Ludwig-Maximilians-Universität München
Institut für Informatik
Prof. Dr. Matthias Schubert
Sebastian Schmoll, Sabrina Friedl

Munich, 26.11.2018

Deep Learning and Artificial Intelligence

WS 2018/19

Exercise 6: Model Uncertainty and LSTMs

Exercise 6-1 Variational Method: Evidence Lower Bound (ELBO)

Assume that D is a set of observations (data) and θ is a hidden variable (e.g. a parameter). According to Bayes' theorem, the posterior distribution of the hidden variable (after having observed D) can be written as:

$$P(\theta|D) = \frac{P(D,\theta)}{P(D)} = \frac{P(D|\theta) \cdot P(\theta)}{\int P(D,\theta)\, d\theta},$$

where $P(D|\theta)$ is the likelihood of the data and $P(\theta)$ is the prior (the probability distribution of θ before seeing any evidence). In many cases, the computation of the denominator $P(D)$ and thus of the the whole posterior is intractable. The idea behind the variational method is thus to find some easier distribution $Q(\theta)$ that approximates the true posterior distribution $P(\theta|D)$. A common metric to measure the closeness between two distributions is the Kullback-Leibler (KL) divergence:

$$KL(Q(\theta) \mid\mid P(\theta|D)) = \int Q(\theta) \log \frac{Q(\theta)}{P(\theta|D)}\, d\theta = -\int Q(\theta) \log \frac{P(\theta|D)}{Q(\theta)} d\theta \quad {}^{1}$$

.

(a) By dissecting the above term, show that

$$\log P(D) = KL(Q(\theta) \mid\mid P(\theta|D)) + L,$$

where $L = \int Q(\theta) \log \frac{P(\theta,D)}{Q(\theta)}\, d\theta$.

(b) The term L is called *evidence lower bound* or *variational lower bound*. Why is it a lower bound? When is L the same as $\log P(D)$?

[1] $\log(\frac{1}{x}) = \log(x^{-1}) = -\log(x)$

Exercise 6-2 Metropolis-Hastings Algorithm

Given a set D of iid (identical and independently distributed) samples $d_1, \ldots d_n$ that are distributed according to a normal distribution with mean μ and variance σ^2, i.e. , $\forall d_i \in D : d_i \sim \mathcal{N}(\mu, \sigma^2)$. The probability density function (pdf) of the normal distribution is given as $\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$. Suppose that we know σ^2 but want to infer the mean $\mu = \theta$ given a set of observations D.

(a) Calculate the likelihood $P(D|\theta)$!

(b) Let the prior $P(\theta)$ of the parameter θ be a standard normal distribution, i.e. $\theta \sim \mathcal{N}(\mu_p, \sigma_p)$ with $\mu_p = 0$ and $\sigma_p = 1$ and let $\sigma^2 = 1$ as well.

Calculate the posterior $P(\theta|D)$! *Hint:* Note that we chose $P(\theta)$ to be a conjugate prior [2], for which the posterior is also a normal distribution given by:

$$P(\theta|D) = \mathcal{N}(\theta|\mu_m, \sigma_m^2)$$

with

$$\begin{aligned}\mu_m &= \frac{\sigma^2}{n\sigma_p^2 + \sigma^2}\mu_p + \frac{n\sigma_p^2}{n\sigma_p^2 + \sigma^2}\left(\frac{1}{n}\sum_{i=1}^{n} d_i\right) \\ \frac{1}{\sigma_m^2} &= \frac{1}{\sigma_p^2} + \frac{n}{\sigma^2} \quad .\end{aligned}$$

(c) Let's assume we used a different prior distribution. What would change?

(d) Use the corresponding Jupyter notebook file from the lecture web-site to implement the analytic solution and the Metropolis-Hastings algorithm in Python. For more information, please consult the notebook.

Exercise 6-3 LSTM

In this exercise we will use Tensorflow to look inside a LSTM cell. We will train it to predict the next element of a simple time series (every n^{th} element is 1) and then consult the gates of the LSTM in all possible states of the sequence to see how the LSTM learns (what it forgets/remembers). Please download and open the corresponding Jupyter notebook from the lecture web-site and follow the instructions.

[2] For more details refer to the following link: Bishop - Pattern Recognition And Machine Learning, page 98.