2413, Machine Learning, Tutorial 10
Universität Bern

Simon Jenni (jenni@inf.unibe.ch)

PCA

PCA is a dimensionality reduction algorithm where each data point is projected to the first k eigenvectors of their covariance matrix. It is important that the data is first normalised, as follows:

- Let $\mu = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$
- Replace each $x^{(i)}$ with $x^{(i)} - \mu$
- Let $\sigma_j^2 = \frac{1}{m}\sum_i (x_j^{(i)})^2$
- Replace each $x_j^{(i)}$ with $x_j^{(i)}/\sigma_j$

If you have prior knowledge that the input data has zero mean, you can skip the first two steps. If you have prior knowledge that the dimension of the data coordinates have the same scale, you can skip the third and fourth step.

1. There was a beer drinking competition at the Octoberfest. There were two rounds, one on Saturday and one on Sunday. The participants needed to drink as many mugs of beer as possible in 30 minutes. The data below shows their results. The first row shows the amount they drank on Saturday, and the second show their result on Sunday.

	$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$
$mugs(Saturday)$	3	2	5	4
$mugs(Sunday)$	1	2	3	4

- Find the first principal component.

- Find the second principal component.

- Model the data as the sum of one principal component and some noise. Remove the noise from the data.

2. There is a nice theorem in numerical linear algebra, that says the following. Suppose M is an $m \times n$ matrix whose entries are real numbers. Then there exists a factorisation of the form

$$M = USV^T, \tag{1}$$

where U and V are orthogonal matrices, and S is diagonal, and $s_i \geq 0$ for every diagonal element of S. The s_i diagonal elements are called singular values, and the factorisation itself is called singular value decomposition (SVD). This decomposition is unique (in most cases). How would you compute the principal components using SVD?

Factor Analysis

3. Let us assume that the random variable x given z is defined by the following model

$$\epsilon \sim \mathcal{N}(0, \Psi)$$
$$x = \mu + \Lambda R z + \epsilon,$$

where R is a known rotation matrix, $R^\top R = I$, and $z \in \mathbb{R}^n$ and $\epsilon \in \mathbb{R}^m$, with $n < m$, are independent random variables, and $\mathcal{N}(0, \Psi)$ denotes the zero-mean Gaussian distribution with covariance Ψ. Let $\Lambda \in \mathbb{R}^{m \times n}$, $\mu \in \mathbb{R}^m$ and $\Psi \in \mathbb{R}^{m \times m}$ be (unknown) model parameters. Let us assume that the expectations $\mathrm{E}_{z \sim Q}[z]$ and $\mathrm{E}_{z \sim Q}[z z^\top]$ are given, where Q is some probability density function of the vector z.

Let us assume that we are given m independent and identically distributed samples $x^{(1)}, \ldots, x^{(m)}$. Each sample $x^{(i)}$ will have an associated vector $z^{(i)}$, with its own probability density function Q_i. Compute Λ by maximizing the likelihood

$$\sum_{i=1}^{m} \mathrm{E}_{z^{(i)} \sim Q_i} \left[\log p(x^{(i)} | z^{(i)}; \mu, \Lambda, \Psi) \right] =$$
$$\sum_{i=1}^{m} \mathrm{E}_{z^{(i)} \sim Q_i} \left[-\frac{1}{2} \log |\Psi| - \frac{n}{2} \log(2\pi) - \frac{1}{2} \left(x^{(i)} - \mu - \Lambda R z^{(i)} \right)^\top \Psi^{-1} \left(x^{(i)} - \mu - \Lambda R z^{(i)} \right) \right]$$

Hint: Make use of the following *trace* formulas: $\nabla_A \mathrm{tr}(A^\top B A C) = B A C + B^\top A C^\top$ and $\mathrm{tr}(AB) = \mathrm{tr}(BA)$.