Machine Learning Exercise Sheet 3

Probabilistic Inference

Homework

Optimizing Likelihoods: Monotonic Transforms

Usually we maximize the *log-likelihood*, $\log p(x_1, \dots, x_n \mid \theta)$ 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 every local maximum of $\log f(\theta)$ is also a local maximum of the differentiable, positive function $f(\theta)$. Considering this and the previous exercise, what is your conclusion?

Properties of MLE and MAP

Problem 3: You model a coin flip f as a Bernoulli distribution with a parameter θ

$$p(f \mid \theta) = \mathrm{Bern}(f \mid \theta) = \theta^{\mathbb{I}[f=T]} (1-\theta)^{\mathbb{I}[f=H]}.$$

That is, the probability of landing tails (T) is θ, and probability of heads (H) is $(1-\theta)$ respectively.

Your prior on θ is a $\mathrm{Beta}(6,4)$ distribution

$$p(\theta) = \mathrm{Beta}(\theta \mid 6, 4).$$

You observe $(M+N)$ coin flips, out of which M are tails and N are heads. After you do maximum a posteriori estimation of θ, you obtain the result $\theta_{MAP} = 0.75$.

Name any possible values of M and N that can lead to such result. Show your work.

Upload a single PDF file with your homework solution to Moodle by 03.11.2019, 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.

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

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

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

Programming Task

Problem 5: Load the notebook `exercise_03_notebook.ipynb` from Piazza. Fill in the missing code and run the notebook. Export (download) the evaluated notebook as PDF and add it to your 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.

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

In-class Exercises

Consider the following probabilistic model

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

Note that here we don't parametrize the variance σ^2, but rather specify the *precision* parameter $\alpha = 1/\sigma^2$.

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

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

Problem 6: Derive the maximum likelihood estimate μ_{MLE}. Show your work.

Problem 7: Derive the maximum a posteriori estimate μ_{MAP}. Show your work.

Problem 8: Does there exist a prior distribution over μ such that $\mu_{\text{MLE}} = \mu_{\text{MAP}}$? Justify your answer.

Problem 9: Derive the posterior distribution $p(\mu \mid \mathcal{D}, \alpha)$. Show your work.

Problem 10: Derive the posterior predictive distribution $p(x_{new} \mid \mathcal{D}, \alpha)$. Show your work.