**Ludwig-Maximilians-Universitaet Muenchen** 20.05.2014
**Institute for Informatics**
Prof. Dr. Volker Tresp
Gregor Jossé
Johannes Niedermayer

# Machine Learning and Data Mining

Summer 2015

## Exercise Sheet 5

*Presentation of Solutions to the Exercise Sheet on the 27.05.2014*

**Exercise 5-1** Probability Calculus

Let $X$ and $Y$ be random variables with the following data:

| | | Y | | |
|---|---|---|---|---|
| | | 1 | 2 | 3 |
| X | 1 | 0,1 | 0,15 | 0,25 |
| | 2 | 0,05 | 0,3 | 0,15 |

a) Compute the marginal distributions $P(X = x_i)$ and $P(Y = y_i)$

b) Compute the expected values $E(X)$, $E(Y)$

c) Compute the variances $var(X)$, $var(Y)$ as well as the covariance $cov(X,Y)$.

d) Compute the correlation $\rho = \frac{cov(X,Y)}{\sqrt{var(X) \cdot var(Y)}}$

e) Compute if the variables $X,Y$ are independent.

**Exercise 5-2** Conditional Probability I

Assume that a certain country's population is equally male and female (and that there exist no other sexes). Furthermore, assume that 10% of all men are color blind, but only 1% of all women.

(a) Compute the probability that a person is color blind.

(b) Compute the probability that a color blind person is male.

**Exercise 5-3** Conditional Probability II

If screening for a disease, there are several possible outcomes. Let $T+$, $T-$ denote the events that the test is positive and negative, respectively, and $D$, $\neg D$ denote the events of having and not having the disease, respectively. There are two major criteria to evaluate tests by:

- Sensitivity: Probability (in practice more likely: ratio) of positively tested people having the disease, i.e., $P(T+ \mid D)$.
- Specificity: Probability (or ratio) of negatively tested people not having the disease, i.e., $P(T- \mid \neg D)$.

Now, assume a (realistic) test for HIV with a sensitivity and specificity of 99.9%. Suppose that a person is randomly selected from a population where 1% are infected with HIV and tested with the aforementioned test. Compute the probability that the person has HIV if:

(a) The test is positive.

(b) The test is negative.

**Exercise 5-4** Interpretation of the Standard Deviation

Sketch the graph of the standardized normal distribution with the following parameters

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

where $\mu = 0$ and $\sigma = 1$, in the intervall $x \in [-4, 4]$.
Mark and intepret the intervalls $0 \pm \sigma;\ 0 \pm 2\sigma;\ 0 \pm 3\sigma$.

**Exercise 5-5** Kernel Theory I

Inner products play a central role in machine learning algorithms, some of them even rely solely on inner products. This is the reason for the importance of what is known as the "kernel trick". Note that the word "kernel" is used differently throughout mathematics and also in machine learning not always consistently. The idea of the kernel trick is to construct non-linear variants of a linear algorithms by substituting inner products with non-linear kernels. If such a non-linear kernel is symmetric and positive semi-definite, then it is equivalent to computing an inner product in some vector space. This corresponding vector space is often of high dimensionality or even infinitely dimensional. Hence, we obtain better separation properties. And, conveniently, we do not have to compute the representation of our data in this vector space (mostly referred to as "feature space"). Instead, we can restrict ourselves to computing the kernel values, which serves as a similarity measure in the implicitly given vector space. In this exercise we want to compute the explicit representation of some kernels.

(a) The homogeneous quadratic kernel $K(x,y) = \langle x,y\rangle^2$ defined on the 2-dimension real vector space.

(b) The gaussian radial basis function kernel $K(x,y) = \exp\left(-\frac{\|x-y\|^2}{2}\right)$ defined on an $n$-dimensional real vector space.

**Exercise 5-6** Kernel Theory II

In general, the validity of a kernel can be shown with Mercer's Theorem. However, this is in practice often complicated.We therefore present alternative methods for constructing kernels and for proving that a kernel is valid.

(a) Prove that for any matrix $A^{n\times m}$ holds: $K(x,y) := x^T A^T A y$ is a valid kernel.

(b) *Remark:* It is also possible to construct kernel by combining valid kernels via some elementary operations. For valid kernels $k_l(\mathbf{x}_i,\mathbf{x}_j)$ holds:

(i) **Scaling**: For $a > 0$: $k(\mathbf{x}_i,\mathbf{x}_j) := a \cdot k_1(\mathbf{x}_i,\mathbf{x}_j)$ is a kernel.

(ii) **Sum**: $k(\mathbf{x}_i,\mathbf{x}_j) := k_1(\mathbf{x}_i,\mathbf{x}_j) + k_2(\mathbf{x}_i,\mathbf{x}_j)$ is a kernel.

(iii) **Linear combination**: For $w \in \mathbb{R}^d_+$: $k(\mathbf{x}_i,\mathbf{x}_j) := \sum_{l=1}^{d} w_l \cdot k_l(\mathbf{x}_i,\mathbf{x}_j)$is a kernel.

(iv) **Product**: $k(\mathbf{x}_i,\mathbf{x}_j) := k_1(\mathbf{x}_i,\mathbf{x}_j) \cdot k_2(\mathbf{x}_i,\mathbf{x}_j)$ is a kernel.

(v) **Power**: For a $p \in \mathbb{N}_+ : k(\mathbf{x}_i,\mathbf{x}_j) := (k_1(\mathbf{x}_i,\mathbf{x}_j))^p$ is a kernel.

It is an optional exercise to prove the above implications.

(c) Prove that $K(x,y) = k_1(x,y) - k_2(x,y)$ is not a valid kernel.

(d) If I cannot prove or disprove the validity of a kernel, is there a way to experimentally check the validity?