Exercises for the course

Machine Learning 2

Summer semester 2017

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 (10+10+5+5 P)

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

$$H(x) = -\int p(x)\,\log p(x)\,dx$$

We would like to find the probability density function $p(x)$ that maximizes the differential entropy under the following constraints

$$\forall x : p(x) \geq 0 \quad , \quad \textstyle\int p(x)\,dx = 1 \quad , \quad \mathrm{E}[x] = 0 \quad , \quad \mathrm{Var}[x] = \sigma^2.$$

The first set of inequality constraints is handled by rewriting the unknown density function as $p(x) = \exp(s(x))$ and searching for a function $s(x)$ that maximizes the objective. Here, we view the function $s(x)$ as an infinite-dimensional vector, and therefore can write

$$\frac{\partial \int f(s(x))\,dx}{\partial s(x)} = f'(s(x)).$$

(a) *Write* the Lagrange function $\Lambda(s(x), \lambda_1, \lambda_2, \lambda_3)$ 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 function $s(x)$ that maximizes the objective $H(x)$ is quadratic in x.

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

(d) *Show* that for every univariate random variable x of mean 0 and variance σ^2, and assuming $x^* \sim \mathcal{N}(0, \sigma^2)$, the negentropy $J(x) = H(x^*) - H(x)$ that independent component analysis seeks to maximize is always greater or equal to 0, and is equal to 0 when x is Gaussian-distributed.

Exercise 2: Finding Independent Components (10+15+15 P)

We consider the 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, linear component analysis such as PCA or ICA in the two-dimensional space is reduced to finding the parameters $\theta \in [0, 2\pi[$ 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 parameters $\theta \in [0, 2\pi[$ that maximize the variance of the projected data $z(\theta)$.

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

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

Exercise 3: Programming (30 P)

Download the programming files on ISIS and follow the instructions.