Machine Learning Exercise Sheet 3

Probabilistic Inference

Exercise sheets consist of two parts: homework and in-class exercises. You solve the homework exercises on your own or with your registered group and upload it to Moodle for a possible grade bonus. The in-class exercises will be solved and discussed during the tutorial along with some difficult and/or important homework exercises. You do not have to upload any solutions of the in-class exercises.

In-class Exercises

Consider the probabilistic model

$$\begin{aligned}
\mathrm{p}(\mu \mid \alpha) &= \mathcal{N}(\mu \mid 0, \alpha^{-1}) = \sqrt{\frac{\alpha}{2\pi}} \exp\left(-\frac{\alpha}{2}\mu^2\right) \\
\mathrm{p}(x \mid \mu) &= \mathcal{N}(x \mid \mu, 1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x-\mu)^2\right)
\end{aligned}$$

and a set of observations $\mathcal{D} = \{x_1, ..., x_N\}$ consisting of N samples $x_i \in \mathbb{R}$.

Note: We parametrize $\mu \mid \alpha$ with the *precision* parameter $\alpha = 1/\sigma^2$ instead of the usual variance σ^2 because it leads to a nicer solution.

Problem 1: Derive the maximum likelihood estimate μ_{MLE}. Show your work.

Our goal is to find

$$\begin{aligned}
\mu_{\mathrm{MLE}} &= \underset{\mu \in \mathbb{R}}{\arg\max}\, \mathrm{p}(\mathcal{D} \mid \mu) \\
&= \underset{\mu \in \mathbb{R}}{\arg\max}\; \log \mathrm{p}(\mathcal{D} \mid \mu)
\end{aligned}$$

We solve this problem in two steps:

1. Write down & simplify the expression for $\log \mathrm{p}(\mathcal{D} \mid \mu)$.
2. Solve $\frac{\partial}{\partial \mu} \log \mathrm{p}(\mathcal{D} \mid \mu) \stackrel{!}{=} 0$ for μ.

$$
\begin{aligned}
\log p(\mathcal{D} \mid \mu) &= \log \mathrm{p}(x_1, ..., x_N \mid \mu) \\
&= \log \left(\prod_{i=1}^{N} \mathrm{p}(x_i \mid \mu) \right) && \text{iid assumption} \\
&= \sum_{i=1}^{N} \log \mathrm{p}(x_i \mid \mu) \\
&= \sum_{i=1}^{N} \left[\log \left(\frac{1}{\sqrt{2\pi}} \right) + \log \left(\exp \left(-\frac{1}{2}(x_i - \mu)^2 \right) \right) \right] \\
&= \sum_{i=1}^{N} \left[-\frac{1}{2}(x_i - \mu)^2 \right] + \text{const.} \\
&= -\frac{1}{2} \sum_{i=1}^{N} (x_i^2 - 2x_i\mu + \mu^2) + \text{const.} \\
&= \left[-\frac{1}{2} \sum_{i=1}^{N} x_i^2 \right] + \left[\sum_{i=1}^{N} x_i\mu \right] - \left[\frac{1}{2} \sum_{i=1}^{N} \mu^2 \right] + \text{const.} \\
&= \mu \sum_{i=1}^{N} x_i - \frac{N}{2}\mu^2 + \text{const.}
\end{aligned}
$$

Now compute the derivative and set it to zero.

$$
\begin{aligned}
\frac{\partial}{\partial \mu} \log \mathrm{p}(\mathcal{D} \mid \mu) &= \frac{\partial}{\partial \mu} \left(\mu \sum_{i=1}^{N} x_i - \frac{N}{2}\mu^2 + \text{const.} \right) \\
&= \sum_{i=1}^{N} x_i - N\mu \overset{!}{=} 0
\end{aligned}
$$

Solving for μ we obtain

$$
\mu_{\text{MLE}} = \frac{1}{N} \sum_{i=1}^{N} x_i
$$

That is, μ_{MLE} is just the average of the datapoints.

Problem 2: Derive the maximum a posteriori estimate μ_{MAP}. Show your work.

Our goal is to find

$$\begin{aligned}\mu_{\text{MAP}} &= \underset{\mu \in \mathbb{R}}{\arg\max}\, \mathrm{p}(\mu \mid \mathcal{D}, \alpha) \\ &= \underset{\mu \in \mathbb{R}}{\arg\max} \log \mathrm{p}(\mu \mid \mathcal{D}, \alpha) \\ &= \underset{\mu \in \mathbb{R}}{\arg\max} \left[\log \mathrm{p}(\mathcal{D} \mid \mu) + \log \mathrm{p}(\mu \mid \alpha)\right]\end{aligned}$$

We solve this problem in two steps:

1. Write down & simplify the expression for $\log \mathrm{p}(\mathcal{D} \mid \mu) + \log \mathrm{p}(\mu \mid \alpha)$.
2. Solve $\frac{\partial}{\partial \mu}\left(\log \mathrm{p}(\mathcal{D} \mid \mu) + \log \mathrm{p}(\mu \mid \alpha)\right) \stackrel{!}{=} 0$ for μ.

$$\begin{aligned}\log \mathrm{p}(\mu \mid \alpha) &= \log\left(\sqrt{\frac{\alpha}{2\pi}}\right) + \log\left(\exp\left(-\frac{\alpha}{2}\mu^2\right)\right) \\ &= -\frac{\alpha}{2}\mu^2 + \text{const.}\end{aligned}$$

From the previous task, we know that

$$\log \mathrm{p}(\mathcal{D} \mid \mu) = \mu \sum_{i=1}^{N} x_i - \frac{N}{2}\mu^2 + \text{const.}$$

Therefore, we get

$$\log \mathrm{p}(\mathcal{D} \mid \mu) + \log \mathrm{p}(\mu \mid \alpha) = \mu \sum_{i=1}^{N} x_i - \frac{N}{2}\mu^2 - \frac{\alpha}{2}\mu^2 + \text{const.}$$

Now compute the derivative and set it to zero.

$$\begin{aligned}\frac{\partial}{\partial \mu}\left(\log \mathrm{p}(\mathcal{D} \mid \mu) + \log \mathrm{p}(\mu \mid \alpha)\right) &= \frac{\partial}{\partial \mu}\left(\mu \sum_{i=1}^{N} x_i - \frac{N}{2}\mu^2 - \frac{\alpha}{2}\mu^2 + \text{const.}\right) \\ &= \sum_{i=1}^{N} x_i - N\mu - \alpha\mu \stackrel{!}{=} 0\end{aligned}$$

Solving for μ we obtain

$$\mu_{\text{MAP}} = \frac{1}{N+\alpha} \sum_{i=1}^{N} x_i$$

By comparing this to μ_{MLE}, we can understand the effect of a 0-mean Gaussian prior on our estimate of μ. Since $\alpha > 0$, we see that μ_{MAP} is always closer to zero than μ_{MLE}.

Problem 3: Does there exist a prior distribution over μ such that $\mu_{\text{MLE}} = \mu_{\text{MAP}}$? Justify your answer.

Let's compare the expressions for μ_{MLE} and μ_{MAP}

$$\mu_{\text{MLE}} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad\qquad \mu_{\text{MAP}} = \frac{1}{N+\alpha}\sum_{i=1}^{N} x_i$$

As α approaches zero ($\alpha \to 0$), μ_{MAP} gets closer to μ_{MLE}. As the *precision* of the prior distribution *decreases*, its *variance increases*. The prior distribution is getting more and more flat, thus being less informative and having a smaller effect on the posterior.

If we could set $\alpha = 0$, we would have a uniform prior on μ, and thus $\mu_{\text{MLE}} = \mu_{\text{MAP}}$. However, technically, we are not allowed to do that — since the distribution $\mathrm{p}(\mu \mid \alpha)$ is defined over all of $\mathbb{R}$, it has to integrate to one ($\int_{-\infty}^{\infty} \mathrm{p}(\mu \mid \alpha) d\mu = 1$).

We can ignore this restriction and assume that we have a uniform prior over μ. Such prior would be called *improper*. While in many cases it's fine to use an improper prior, it might lead to subtle problems in certain situations.

Problem 4: Derive the posterior distribution $\mathrm{p}(\mu \mid \mathcal{D}, \alpha)$. Show your work.

We obtain the posterior distribution using Bayes formula

$$\begin{aligned}
\mathrm{p}(\mu \mid \mathcal{D}, \alpha) &= \frac{\mathrm{p}(\mathcal{D} \mid \mu)\,\mathrm{p}(\mu \mid \alpha)}{\mathrm{p}(\mathcal{D} \mid \alpha)} \\
&\propto \mathrm{p}(\mathcal{D} \mid \mu)\,\mathrm{p}(\mu \mid \alpha) \\
&\propto \left(\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_i - \mu)^2\right)\right) \sqrt{\frac{\alpha}{2\pi}} \exp\left(-\frac{\alpha}{2}\mu^2\right) \\
&\propto \left(\prod_{i=1}^{N} \exp\left(-\frac{1}{2}(x_i - \mu)^2\right)\right) \exp\left(-\frac{\alpha}{2}\mu^2\right) \\
&\propto \exp\left(-\frac{1}{2}\sum_{i=1}^{N}(x_i - \mu)^2 - \frac{\alpha}{2}\mu^2\right) \\
&\propto \exp\left(-\frac{1}{2}\sum_{i=1}^{N} x_i^2 + \mu\sum_{i=1}^{N} x_i - \frac{1}{2}\sum_{i=1}^{N}\mu^2 - \frac{\alpha}{2}\mu^2\right) \\
&\propto \exp\left(-\frac{N+\alpha}{2}\mu^2 + \mu\sum_{i=1}^{N} x_i\right) \qquad (1)
\end{aligned}$$

We know that the posterior distribution has to integrate to 1, but we don't know the normalizing constant. However, we know that it's proportional to $\exp(a\mu^2 + b\mu)$. This looks very similar to a normal distribution — we have an quadratic form inside the exponential.

How can we use this fact? Consider a normal distribution over μ with mean m and precision β

$$\begin{aligned}\mathcal{N}(\mu \mid m, \beta^{-1}) &= \sqrt{\frac{\beta}{2\pi}} \exp\left(-\frac{\beta}{2}(\mu - m)^2\right) \\ &\propto \exp\left(-\frac{\beta}{2}\mu^2 + \beta m \mu\right) \qquad (2)\end{aligned}$$

If we find β and m such that Equations 1 and 2 are equal, we will know that our posterior $\mathrm{p}(\mu \mid \mathcal{D}, \alpha)$ is a normal distribution with mean m and precision β.

First we observe that

$$\beta = N + \alpha$$

Now we need to find m such that

$$\begin{aligned}\beta m &= \sum_{i=1}^{N} x_i \\ m &= \frac{1}{\beta} \sum_{i=1}^{N} x_i \\ m &= \frac{1}{N + \alpha} \sum_{i=1}^{N} x_i\end{aligned}$$

Putting everything together we see that

$$\mathrm{p}(\mu \mid \mathcal{D}, \alpha) = \mathcal{N}\left(\mu \,\middle|\, \frac{1}{N + \alpha} \sum_{i=1}^{N} x_i, (N + \alpha)^{-1}\right)$$

Since the posterior is a normal distribution, its mean coincides with its mode — this means that $\mathbb{E}_{\mathrm{p}(\mu|\mathcal{D},\alpha)}[\mu] = \mu_{\mathrm{MAP}}$. We can see that this is indeed the case, which is a good sanity check.

Problem 5: Derive the posterior predictive distribution $\mathrm{p}(x_{new} \mid \mathcal{D}, \alpha)$. Show your work.

The posterior over μ is $\mathrm{p}(\mu \mid \mathcal{D}, \alpha) = \mathcal{N}(\mu \mid m, \beta^{-1})$. Our goal is to find the *posterior predictive* distribution over the next sample $\mathrm{p}(x_{new} \mid \mathcal{D}, \alpha)$. For brevity, we will drop the $_{new}$ subscript.

From the lecture we remember that thanks to the conditional independence assumption the posterior predictive is

$$\mathrm{p}(x \mid \mathcal{D}, \alpha) = \int_{-\infty}^{\infty} \mathrm{p}(x \mid \mu)\, \mathrm{p}(\mu \mid \mathcal{D}, \alpha) d\mu$$

There are two (equivalent) ways to approach this problem.

Approach 1. Basically, we are modeling the following process

- We draw μ from the posterior distribution $\mu \sim \mathcal{N}(m, \beta^{-1})$.
- We draw x from the conditional distribution $x \sim \mathcal{N}(\mu, 1)$.

This process is identical to the following procedure

- We draw μ from the posterior distribution $\mu \sim \mathcal{N}(m, \beta^{-1})$.
- We draw y from the standard normal distribution $y \sim \mathcal{N}(0, 1)$.
- We calculate x as $\mu + y$.

Clearly, x is a sum of two *independent* normally distributed random variables. Hence, x also follows a normal distribution with mean $m + 0$ and precision $(\beta^{-1} + 1)^{-1}$.

$$\mathrm{p}(x \mid \mathcal{D}, \alpha) = \mathcal{N}(x \mid m, \beta^{-1} + 1)$$

where m and β were computed in the previous problem.

Approach 2. We can directly look at the integral

$$\begin{aligned}
\mathrm{p}(x \mid \mathcal{D}, \alpha) &= \int_{-\infty}^{\infty} \mathrm{p}(x \mid \mu)\, \mathrm{p}(\mu \mid \mathcal{D}, \alpha) d\mu \\
&= \int_{-\infty}^{\infty} \mathcal{N}(x \mid \mu, 1) \mathcal{N}(\mu \mid m, \beta^{-1}) d\mu \\
&= \int_{-\infty}^{\infty} \mathcal{N}(x - \mu \mid 0, 1) \mathcal{N}(\mu \mid m, \beta^{-1}) d\mu
\end{aligned}$$

This a convolution of two Gaussian densities — the result is a Gaussian density as well

$$= \mathcal{N}(x \mid m, \beta^{-1} + 1)$$

You can find the proof on Wikipedia `https://en.wikipedia.org/wiki/Sum_of_normally_distributed_random_variables#Proof_using_convolutions`.

The two approaches are effectively identical, and both rely on two facts:

1. μ is the location parameter of the normal distribution. That means that if $\mathrm{p}(x) = \mathcal{N}(x \mid \mu, \sigma^2)$ and $y = x + a$ (for a fixed $a \in \mathbb{R}$), then $\mathrm{p}(y) = \mathcal{N}(y \mid \mu + a, \sigma^2)$.
2. the sum of two normally distributed RVs is a normally distributed RV