Machine Learning Exercise Sheet 3

Probabilistic Inference

Exercise sheets consist of two parts: In-class exercises and homework. The in-class exercises will be solved and discussed during the tutorial. The homework is for you to solve at home and further engage with the lecture content. There is no grade bonus and you do not have to upload any solutions. Note that the order of some exercises might have changed compared to last year's recordings.

In-class Exercises

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

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

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

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

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

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

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

P1

① $\mu_{MLE} = \arg\max_{\mu} \log p(D|\mu)$

$= \arg\max_{\mu} \log p(x_1 \cdots x_n | \mu)$

$= \arg\max_{\mu} \log \left(\prod_{i=1}^{n} p(x_i|\mu) \right)$

$= \arg\max_{\mu} \sum_{i=1}^{n} \left(\log p(x_i|\mu) \right)$

$= \arg\max_{\mu} \sum_{i=1}^{n} \left(\log \left(\frac{1}{\sqrt{2\pi}} \right) + \log \left(e^{(-\frac{1}{2}(x-\mu)^2)} \right) \right)$

$= \arg\max_{\mu} \sum_{i=1}^{n} \cdots + \sum_{i=1}^{n} \left(-\frac{1}{2}(x_i - \mu)^2 \right)$

$= \arg\max_{\mu} \underbrace{\sum_{i=1}^{n} \left(+ \sum_{i=1}^{n} \left(-\frac{1}{2} x_i^2 \right) + \sum_{i=1}^{n} \left(x_i \mu \right) + \sum_{i=1}^{n} \left(-\frac{1}{2}\mu^2 \right) \right)}_{\text{no } \mu}$

$= \arg\max_{\mu} \sum_{i=1}^{n} \left(+ \mu \sum_{i=1}^{N} x_i - \frac{N\mu^2}{2} \right)$

$\frac{\partial \arg\max_{\mu} f(D|\mu)}{\partial \mu} = \sum_{i=1}^{n} x_i - N\mu = 0$

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

② $\mu_{MAP} = \arg\max_{\mu} \log p(\mu|D,\alpha)$

$= \arg\max_{\mu} \log p(D|\mu) \cdot \underbrace{p(\mu|\alpha)}_{\text{prior.}}$

$= \arg\max_{\mu} \left(\log p(D|\mu) + \log p(\mu|\alpha) \right)$

$= \arg\max_{\mu} \sum_{i=1}^{n} \left(+ \mu \sum_{i=1}^{N} x_i - \frac{N\mu^2}{2} - \frac{\alpha}{2}\mu^2 \right)$

$\log p(\mu|\alpha) = \log \left(\sqrt{\frac{\alpha}{2\pi}} \right) + \log \left(e^{(-\frac{\alpha}{2}\mu^2)} \right)$

$= C - \frac{\alpha}{2}\mu^2$

$\frac{\partial \arg\max_{\mu} p(\mu|D,\alpha)}{\partial \mu} = \sum_{i=1}^{n} x_i - N\mu - \alpha\mu = 0$

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

③ μ ? $\Rightarrow \mu_{MLE} = \mu_{MAP}$.

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

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

When $\alpha = 0$, $\mu_{MLE} = \mu_{MAP}$

but $\alpha = \frac{1}{\sigma^2}$ σ^2 never equal to 0

so impossible

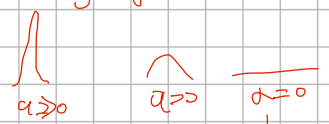
X.

① because $\int_{-\infty}^{\infty} p(\mu|\alpha) d\mu = 1$

② ingnore this limitation

We can use a uniform prior

α is very large



$\alpha \gg 0$ $\alpha \gg$ $\alpha = 0$

implies there is no assumption

④ $p(\mu|D,\alpha) = \frac{p(D|\mu)p(\mu|\alpha)}{p(D)}$

$\propto p(D|\mu) p(\mu|\alpha)$

$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x-\mu)^2 \right) \cdot \sqrt{\frac{\alpha}{2\pi}} \exp\left(-\frac{\alpha}{2}\mu^2 \right)$

$\propto \exp\left(-\frac{1}{2} \sum_{i=1}^{N} (x_i - \mu)^2 - \frac{\alpha}{2}\mu^2 \right)$

$= \exp\left(-\frac{N+\alpha}{2}\mu^2 + \mu \sum_{i=1}^{n} x_i \right)$

$N(x|m,\beta^{-1}) = \sqrt{\frac{\beta}{2\pi}} \exp\left(-\frac{\beta}{2}(x-\mu)^2 \right)$

$\propto \exp\left(-\frac{\beta}{2}u^2 + \beta m \mu \right)$

$\beta = N + \alpha$

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

$m = \frac{1}{\beta} \sum_{i=1}^{N} \alpha_n \Rightarrow m = \frac{1}{N+\alpha} \sum_{i=1}^{N} x_n$.

$p(\mu|x_1) = N\left(m, \frac{1}{\beta}\right) = N\left(\frac{1}{N+\alpha} \sum_{n=1}^{N} \alpha_n, \frac{1}{n+\alpha} \right)$

⑤ $p(x|D,\alpha) = \int_{-\infty}^{\infty} p(x|\mu) p(\mu|D,\alpha) d\mu$

$p(x|\mu) p(\mu|D,\alpha) d\mu$

(m) $p(x) = \int p(x|\mu) p(\mu) d\mu$.

(a)

$\mu \sim N(m, \beta^{-1})$

$x \sim N(\mu, 1)$ $\Rightarrow x = \mu + y \sim N(0+m, 1)$

$m \sim N(m, \beta^{-1})$

$y \sim N(0,1)$

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

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?

Properties of MLE and MAP

Problem 8: 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 9: 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.

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.

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.

For more information on Jupyter notebooks, consult the Jupyter documentation.

P6

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

② $\frac{d^2\cdots}{d\theta^2} = t\cdot(t-1)\cdot\theta^{t-2}\cdot(1-\theta)^n + t\theta^{t-1}\cdot n(1-\theta)^{n-1}$
$+ t\cdot\theta^{t-1}\cdot n(1-\theta)^{n-1} + \theta^t\cdot n\cdot(n-1)\cdot(1-\theta)^{n-2}$

③ $f(\theta) = \log\theta^t(1-\theta)^n = \log\theta^t + \log(1-\theta)^n = t\log\theta + n\log(1-\theta)$

$\frac{d f(\theta)}{d\theta} = \frac{t}{\theta} + \frac{n}{1-\theta}$

$\frac{d^2 f(\theta)}{d\theta^2} = -\frac{t}{\theta^2} - \frac{n}{(1-\theta)^2}$

P7 $g(x) = f(x)$ $h(x) = \log f(\theta)$

$\frac{dg(x)}{dx} = f'(x) = 0$ 时 $f(x)$ 取极值.

$\frac{dh(x)}{dx} = \frac{1}{f(x)} = 0$ 时 $f(x) \to \pm\infty$

$\Leftrightarrow \theta^*$ is maximum of $g(x)$

$\Rightarrow g(\theta^*) \geq g(\theta)$.

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

↑ also maximum

For log and no log, maximum and minimum is equal, so its easy to use log for calculation

P7 $X \sim Ber \quad \begin{cases} x=1, & m \\ x=0 & l \end{cases} \quad m+l=N$

prior Distribution $p(\theta) \sim Beta\ a,b$

posterior mean $E[\theta|D]$ lies between mean of θ and θ_{MLE}

$p(\theta) = Beta(\theta|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$

$p(X=m|\theta) = Binomial = \theta^m(1-\theta)^{N-m}$

$p(\theta|X) = \frac{p(X|\theta)\cdot p(\theta)}{p(X)} \propto \theta^m(1-\theta)^{N-m}\cdot\theta^{a-1}(1-\theta)^{b-1}$

$= \theta^{m+a+1}(1-\theta)^{N-m+b-1}$

$\Rightarrow p(\theta|x) \sim Beta(\theta|m+a, N-m+b)$

$\Rightarrow E(Beta) = \frac{m+a}{N+a+b} = \frac{m}{N+a+b} + \frac{a}{N+a+b}$.

mean of $p(\theta) = \frac{a}{a+b}$ $\theta_{MLE} = \frac{m}{N}$

$\Rightarrow \lambda\cdot\frac{a}{a+b} + (1-\lambda)\frac{m}{N} = \frac{m+a}{N+a+b}$

$\lambda = \frac{a+b}{N+a+b}$ $1-\lambda = \frac{N}{N+a+b}$

P8 $p(\lambda|x,a,b) = \frac{p(x|\lambda)\,p(\lambda|a,b)}{p(x)} \propto \underset{\lambda}{\operatorname{argmax}}\ p(x|\lambda)\,p(\lambda|a,b)$

(argmax 下标 λ)

$= \underset{\lambda}{\operatorname{argmax}}\ \log p(x|\lambda) + \log(\lambda|a,b)$

$= \underset{\lambda}{\operatorname{argmax}}\ \log\lambda^x + (-\lambda) + \cancel{\log(x!)^{-1}} + \log\frac{b^a}{\Gamma(a)} + \log\lambda^{a-1} - b\lambda$

$= \underset{\lambda}{\operatorname{argmax}}\ x(\log\lambda - \lambda + (a-1)\log\lambda - b\lambda + C$

$\frac{\partial -}{\partial\lambda} = \frac{x+a-1}{\lambda} - (b+1) = 0.$

$\lambda_{MAP} = \frac{x+a-1}{b+1}$