Machine Learning – SS 2024 – Exercise Sheet 3

Probabilistic Inference and Evaluation

Deadline: 29. Apr. 2024, 23:55 (CET)

Important. Upload a single PDF file and a single .ipynb notebook with your homework solution to ILIAS by 29. Apr. 2024, 23:55 (CET). We recommend typesetting your solution (using LaTeX or Word), but handwritten solutions are also accepted. If your handwritten solution is illegible, it wont be graded and you waive your right to dispute that. You only need to submit homework problems. The in-class problems will be discussed in the tutorial but will not be graded.

Policy. Any resources that you use to solve the homework must be cited:

- When taking inspiration from online forums (like StackOverflow), you should include the URL where you found the information;
- If using large language models (like ChatGPT) for formulation/editing your answers, you should either provide the link when available (ChatGPT supports this feature) or detail how you did use it, whether for rephrasing, explaining an unclear concept, or seeking assistance in proving a point, among other uses.
- When taking inspiration from previous years' homeworks, you should include both the number of the exercise and the year of the homework along with how you did use the solution, e.g. "analogous exercise from the 2023 homework 4 ex. 2, differently from that we [...]. Similarly, to the exercise we [...]".
- If you discuss the homework with another team, you should include the name of the team and the members you discussed with – note that this does not mean that the solutions should be the same, just that you discussed the homework together. Write your own solution.

Homework Problems

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

Problem 2. 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?

Properties of MLE and MAP

Problem 3. Show that θ_{MLE} can be interpreted as a special case of θ_{MAP} in the sense that there always exists a prior $\mathrm{p}(\theta)$ such that $\theta_{\text{MLE}} = \theta_{\text{MAP}}$.

Problem 1:

$$\theta^t(1-\theta)^h$$

First derivate: $\frac{d}{d\theta}\left[\theta^t(1-\theta)^h\right] = \frac{d}{d\theta}\left[\theta^t\right]\cdot(1-\theta)^h + \theta^t\cdot\frac{d}{d\theta}\left[(1-\theta)^h\right] = t\,\theta^{t-1}\cdot(1-\theta)^h - h(1-\theta)^{h-1}\cdot\theta^t$

Second derivate: $\frac{d}{d\theta}\left[t\,\theta^{t-1}\cdot(1-\theta)^h - h(1-\theta)^{h-1}\cdot\theta^t\right] = t\cdot\frac{d}{d\theta}\left[\theta^{t-1}\cdot(1-\theta)^h\right] - h\cdot\frac{d}{d\theta}\left[\theta^t\cdot(1-\theta)^{h-1}\right]$

$= t\left(\frac{d}{d\theta}\left[(1-\theta)^h\right]\cdot\theta^{t-1} + (1-\theta)^h\frac{d}{d\theta}\left[\theta^{t-1}\right]\right) - h\left(\frac{d}{d\theta}\left[(1-\theta)^{h-1}\right]\cdot\theta^t + (1-\theta)^{h-1}\frac{d}{d\theta}\left[\theta^t\right]\right)$

$= t\left((t-1)(1-\theta)^h\theta^{t-2} - h(1-\theta)^{h-1}\theta^{t-1}\right) - h\left(t(1-\theta)^{h-1}\cdot\theta^{t-1} - (h-1)(1-\theta)^{h-2}\cdot\theta^t\right)$

$= t(t-1)(1-\theta)^h\theta^{t-2} - ht(1-\theta)^{h-1}\cdot\theta^{t-1} - ht(1-\theta)^{h-1}\theta^{t-1} + h(h-1)(1-\theta)^{h-2}\theta^t$

$= t(t-1)(1-\theta)^h\theta^{t-2} - 2ht(1-\theta)^{h-1}\cdot\theta^{t-1} + h(h-1)(1-\theta)^{h-2}\theta^t$

log $\theta^t(1-\theta)^h$

First derivate: $\frac{1}{(1-\theta)^h\theta^t}\cdot\frac{d}{d\theta}\left[\theta^t\cdot(1-\theta)^h\right] = \frac{\frac{d}{d\theta}\left[\theta^t\right]\cdot(1-\theta)^h + \theta^t\cdot\frac{d}{d\theta}\left[(1-\theta)^h\right]}{(1-\theta)^h\theta^t}$

$= \frac{t\cdot\theta^{t-1}\cdot(1-\theta)^h - \theta^t\cdot h(1-\theta)^{h-1}}{(1-\theta)^h\theta^t}$

$= \frac{\theta^t\cdot(1-\theta)^h\cdot\left(t\cdot\theta^{-1} - h(1-\theta)^{-1}\right)}{(1-\theta)^h\theta^t}$

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

Second derivate: $\frac{d}{d\theta}\left[\frac{t}{\theta} - \frac{h}{(1-\theta)}\right] = \frac{d}{d\theta}\left[\frac{t}{\theta}\right] - \frac{d}{d\theta}\left[\frac{h}{(1-\theta)}\right]$

$= t\cdot\frac{d}{d\theta}\left[\frac{1}{\theta}\right] - h\cdot\frac{d}{d\theta}\left[\frac{1}{(1-\theta)}\right]$

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

$= -\frac{t}{\theta^2} - \frac{h}{(1-\theta)^2}$

Problem 4. Consider a Bernoulli random variable X and suppose we have observed m occurrences of $X = 1$ and l occurrences of $X = 0$ in a sequence of $N = m + l$ Bernoulli experiments. We are only interested in the number of occurrences of $X = 1$—we will model this with a Binomial distribution with parameter θ. A prior distribution for θ is given 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 $0 \leq \lambda \leq 1$. 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.

Problem 5. Consider the following probabilistic model

$$\mathrm{p}(\lambda \mid a, b) = \mathrm{Gamma}(\lambda \mid a, b) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} \exp(-b\lambda)$$
$$\mathrm{p}(x \mid \lambda) = \mathrm{Poisson}(x \mid \lambda) = \frac{\lambda^x \exp(-\lambda)}{x!}$$

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.

Evaluation

Problem 6. You are given a dataset `[1, 4, 2, 4, 3, 3, 4, 3, 2, 2]` that represents the test scores of a sample of students who were taught using a new teaching method. You want to compute the 95% confidence interval of the mean of the dataset, the standard error of the mean, and the standard deviation of the dataset. *Recall that the t-value for 95% confidence interval is 1.96.*

Programming problems

Problem 7. Download `exercise_03_notebook.ipynb` from ILIAS. Fill in the missing code and run the notebook. Use the given docstring of each function as a hint on how you should implement the function and what is going to be the input and the output. Upload the final `.ipynb` along with the PDF to ILIAS.

In-Class Problems

Consider the probabilistic model

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

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 8. Derive the maximum likelihood estimate μ_{MLE}. Show your work.

Problem 9. Derive the maximum a posteriori estimate μ_{MAP}. Show your work.

Problem 10. Derive the posterior distribution $\mathrm{p}(\mu \mid \mathcal{D}, \alpha)$. Show your work.

Problem 11. Derive the posterior predictive distribution $\mathrm{p}(x_{new} \mid \mathcal{D}, \alpha)$. Show your work.

Problem 6: $\bar{x} \pm 1.96 \cdot \frac{\sigma_x}{\sqrt{N}}$, $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$

$N = 10$

$\bar{x} = \frac{1}{10} \cdot (1 + 4 + 2 + 4 + 3 + 3 + 4 + 3 + 2 + 2) = \frac{28}{10} = 2.8$

$\sigma_x^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2 = \frac{1}{9} (3.24 + 1.44 + 0.64 + 1.44 + 0.04 + 0.04 + 1.44 + 0.04 + 0.64 + 0.64)$

$= 1.067$

$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{N}} = \frac{1.067}{\sqrt{10}} = 0.337$

CI: $\bar{x} \pm 1.96\, \sigma_{\bar{x}} = 2.8 \pm 1.96 \cdot 0.337$ ⇒ Upper boundary: 3.461

Lower boundary: 2.139