Exercises for the course

Machine Learning 2

Summer semester 2020

Abteilung Maschinelles Lernen
Institut für Softwaretechnik und theoretische Informatik
Fakultät IV, Technische Universität Berlin
Prof. Dr. Klaus-Robert Müller
Email: klaus-robert.mueller@tu-berlin.de

Exercise Sheet 3

Exercise 1: Maximum Entropy Distributions (20 P)

The differential entropy $H(X)$ for a random variable $X \in \mathbb{R}$ with probability density function p is given by

$$H(X) = -\int_{-\infty}^{\infty} p(x) \, \log p(x) \, dx$$

We would like to find the probability density function p that maximizes the differential entropy under the constraints

$$\forall x \in \mathbb{R} : p(x) \geq 0 \quad , \quad \int_{-\infty}^{\infty} p(x) \, dx = 1 \quad , \quad \mathrm{E}[X] = 0 \quad , \quad \mathrm{Var}[X] = \sigma^2.$$

In the following, we approximate the integrals by a Riemanian sum over a regular partition $\mathcal{G}$ of $\mathbb{R}$. Each element of $\mathcal{G}$ is an interval of size δ represented by its center point x. This approximation lets us view the probability density function p as a collection of $|\mathcal{G}|$ variables $(p(x))_{x \in \mathcal{G}}$ and the objective can be rewritten as:

$$H(X) = -\sum_{x \in \mathcal{G}} p(x) \, \log p(x) \, \delta$$

Similarly, integrals in the constraints can also be written as sums. To handle the inequality constraint, we can apply the reparameterization $p(x) = \exp(s(x))$. To handle the equality constraints, we will make use of the method of Lagrange multipliers.

(a) *Write* the Lagrange function $\mathcal{L}((s(x))_{x \in \mathcal{G}}, \lambda_1, \lambda_2, \lambda_3)$ corresponding to the constrained optimization problem above, where $\lambda_1, \lambda_2, \lambda_3$ are used to incorporate the three equality constraints.

(b) *Show* that the probability distribution p that maximizes the objective $H(X)$ is a Gaussian probability density function with mean 0 and variance σ^2.

Exercise 2: Independent Components in Two Dimensions (25 P)

High entropy can be seen as the result of superposing many independent low-entropy sources. Independent component analysis (ICA) aims to recover the independent sources from the data by finding projections of the data that have low entropy, i.e. that diverge the most from a Gaussian probability distribution.

In the following, we consider a simple two-dimensional problem where we are given a joint probability distribution $p(x, y) = p(x) \, p(y|x)$ with

$$p(x) \sim \mathcal{N}(0, 1),$$
$$p(y|x) = \frac{1}{2} \, \delta(y - x) + \frac{1}{2} \, \delta(y + x),$$

where $\delta(\cdot)$ denotes the Dirac delta function. A useful property of linear component analysis for two-dimensional probability distributions is that the set of all possible directions to look for in $\mathbb{R}^2$ is directly given by

$$\left\{ \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix}, \quad 0 \leq \theta < 2\pi \right\}.$$

The projection of the random vector (x, y) on a particular component can therefore be expressed as a function of θ:

$$z(\theta) = x \, \cos(\theta) + y \, \sin(\theta).$$

As a result, ICA in the two-dimensional space is reduced to finding the values of the parameter $\theta \in [0, 2\pi[$ that maximize a certain objective $J(z(\theta))$.

(a) *Sketch* the joint probability distribution $p(x, y)$, along with the projections $z(\theta)$ of this distribution for angles $\theta = 0$, $\theta = \pi/8$ and $\theta = \pi/4$.

(b) *Find* the principal components of $p(x, y)$. That is, find the values of the parameter $\theta \in [0, 2\pi[$ that maximize the variance of the projected data $z(\theta)$.

(c) *Find* the independent components of $p(x, y)$, more specifically, find the values of the parameter $\theta \in [0, 2\pi[$ that maximize the non-Gaussianity of $z(\theta)$. We use as a measure of non-Gaussianity the excess kurtosis defined as

$$\text{kurt}\,[z(\theta)] = \frac{\mathrm{E}\big[(z(\theta) - \mathrm{E}[z(\theta)])^4\big]}{(\mathrm{Var}[z(\theta)])^2} - 3.$$

Exercise 3: Deriving a Special Case of FastICA (25 P)

We consider our input data x to be in $\mathbb{R}^d$ and coming from some distribution $p(x)$. We assume we have applied as an initial step the centering and whitening procedures so that:

$$\mathrm{E}[x] = 0 \qquad \text{and} \qquad \mathrm{E}[xx^\top] = I.$$

To extract an independent component, we would like to find a unit vector w (i.e. $\|w\|^2 = 1$) such that the excess kurtosis of the projected data:

$$\text{kurt}\,[w^\top x] = \frac{\mathrm{E}\big[(w^\top x - \mathrm{E}[w^\top x])^4\big]}{(\mathrm{Var}[w^\top x])^2} - 3.$$

is maximized.

(a) *Show* that for any w subject to $\|w\|^2 = 1$, the projection $z = w^\top x$ has mean 0 and variance 1.

(b) Show using the method of Lagrange multipliers that the projection w that maximizes $\text{kurt}\,[w^\top x]$ is a solution of the equation

$$\lambda w = \mathrm{E}[x \cdot (w^\top x)^3]$$

where λ can be resolved by enforcing the constraint $\|w\|^2 = 1$.

(c) The solution of the previous equation cannot be found analytically, and must instead be solved iteratively using e.g. the Newton's method. The Newton's method assumes that the equation is given in the form $F(w) = 0$ and then defines the iteration as $w^+ = w - J(w)^{-1}F(w)$ where w^+ denotes the next iterate and where J is the Jacobian associated to the function F. *Show* that the Newton iteration in our case takes the form:

$$w^+ = w - (\mathrm{E}[3xx^\top(w^\top x)^2 - \lambda I])^{-1} \cdot (\mathrm{E}[x \cdot (w^\top x)^3] - \lambda w)$$

(d) *Show* that when making the decorrelation approximation $\mathrm{E}[xx^\top(w^\top x)^2] = \mathrm{E}[xx^\top] \cdot \mathrm{E}[(w^\top x)^2]$, the Newton iteration further reduces to

$$w^+ = \gamma \cdot (\mathrm{E}[x \cdot (w^\top x)^3] - 3w)$$

where γ is some constant factor. (The constant factor does not need to be resolved since we subsequently apply the normalization step $w^+ \leftarrow w^+/\|w^+\|$.)

Exercise 4: Programming (30 P)

Download the programming files on ISIS and follow the instructions.