Machine Learning Exercise Sheet 3

Probabilistic Inference

Exercise sheets consist o[illegible]k and in-class exercises. You solve the homework exercises on your own or with you[illegible]d upload it to Moodle for a possible grade bonus. The in-class exercises will be sol[illegible]ing 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

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

$$p(x \mid \mu) = \mathcal{N}(x \mid \mu, 1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x-\mu)^2\right)$$

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

Note: We parametrize $p(\mu \mid \alpha)$ 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

$$\begin{aligned}\mu_{\text{MLE}} &= \arg\max_{\mu \in \mathbb{R}} p(\mathcal{D} \mid \mu) \\ &= \arg\max_{\mu \in \mathbb{R}} \log p(\mathcal{D} \mid \mu)\end{aligned}$$

We solve this problem in two steps:

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

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

Now compute the derivative and set it to zero.

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

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

$$\begin{aligned}\mu_{\text{MAP}} &= \arg\max_{\mu} \mathrm{p}(\mu \mid \mathcal{D}, \alpha) \\ &= \arg\max_{\mu} \log \mathrm{p}(\mu \mid \mathcal{D}, \alpha) \\ &= \arg\max_{\mu} [\log \mathrm{p}(\mathcal{D} \mid \mu) + \log \mathrm{p}(\mu \mid \alpha)]\end{aligned}$$

We solve this problem [illegible]

1. Write down & si[illegible] for $\log \mathrm{p}(\mathcal{D} \mid \mu) + \log \mathrm{p}(\mu \mid \alpha)$.
2. Solve $\frac{\partial}{\partial \mu} (\log \mathrm{p}(\mathcal{D} \mid \mu) + \log \mathrm{p}(\mu \mid \alpha)) \stackrel{!}{=} 0$ 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.

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

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}

μ_{M} [illegible]

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

As α approaches zero [illegible]loser to μ_{MLE}. As the *precision* of the prior distribution *decreases*, its *variance* [illegible]listribution is getting more and more flat, thus being less informative and having [illegible]he posterior.

If we could set $\alpha =$ [illegible]niform 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.

Upload a single PDF file with your homework solution to Moodle by 10.11.2021, 23:59 CET. We recommend to typeset your solution (using LaTeX or Word), but handwritten solutions are also accepted. If your handwritten solution is illegible, it won't be graded and you waive your right to dispute that.

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

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

If we find β and m suc[illegible]d 2 are equal, we will know that our posterior $\mathrm{p}(\mu \mid \mathcal{D}, \alpha)$ is a normal distributio[illegible]precision β.

First we observe that

$$\beta = N + \alpha$$

Now we need to find m such that

$$\beta m = \sum_{i=1}^{N} x_i$$

$$m = \frac{1}{\beta}\sum_{i=1}^{N} x_i$$

$$m = \frac{1}{N+\alpha}\sum_{i=1}^{N} x_i$$

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 a standard normal distribution $y \sim \mathcal{N}(0, 1)$.
- We calculate x as $x = \mu + y$.

Clearly, x is a sum of two 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

Homework

Optimizing Likelihoods: Monotonic Transforms

Usually we maximize the *log-likelihood*, $\log \mathrm{p}(x_1, \ldots, x_n \mid \theta)$ instead of the likelihood. The next two problems provide a justification for this.

In the lecture, we encountered the likelihood maximization problem

$$\arg\max_{\theta \in [0,1]} \theta^t (1-\theta)^h,$$

where t and h denoted the number of tails and heads in a sequence of coin tosses, respectively.

Problem 6: Compute the first and second derivative of this likelihood w.r.t. θ. Then compute first and second derivative of the log-likelihood $\log \theta^t(1-\theta)^h$.

To solve this, we need to apply the product rule.

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\theta^t(1-\theta)^h = \theta^{t-1}(1-\theta)^{h-1}((1-\theta)\,t-\theta h)$$

$$\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\theta^t(1-\theta)^h = \theta^{t-2}(1-\theta)^{h-2}\cdot((1-\theta)\,(t-1)-\theta(h-1))\cdot((1-\theta)\,t-\theta h)-\theta^{t-1}(1-\theta)^{h-1}\,(t+h)$$

Without the logarithm, the product rule leads to an explosion of terms because it introduces more terms that themselves contain products which multiply again in the next derivative. This quickly renders the expressions long and confusing.

The logarithm decomposes products into sums and we only need to take the derivative of each of the summands. Do not forget the change of sign from taking the derivative of $1-\theta$.

$$g(\theta) := \log \theta^t(1-\theta)^h = t\log\theta + h\log(1-\theta)$$

$$\frac{\mathrm{d}}{\mathrm{d}\theta}g(\theta) = \frac{t}{\theta} - \frac{h}{1-\theta}$$

$$\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}g(\theta) = -\left(\frac{t}{\theta^2} + \frac{h}{(1-\theta)^2}\right)$$

Problem 7: Show that for *any* differentiable, positive function $f(\theta)$ every local maximum of $\log f(\theta)$ is also a local maximum of $f(\theta)$. Considering this and the previous exercise, what is your conclusion?

Let θ^* be an arbitrary local maximum of $g(\theta) = \log f(\theta)$, i.e., for any θ in a small neighborhood of θ^*, we have that $g(\theta^*) \geq g(\theta)$. Since exp is a monotonic transform, we also have

$$f(\theta^*) = \exp(g(\theta^*)) \geq \exp(g(\theta)) = f(\theta).$$

Hence, θ^* is also a maximum of f.

With the help of the previous exercise, we can now safely apply the logarithm and any maximum or minimum remains preserved (its position only, of course). Moreover, we have seen that the logarithmic domain can greatly simplify the computational effort to arrive at critical points. This also leads to improved numerical stability. Thus, it is often worth switching to the log domain when analyzing likelihoods.

Notice that the exercise left out a part of the argument: We only showed that a maximum of the log likelihood is also a maximum of the likelihood. We would still need to prove that taking the logarithm does not eliminate maxima of the likelihood. This is done by showing that monotonic transforms preserve critical points and observing that the logarithm is monotonic—we will not do this here.

Properties of MLE [illegible]

Problem 8: Consider a [illegible]ariable X and suppose we have observed m occurrences of $X = 1$ and l occurrence[illegible]uence of $N = m + l$ Bernoulli experiments. We are only interested in the number [illegible] $= 1$—we will model this with a Binomial distribution with parameter θ. A prior dis[illegible]n by the Beta distribution with parameters a, b. Show that the posterior *mean* value $\mathbb{E}[\theta \mid \mathcal{D}]$ (not the MAP estimate) of θ lies between the prior mean of θ and the maximum likelihood estimate for θ.

To do this, show that the posterior mean can be written as λ times the prior mean plus $(1-\lambda)$ times the maximum likelihood estimate, with $0 \le \lambda \le 1$. This illustrates the concept of the posterior mean being a compromise between the prior distribution and the maximum likelihood solution.

The probability mass function of the Binomial distribution for some $m \in \{0, 1, \ldots, N\}$ is

$$\mathrm{p}(x = m \mid N, \theta) = \binom{N}{m} \theta^m (1-\theta)^{N-m}.$$

Hint: Identify the posterior distribution. You may then look up the mean rather than computing it.

We have the following random variables.

$$\theta \sim \mathrm{Beta}(a, b) \propto \theta^{a-1}(1-\theta)^{b-1}$$
$$x = m \mid \theta \sim \mathrm{Binom}(N, \theta) \propto \theta^m (1-\theta)^{N-m}$$

The posterior $\theta \mid x$ is

$$\mathrm{p}(\theta \mid x) = \frac{\mathrm{p}(x \mid \theta)\, \mathrm{p}(\theta)}{\mathrm{p}(x)} \propto \mathrm{p}(x \mid \theta)\, \mathrm{p}(\theta) \propto \theta^{m+a-1}(1-\theta)^{(N-m)+b-1}.$$

This is the form of an unnormalized Beta distribution and we conclude

$$\theta \mid x \sim \mathrm{Beta}(m + a, l + b) \text{ where } l = N - m.$$

We can now look up the posterior mean of θ as the mean of this Beta distribution:

$$\mathbb{E}[\theta \mid \mathcal{D}] = \frac{m+a}{m+l+a+b} = \frac{m}{m+l+a+b} + \frac{a}{m+l+a+b}$$

But:

$$\frac{m}{m+l+a+b} = \underbrace{\frac{m+l}{m+l+a+b}}_{1-\lambda} \cdot \frac{m}{m+l}$$

and

$$\frac{a}{m+l+a+b} = \underbrace{\frac{a+b}{m+l+a+b}}_{\lambda} \cdot \frac{a}{a+b}$$

producing what was asked, because $\frac{m}{m+l}$ is the maximum likelihood estimate and $\frac{a}{a+b}$ is the prior mean value of θ.

Problem 9: Consider the following probabilistic model

$$\operatorname{Gamma}(\lambda \mid a, b) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} \exp(-b\lambda)$$

$$\operatorname{Poisson}(x \mid \lambda) = \frac{\lambda^x \exp(-\lambda)}{x!}$$

where $a \in (1, \infty)$ and $b \in (0, \infty)$. We have observed a single data point $x \in \mathbb{N}$. Derive the maximum a posteriori (MAP) estimate of the parameter λ for the above probabilistic model. Show your work.

The MAP estimate of λ is defined as

$$\begin{aligned}
\lambda_{\text{MAP}} &= \arg\max_{\lambda} \mathrm{p}(\lambda \mid x, a, b) \\
&= \arg\max_{\lambda} \log \mathrm{p}(\lambda \mid x, a, b) \\
&= \arg\max_{\lambda} \left(\log \mathrm{p}(x \mid \lambda) + \log \mathrm{p}(\lambda \mid a, b)\right) \\
&= \arg\min_{\lambda} -\left(\log \mathrm{p}(x \mid \lambda) + \log \mathrm{p}(\lambda \mid a, b)\right) \\
&= \arg\min_{\lambda} \left(-x \log(\lambda) + \lambda + \log x! - a \log b + \log \Gamma(a) - (a-1)\log \lambda + b\lambda\right) \\
&= \arg\min_{\lambda} \left((1 - x - a)\log(\lambda) + (b+1)\lambda\right)
\end{aligned}$$

This a convex function of λ. To minimize it, compute the derivative

$$\frac{\partial}{\partial \lambda}\left((1 - x - a)\log(\lambda) + (b+1)\lambda\right) = \frac{1 - x - a}{\lambda} + b + 1$$

and find its root in λ.

$$\frac{1 - x - a}{\lambda} + b + 1 \overset{!}{=} 0 \iff \frac{x + a - 1}{\lambda} = b + 1 \iff \lambda = \frac{x + a - 1}{b + 1}$$

Therefore, $\lambda_{\text{MAP}} = \frac{x+a-1}{b+1}$.

Programming Task

Problem 10: Download the notebook `exercise_03_notebook.ipynb` from Moodle. Fill in the missing code and follow the instructions in the notebook to append the solution to your PDF submission.

Note: We suggest that you use Anaconda for installing Python and Jupyter, as well as for managing packages. We recommend that you use Python 3.

Upload a single PDF file with your homework solution to Moodle by 10.11.2021, 23:59 CET. We recommend to typeset your solution (using LaTeX or Word), but handwritten solutions are also accepted. If your handwritten solution is illegible, it won't be graded and you waive your right to dispute that.

For more information on Jupyter notebooks, consult the Jupyter documentation. Instructions for converting the Jupyter notebooks to PDF are provided within the notebook.

Upload a single PDF file with your homework solution to Moodle by 10.11.2021, 23:59 CET. We recommend to typeset your solution (using LaTeX or Word), but handwritten solutions are also accepted. If your handwritten solution is illegible, it won't be graded and you waive your right to dispute that.