gamma^2}\,. \tag{5}$$

From eq. (4) and point i), we can write

$$\begin{aligned} f_Q(q) &= \frac{1}{2\pi\sigma_x\sigma_y}\int_{-\infty}^{\infty} |x|\,\exp\left(-\frac{x^2}{2\sigma_x^2} - \frac{q^2x^2}{2\sigma_y^2}\right)\mathrm{d}x \\ &= \frac{1}{2\pi\sigma_x\sigma_y}\int_{-\infty}^{\infty} |x|\,\exp\left(-x^2\left(\frac{1}{2\sigma_x^2} + \frac{q^2}{2\sigma_y^2}\right)\right)\mathrm{d}x \\ &= \frac{1}{\pi\sigma_x\sigma_y}\int_{0}^{\infty} x\,\exp\left(-x^2\left(\frac{1}{2\sigma_x^2} + \frac{q^2}{2\sigma_y^2}\right)\right)\mathrm{d}x\,, \end{aligned}$$

where in the last step we used the fact that the integrand is symmetric around zero. Using the fact that

$$\int_0^{\infty} x\,\exp\left(-kx^2\right)\mathrm{d}x = \frac{1}{2k}\,,$$

for $k = \frac{\sigma_y^2 + q^2\sigma_x^2}{2\sigma_x^2\sigma_y^2}$,

$$f_Q(q) = \frac{1}{\pi}\frac{\sigma_x/\sigma_y}{q^2 + (\sigma_x/\sigma_y)^2}\,.$$

Task 2: Bayesian Inference

You are given a set of points $\boldsymbol{d} = \{d_i\}_{i=1}^N$ with $d_i \in \mathbb{R}$. You make the *modelling assumption* that the points come from N realisations of N *independent* random variables X_i, $i = 1, \dots, N$, that follow normal distribution with unknown parameter μ and known parameter $\sigma = 1$.

a) [10pts] Formulate the *likelihood function* of μ,

$$\mathcal{L}(\mu) := p(\boldsymbol{d}|\mu)\,, \tag{6}$$

where p is the conditional pdf of $\boldsymbol{d}$ conditioned on μ.

From the modelling assumption, d_i are independent,

$$p(\boldsymbol{d}|\mu) = \prod_{i=1}^{N} p(d_i|\mu)\,,$$

and $X_i \sim \mathcal{N}(\mu, 1)$,

$$\begin{aligned} p(\boldsymbol{d}|\mu) &= \prod_{i=1}^{N} \mathcal{N}(d_i|\mu, 1) \\ &= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(d_i-\mu)^2\right) \\ &= (2\pi)^{-N/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{N}(d_i-\mu)^2\right)\,. \end{aligned} \tag{7}$$

The log-likelihood is given by,

$$\log p(\boldsymbol{d}|\mu) = -\frac{N}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{N}(d_i-\mu)^2\,.$$

b) [15pts] Find the *maximum likelihood estimate* (MLE) of μ, i.e.,

$$\hat{\mu} = \arg\max_{\mu} \mathcal{L}(\mu). \tag{8}$$

You may find useful that $\arg\max_\mu \mathcal{L}(\mu) = \arg\max_\mu \log \mathcal{L}(\mu)$.

It holds that,

$$\begin{aligned} \hat{\mu} &= \arg\max_{\mu} p(\boldsymbol{d}|\mu) \\ &= \arg\max_{\mu} \log p(\boldsymbol{d}|\mu) \\ &= \arg\max_{\mu} \left(-\frac{N}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{N}(d_i-\mu)^2\right) \\ &= \arg\max_{\mu} -\frac{1}{2}\sum_{i=1}^{N}(d_i-\mu)^2\,. \end{aligned}$$

The optimum $\hat{\mu}$ satisfies,

$$
\begin{aligned}
0 &= \frac{\mathrm{d}}{\mathrm{d}\mu} \frac{1}{2} \sum_{i=1}^{N} (d_i - \mu)^2 \Bigg|_{\mu=\hat{\mu}} \\
&= \sum_{i=1}^{N} d_i - N\hat{\mu} \\
\Rightarrow \hat{\mu} &= \frac{1}{N} \sum_{i=1}^{N} d_i \,.
\end{aligned}
$$

c) [30pts] Before observing any data $\boldsymbol{d}$ you had the belief that μ follows a normal distribution with mean μ_0 and variance σ_0^2. After observing the dataset $\boldsymbol{d}$ you *update your belief* by using Bayes' theorem. Identify the *posterior distribution* $p(\mu|\boldsymbol{d})$ of μ conditioned on $\boldsymbol{d}$. Calculate the mean and the variance of $p(\mu|\boldsymbol{d})$.

The posterior is given by Bayes' theorem,

$$
p(\mu|\boldsymbol{d}) = \frac{p(\boldsymbol{d}|\mu)\, p(\mu)}{p(\boldsymbol{d})}.
$$

where $p(\mu) = \mathcal{N}(\mu|\mu_0, \sigma_0^2)$. Working with the numerator and forgetting for the moment all the proportionality constants, we have

$$
\begin{aligned}
p(\boldsymbol{d}|\mu)\, p(\mu) &= \prod_{i=1}^{N} \mathcal{N}(d_i|\mu, 1)\, \mathcal{N}(\mu|\mu_0, \sigma_0^2) \\
&= (2\pi)^{-N/2} \exp\left(-\frac{1}{2} \sum_{i=1}^{N} (d_i - \mu)^2 \right) (2\pi\sigma_0^2)^{-1/2} \exp\left(-\frac{1}{2\sigma_0^2} (\mu - \mu_0)^2 \right) \\
&\propto \exp\left(-\frac{1}{2} \sum_{i=1}^{N} (d_i^2 - 2\mu d_i + \mu^2) - \frac{1}{2\sigma_0^2} (\mu^2 - 2\mu\mu_0 + \mu_0^2) \right) \\
&= \exp\left(-\frac{(1 + N\sigma_0^2)\mu^2 - 2\mu(\mu_0 + \sigma_0^2 \sum_i d_i) + \mu_0 + \sigma_0^2 \sum_i d_i}{2\sigma_0^2} \right) \\
&= \exp\left(-\frac{1}{2} \frac{\mu^2 - 2\mu \frac{\mu_0 + \sigma_0^2 \sum_i d_i}{1 + N\sigma_0^2} + \frac{\mu_0 + \sigma_0^2 \sum_i d_i}{1 + N\sigma_0^2}}{\frac{\sigma_0^2}{1 + N\sigma_0^2}} \right) \,.
\end{aligned}
\tag{9}
$$

In order to lighten the notation, we set

$$
\bar{\mu} = \frac{\mu_0 + \sigma_0^2 \sum_i d_i}{1 + N\sigma_0^2} \qquad \text{and} \qquad \bar{\sigma}^2 = \frac{\sigma_0^2}{1 + N\sigma_0^2} \,.
$$

Now we work only with the numerator in the exponential eq. (9) and complete the square,

$$
\begin{aligned}
\mu^2 - 2\mu\bar{\mu} + \bar{\mu} &= \mu^2 - 2\mu\bar{\mu} + \bar{\mu}^2 - \bar{\mu}^2 + \bar{\mu} \\
&= (\mu - \bar{\mu})^2 - \bar{\mu}^2 + \bar{\mu} \,.
\end{aligned}
\tag{10}
$$

Substituting eq. (10) in eq. (9),

$$
\begin{aligned}
p(\boldsymbol{d}|\mu)\, p(\mu) &\propto \exp\left(-\frac{1}{2}\frac{(\mu-\bar{\mu})^2-\bar{\mu}^2+\bar{\mu}}{\bar{\sigma}^2}\right) \\
&\propto \exp\left(-\frac{1}{2}\frac{(\mu-\bar{\mu})^2}{\bar{\sigma}^2}\right)\,,
\end{aligned} \tag{11}
$$

where the last proportionality is true due to the fact that $\bar{\mu}$ does not depend on μ. At this point, we use equality in eq. (11) and find the proportionality constant C such that the posterior integrates to 1 (and therefore is a distribution),

$$
\begin{aligned}
p(\mu|\boldsymbol{d}) &= C\,\exp\left(-\frac{1}{2}\frac{(\mu-\bar{\mu})^2}{\bar{\sigma}^2}\right) \\
&= \frac{1}{\sqrt{2\pi\bar{\sigma}^2}}\exp\left(-\frac{1}{2}\frac{(\mu-\bar{\mu})^2}{\bar{\sigma}^2}\right) \\
&= \mathcal{N}(\mu|\bar{\mu},\bar{\sigma}^2)\,.
\end{aligned} \tag{12}
$$

d) [5pts] Find the *maximum a posteriori* (MAP) estimate of μ, i.e.,

$$
\hat{\mu} \;=\; \arg\max_{\mu}\; p(\mu|\boldsymbol{d})\,. \tag{13}
$$

The normal distribution attains its maximum at the mean. From eq. (12)

$$
\hat{\mu} = \frac{\mu_0+\sigma_0^2\sum_i d_i}{1+N\sigma_0^2}. \tag{14}
$$

e) [10pts] Perform (c) and (d) using as prior an uninformative distribution, i.e. a uniform distribution in $\mathbb{R}$, and compare the MAP with the MLE. Although this is not a distribution, since it is not integrable over $\mathbb{R}$, we are allowed to use it in Bayes' theorem as long as the posterior is a distribution. These priors are called *improper priors* and a common choice in practical applications when there is no prior information on the parameters.

The prior is proportional to 1, $p(\mu)\propto 1$, and the likelihood is given by eq. (7). Thus, the posterior is proportional to,

$$
\begin{aligned}
p(\mu|\boldsymbol{d}) &\propto (2\pi)^{-N/2}\,\exp\left(-\frac{1}{2}\sum_{i=1}^{N}(d_i-\mu)^2\right) \\
&\propto \exp\left(-\frac{1}{2N}\left(\mu-\frac{1}{N}\sum_{i=1}^{N}d_i\right)^2\right)\,,
\end{aligned}
$$

where we have rearranged the terms in the exponential and completed the square, in the same way as in Task 2c. It is easy to find the proportionality constant and write,

$$
\begin{aligned}
p(\mu|\boldsymbol{d}) &= \frac{1}{\sqrt{2\pi N^{-1}}} \exp\left(-\frac{1}{2N}\left(\mu - \frac{1}{N}\sum_{i=1}^{N} d_i\right)^2\right) \\
&= \mathcal{N}\left(\mu \,|\, \frac{1}{N}\sum_{i=1}^{N} d_i, \frac{1}{N}\right) .
\end{aligned}
$$

For this choice of prior, the MAP coincides with the MLE.

Task 3: Bayesian Inference: Linear Model

You are given the linear regression model that describes the relation between variables x and y,

$$y = \beta x + \epsilon,$$

where β is the regression parameter, y is the output quantity of interest (QoI) of the system, x is the input variable and ϵ is the random variable accounting for model and measurement errors. The model error is quantified by a Gaussian distribution $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

You are given one measurement data point, $D = \{x_0, y_0\}$.

a) [20pts] Consider an uninformative prior for β and identify the posterior distribution of β after observing D. Calculate the MAP and the standard deviation of $\beta|D$.

The posterior distribution of β, given the dataset D is given by Bayes' therorem,

$$p(\beta|D) = \frac{p(y_0|\beta)\, p(\beta)}{p(y = y_0)}.$$

Since $\epsilon \sim \mathcal{N}(0, \sigma^2)$, $p(y_0|\beta, I) = \mathcal{N}(y_0|\beta x, \sigma^2)$ and $p(\beta) \propto 1$, it follows that

$$p(\beta|D) \propto \exp\left(-\frac{1}{2\sigma^2}(y_0 - \beta x_0)^2\right),$$

or equivalently, by reordering the terms in the exponential,

$$p(\beta|D) \propto \exp\left(\frac{x_0^2}{2\sigma^2}\left(\beta - \frac{y_0}{x_0}\right)^2\right).$$

We can conclude that the posterior distribution is the normal distribution $\mathcal{N}\left(\frac{y_0}{x_0}, \frac{\sigma^2}{x_0^2}\right)$. And hence the MAP of β is $\hat{\beta}_{MAP} = \frac{y_0}{x_0}$ and the standard deviation of $p(\beta|D)$ is $\frac{\sigma}{x_0}$.

b) [20pts] Now, consider a Gaussian prior for β with mean 0 and variance τ^2, i.e. $\beta \sim \mathcal{N}(0, \tau^2)$, identify the posterior distribution $p(\beta|D)$. This form of regression is also known as Bayesian linear regression.

With $\epsilon \sim \mathcal{N}(0, \sigma^2)$, $p(y_0|\beta, I) = \mathcal{N}(y_0|\beta x, \sigma^2)$ and $p(\beta) = \mathcal{N}(0|\tau^2)$ and Bayes' formula we obtain

$$p(\beta|D) \propto \exp\left(-\frac{1}{2\sigma^2}(y_0 - \beta x_0)^2 - \frac{1}{2\tau^2}\beta^2\right)$$

Now through reordering the terms in the exponent we again obtain the form of a normal distribution:

$$\begin{aligned} p(\beta|D) &\propto \exp\left(-\frac{1}{2\sigma^2}(y_0^2 - 2y_0\beta x_0 + \beta^2(x_0^2 + \frac{\sigma^2}{\tau^2}))\right), \\ &\propto \exp\left(-\frac{x_0^2 + \frac{\sigma^2}{\tau^2}}{2\sigma^2}(\frac{1}{x_0^2 + \frac{\sigma^2}{\tau^2}}y_0^2 - \frac{2}{x_0^2 + \frac{\sigma^2}{\tau^2}}y_0\beta x_0 + \beta^2)\right), \\ &\propto \exp\left(-\frac{x_0^2 + \frac{\sigma^2}{\tau^2}}{2\sigma^2}(\frac{x_0 y_0}{x_0^2 + \frac{\sigma^2}{\tau^2}} - \beta)^2\right). \end{aligned}$$

And hence $p(\beta|D) = \mathcal{N}\left(\frac{x_0 y_0}{x_0^2 + \frac{\sigma^2}{\tau^2}}, \frac{\sigma^2}{x_0^2 + \frac{\sigma^2}{\tau^2}}\right)$. Note that in the last step we omitted terms not depending on β.

Guidelines for reports submissions:

- Submit a pdf file of your solution via Moodle until March 09, 2020, 08:00am.