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.

Possible Solution:

a)

		Y			
		1	2	3	
X	1	0,1	0,15	0,25	0,5
	2	0,05	0,3	0,15	0,5
		0,15	0,45	0,4	1

b)

$$E(X) = E_{P(x)}(X) = \sum_i x_i P(X = x_i)$$
$$E(X) = 1 \cdot 0{,}5 + 2 \cdot 0{,}5 = 1{,}5$$
$$E(Y) = 1 \cdot 0{,}15 + 2 \cdot 0{,}45 + 3 \cdot 0{,}4 = 2{,}25$$

c) $var(X) = \sum_i [x_i - E(X)]^2 \cdot P(X = x_i) = E((X - E(X))^2) = E(X^2) - (E(X))^2$

$$var(X) = (1 - 1{,}5)^2 \cdot 0{,}5 + (2 - 1{,}5)^2 \cdot 0{,}5 = 0{,}25$$
$$var(Y) = (1 - 2{,}25)^2 \cdot 0{,}15 + (2 - 2{,}25)^2 \cdot 0{,}45 + (3 - 2{,}25)^2 \cdot 0{,}4 = 39/80 \approx 0{,}4875$$
$$cov(X,Y) = \sum_i \sum_j (x_i - E(X))(y_i - E(Y))P(X = x_i, Y = y_i)$$
$$= 0.1(1 - 1.5)(1 - 2.25) + 0.15(1 - 1.5)(2 - 2.25) + 0.25(1 - 1.5)(3 - 2.25) +$$
$$0.05(2 - 1.5)(1 - 2.25) + 0.3(2 - 1.5)(2 - 2.25) + 0.15(2 - 1.5)(3 - 2.25)$$
$$= -0{,}025$$
$$oder\ mit\ Verschiebungssatz : cov(X,Y) = E(XY) - E(X)E(Y)$$
$$E(XY) = \underbrace{1 * 0.1 + 2 * 0.15 + 3 * 0.25}_{Zeile 1} + \underbrace{2 * 0.05 + 4 * 0.3 + 6 * 0.15}_{Zeile 2} = 3{,}35$$
$$cov(XY) = 3{,}35 - 1{,}5 * 2{,}25 = -0{,}025$$

d) $\rho = \frac{cov(X,Y)}{\sqrt{var(X) \cdot var(Y)}} = \frac{-0{,}025}{\sqrt{0{,}25 \cdot 0{,}4875}} \approx -0{,}07$

e) No, because:

- $\rho \neq 0$ is not a sufficient property! Compare: X,Y independent $\Rightarrow$ X,Y uncorrelated (but the other implication does not hold)
- $P(X = 2, Y = 1) = 0{,}05 \neq 0{,}15 * 0{,}5 = 0{,}075$, so no independence.

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.

Possible Solution:

Use the law of total probability:
Let $(B_n)_n$ be a finite or countably infinite partition of the sample space Ω, i.e., for all $i \neq j$ holds $B_i \cap B_j = \emptyset$ and $\bigcup B_n = \Omega$. Then, for any A (in the same probability space) holds: $P(A) = \sum P(A \mid B_n)$.
Let C denote the event of color blindness, and let F and M denote the event of being female and male, respectively.

$$P(C) = P(C \mid F) + P(C \mid M) = P(C \cap F) \cdot P(F) + P(C \cap M) \cdot P(M) = 0.01 \cdot 0.5 + 0.1 \cdot 0.5 = 0.055$$

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

Possible Solution:

$$P(M \mid C) = \frac{P(M \cap C)}{P(C)} = \frac{P(C \cap M)}{P(C)} = \frac{P(C \mid M) \cdot P(M)}{P(C)} = \frac{0.1 \cdot 0.5}{0.055} = 0.9091$$

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.

Possible Solution:

$$P(D \mid T+) = \frac{P(T+ \mid D) \cdot P(D)}{P(T+)}$$

where

$$P(T+) = P(T+ \mid D) \cdot P(D) + P(T+ \mid \neg D) \cdot P(\neg D) = P(T+ \mid D) \cdot P(D) + (1 - P(T- \mid \neg D)) \cdot P(\neg D)$$

It follows:

$$P(D \mid T+) = 0.9098$$

(b) The test is negative.

Possible Solution:

$$P(D \mid T-) = \frac{P(T- \mid D) \cdot P(D)}{P(T-)}$$

where

$$P(T-) = P(T- \mid D) \cdot P(D) + P(T- \mid \neg D) \cdot P(\neg D) = P(T- \mid D) \cdot P(D) + (1 - P(T+ \mid \neg D)) \cdot P(\neg D)$$

It follows:

$$P(D \mid T-) = 1 \cdot 10^{-5}$$

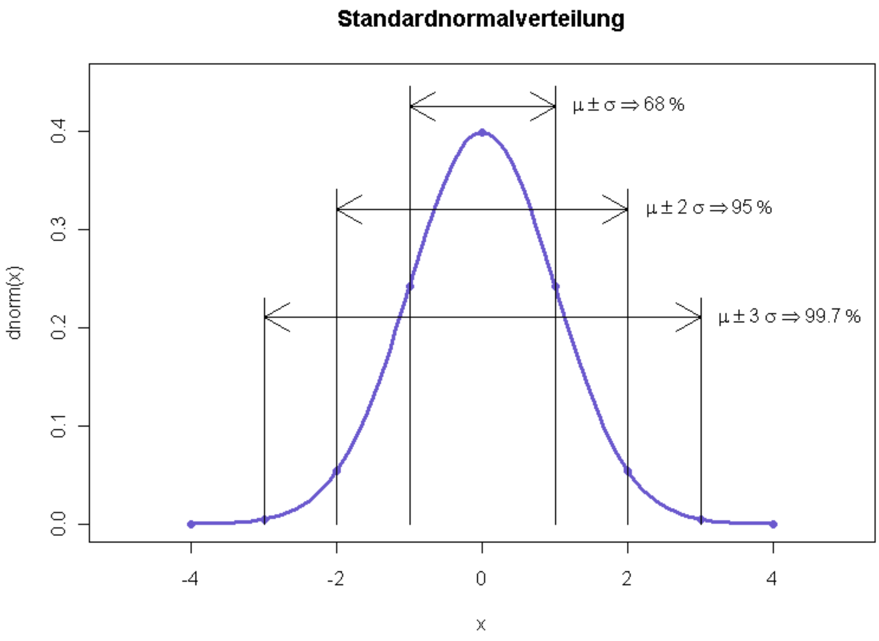
Exercise 5-4 Interpretation of the Standard Deviation

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 $0 \pm \sigma;\ 0 \pm 2\sigma;\ 0 \pm 3\sigma$.

Possible Solution:



Verteilungsfunktionen in `R` unter `dnorm`:

Dichteverteilung: $P(X = x)$ (`dnorm`)

Wahrscheinlichkeitsverteilung: $P(X <= x)$ (`pnorm`)

Quantilverteilung: $\arg_x P(X <= x) = q$ (`qnorm`)

Zufallige Stichprobe aus der Verteilung: Datensatz der entsprechend der Verteilung zufallsverteilt ist. (`rnorm`)

Folgen die Daten einer Standardnormalverteilung, gilt die 68-95-99.7-Prozent-Regel:

- 68% der Beobachtungen liegen im Intervall $\mu \pm \sigma$
(also $P(X \geq \mu - \sigma \wedge X \leq \mu + \sigma) = P(X \leq \mu + \sigma) - P(X \geq \mu - \sigma))$
- 95% der Beobachtungen liegen im Intervall $\mu \pm 2\sigma$
- 99.7% der Beobachtungen liegen im Intervall $\mu \pm 3\sigma$

Wenn sich eine Häufigkeitsverteilung fur die Daten eines Merkmals X gut durch eine Normalverteilung mit $\mu = 0, \sigma = s$ approximieren lasst, gelten die Aussagen wie die 68-95-99.7-Prozent-Regel auch approximativ fur diese Daten.
Quelle: Statistik (Ludwig Fahrmeir, Rita Knstler, Iris Pigeot) S. 92f

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.

Possible Solution:

$$\begin{aligned} K(x,y) =& \langle x,y \rangle^2 \\ =& (x_1 y_1 + x_2 y_2)^2 \\ =& x_1^2 y_1^2 + 2 x_1 x_2 y_1 y_2 + x_2^2 y_2^2 \\ =& \langle (x_1^2, \sqrt{2} x_1 x_2, x_2^2), (y_1^2, \sqrt{2} y_1 y_2, y_2^2) \rangle \\ =& \langle \phi(x), \phi(y) \rangle \end{aligned}$$

Hence, the homogeneous quadratic kernel corresponds to the inner product of a three-dimensional vector space. Or, in other words: the corresponding feature map has three coordinates.
Remark: For general homogeneous polynomial kernels $K(x,y) = \langle x,y \rangle^d$ where x,y are k-dimensional vectors, the corresponding feature map has $\binom{k+d-1}{d}$ coordinates.
Remark: Non-homogeneous polynomial kernels are defined as $K(x,y) = (\langle x,y \rangle + c)^d$ for some constant $c > 0$.

(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.

Possible Solution:

$$
\begin{aligned}
&\exp\Big(-\frac{\|x-y\|^2}{2}\Big)\\
&\underset{(\star)}{=}\exp\Big(-\frac{\|x\|^2}{2}\Big)\exp\Big(-\frac{\|y\|^2}{2}\Big)\exp(x^Ty)\\
&=\sum_{j=0}^{\infty}\frac{(x^Ty)^j}{j!}\exp\Big(-\frac{\|x\|^2}{2}\Big)\exp\Big(-\frac{\|y\|^2}{2}\Big)\\
&=\sum_{j=0}^{\infty}\Big(x^Ty\frac{\exp\big(-\frac{\|x\|^2}{2j}\big)}{\sqrt{j!}^{1/j}}\frac{\exp\big(-\frac{\|y\|^2}{2j}\big)}{\sqrt{j!}^{1/j}}\Big)^j\\
&\underset{(\star\star)}{=}\sum_{j=0}^{\infty}\sum_{\sum_i n_i=j}\frac{\exp\big(-\frac{\|x\|^2}{2j}\big)}{\sqrt{j!}^{1/j}}\frac{\exp\big(-\frac{\|y\|^2}{2j}\big)}{\sqrt{j!}^{1/j}}\prod_{i=0}^{k}\Big(\binom{j}{n_1,\dots,n_k}x_i^{n_i}y_i^{n_i}\Big)\\
&=\sum_{j=0}^{\infty}\sum_{\sum_i n_i=j}\left(\frac{\exp\big(-\frac{\|x\|^2}{2j}\big)}{\sqrt{j!}^{1/j}}\prod_{i=0}^{k}\Big(\binom{j}{n_1,\dots,n_k}^{\frac{1}{2}}x_i^{n_i}\Big)\right)\left(\frac{\exp\big(-\frac{\|y\|^2}{2j}\big)}{\sqrt{j!}^{1/j}}\prod_{i=0}^{k}\Big(\binom{j}{n_1,\dots,n_k}^{\frac{1}{2}}y_i^{n_i}\Big)\right)
\end{aligned}
$$

Where

$$(\star):\ \langle x-y,x-y\rangle^2=\langle x,x\rangle+\langle y,y\rangle-2\langle x,y\rangle$$

and

$$(\star\star):\ \binom{j}{n_1,\dots,n_k}=\frac{j!}{n_1!\cdot\dots\cdot n_k!}$$

is the multinomial coefficient (the number of possibilities for putting j indistinct balls in k bins where the i-th bin holds exactly n_i elements).
Concludingly,

$$\phi(x)=\left(\frac{\exp\big(-\frac{\|x\|^2}{2j}\big)}{\sqrt{j!}^{1/j}}\prod_{i=0}^{k}\Big(\binom{j}{n_1,\dots,n_k}^{\frac{1}{2}}x_i^{n_i}\Big)\right)_{j=0,\dots,\infty,\sum_i n_i=j}$$

which shows that the corresponding feature map to the gaussian RBF kernel has infinite dimensionality.

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: $K(x,y):=x^TA^TAy$ is a valid kernel.

Possible Solution:

It suffices to show that $K(x,y)$ is an inner product on some Hilbert space. Then by the aforementioned equivalence, $K(x,y)$ is symmetric and positive semi-definite, i.e., a kernel. We define $\phi(x):=Ax$. Then

$$\langle\phi(x),\phi(y)\rangle=x^TA^TAy=K(x,y)$$

(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 $w \in \mathbb{R}^d_+$: $k(\mathbf{x}_i,\mathbf{x}_j) := \sum\limits_{l=1}^{d} w_l \cdot k_l(\mathbf{x}_i,\mathbf{x}_j)$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.

Possible Solution:

Let, for example, $k_1(x,y) := x^T y$ be the identity, and let $k_2(x,y) := (x^T y)^2$ be the homogeneous quadratic kernel. Furthermore, let x such that $x^T x > 1$. Then

$$K(x,x) = x^T x - (x^T x)^2 < 0$$

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

Possible Solution:

Given a dataset X (where the rows are the patterns), the kernel matrix is defined as $K = XX^T$. Thereby K is symmetric, hence, K is a valid kernel iff it is positive semi-definite. This can empirically be investigated by generating random data points (according to the underlying [assumed] distribution) and checking whether the according kernel matrix is indeed positive semi-definite. In Matlab for example:

$$\begin{aligned} X &= a + (b - a). * rand(1000,10) \\ K &= X * X' \\ [U,V] &= eig(K) \\ eigvals &= diag(V) \\ [\ ,p] &= chol(K)//p > 0, \text{then not positive semi-definite} \end{aligned}$$