Machine Learning	Prof. Dr. Bernt Schiele
	Dr. Paul Swoboda & Dr. Gerard Pons-Moll
Solution 1 - 22.10.2018	due: 29.10.2018, 14:15

Exercise 1 - Correlation, Independence and all the rest

a. **(1 Point)** Let X be a $\mathbb{R}$-valued random variable with symmetric density function ($p(x) = p(-x)$, $\forall x \in \mathbb{R}$) and define $Y = X^2$. Compute the correlation of X and Y. What does this result tell you about the dependence of X and Y ? Are they independent ?

b. **(2 Points)** Suppose you flip a biased coin which shows head with probability $0 \leq p \leq 1$. You predict head with probability $0 \leq q \leq 1$. What is the probability that your prediction is correct under the assumption that the coin flip and your prediction are independent ? How has q to be chosen in order to maximize this probability (for fixed p) ?

Solution:

a. The covariance of X and Y is given as

$$\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \mathbb{E}[X^3] - \mathbb{E}[X]\mathbb{E}[X^2].$$

As X has a symmetric density function and the integrand x^3 resp. x is antisymmetric, we get

$$\mathbb{E}[X^3] = \mathbb{E}[X] = 0.$$

Thus the covariance is zero, which implies that the correlation is zero. The last conclusion is only true if the variance of X is non-zero, otherwise the correlation is undefined. However, as we assume that X has a density, it must have a non-zero variance.

Obviously, X and Y are dependent (however, note that the functional relationship is not injective, and thus we can not deduce from the knowledge of Y the value of X). The fact that the correlation is zero only shows that there is no **linear** dependence between X and Y. In particular, X and Y are not independent.

b. As the coin flip C and the prediction P are independent, we have

$$\mathrm{P}(C = A,\, P = B) = \mathrm{P}(C = A)\,\mathrm{P}(P = B),$$

for $A, B \in \{H, T\}$. The probability that the prediction is correct is

$$\mathrm{P}(P \text{ correct}) = \mathrm{P}(C = H)\mathrm{P}(P = H) + \mathrm{P}(C = T)\mathrm{P}(P = T) = pq + (1-p)(1-q) = 1 - p - q + 2pq.$$

This probability is an affine function of q, thus the maximum is attained either at $q = 0$ or $q = 1$. The optimal value of q is

$$q = \begin{cases} 0 & \text{if } p < \frac{1}{2}, \\ 1 & \text{if } p > \frac{1}{2} \end{cases}$$

In the case $p = \frac{1}{2}$, all q-values in $[0, 1]$ are optimal.

Exercise 2 - Transformation law

a. **(3 Points)** Let U_1, U_2 be two independent real-valued random variables with uniform distribution on [0, 1]. Compute the density of the variables $X = g(U)$:

$$X_1 = \sqrt{-2\ln U_2}\sin(2\pi U_1),$$
$$X_2 = \sqrt{-2\ln U_2}\cos(2\pi U_1).$$

The result is $p(x_1, x_2) = \frac{1}{2\pi}e^{-\frac{x_1^2+x_2^2}{2}}$ (a centered Gaussian in $\mathbb{R}^2$ with covariance matrix $C = \mathbb{1}$, where $\mathbb{1}$ is the identity matrix in $\mathbb{R}^2$). This is the so-called **Box-Muller**-transform which is used to generate samples of a centered Gaussian with unit variance (this is what you get if you use the command `randn` in Matlab).

Hints:

a. Derive the inverse function g^{-1}, that is derive U_1, U_2 as a function of X_1, X_2 (you have to invert the given nonlinear system of equations). Here, $\sin^2\alpha + \cos^2\alpha = 1$ is quite helpful.

b. Then apply the transformation law from the lecture. First, derive the Jacobian of g^{-1} and then compute its determinant.

Solution:

a. The range of g is $\mathbb{R}^2$ if $c = d = 1$. In this case g is a continuously differentiable, injective mapping from $[0, 1) \times (0, 1) \to \mathbb{R}^2\backslash\{0\}$ with non-vanishing Jacobian. Thus we can apply the transformation formula given in the lecture. Note that the set $[0, 1] \times [0, 1]\backslash[0, 1) \times (0, 1)$ is a set of measure zero.

In order to derive the density of X we have to find the inverse mapping g^{-1} and its Jacobian. First we derive,

$$X_1^2 + X_2^2 = -2\ln U_2, \qquad \text{and} \qquad \tan(2\pi U_1) = \frac{X_1}{X_2}.$$

From these two relations we deduce the inverse mapping g^{-1} defined on $\mathbb{R}^2\backslash\{0\}$,

$$u_1 = \frac{1}{2\pi}\arctan\left(\frac{x_1}{x_2}\right)$$
$$u_2 = e^{-\frac{1}{2}(x_1^2+x_2^2)}.$$

The transformation for u_1 is often written like this but it does not cover all cases. For a one-to-one transformation we have to do a case distinction. We get

$$u_1 = \frac{1}{2\pi}\begin{cases} \arctan(x_1/x_2), & \text{if } x_2 > 0 \text{ and } x_1 > 0 \\ \arctan(x_1/x_2) + \pi, & \text{if } x_2 < 0, \\ \arctan(x_1/x_2) + 2\pi, & \text{if } x_2 > 0 \text{ and } x_1 < 0. \end{cases}$$

Note, that this does not change the derivative since we only add constants.

The Jacobian $J_{g^{-1}}$ of g^{-1} evaluated at (x_1, x_2) is given by

$$(J_{g^{-1}})_{ij} = \frac{\partial g_i^{-1}}{\partial x_j} = \frac{\partial u_i}{\partial x_j} = \begin{pmatrix} \frac{1}{2\pi}\frac{x_2}{x_1^2+x_2^2} & -\frac{1}{2\pi}\frac{x_1}{x_1^2+x_2^2} \\ -x_1 e^{-\frac{x_1^2+x_2^2}{2}} & -x_2 e^{-\frac{x_1^2+x_2^2}{2}} \end{pmatrix}$$

The computation of the determinant of $J_{g^{-1}}$ is then straightforward,

$$|\det J_{g^{-1}}(x_1, x_2)| = \frac{1}{2\pi}e^{-\frac{x_1^2+x_2^2}{2}}.$$

The joint original density $p(u_1, u_2)$ is uniform and $p(u_1, u_2) = 1$ for $u \in [0,1) \times (0,1)$. Therefore the joint density $p(x_1, x_2)$ of (X_1, X_2) is given for $(x_1, x_2) \neq (0,0)$,

$$p(x_1, x_2) = p(u_1(x_1, x_2), u_2(x_1, x_2))|\det J_{g^{-1}}(x_1, x_2)| = \frac{1}{2\pi} e^{-\frac{x_1^2 + x_2^2}{2}}.$$

Thus we have generated the density which equals the density of a Gaussian except for $(x_1, x_2) = (0,0)$ (where it is undefined). However, in the lecture we have discussed that a density is unique only up to sets of measure zero. The computed density is not equal to the Gaussian density for one point, but since a point has measure zero, the induced probability measure on $\mathbb{R}^2$ is the same.

Exercise 3 - Phenomena in high dimensions and sampling

This exercise shows that the geometric intuition we have from living in three dimensions can usually not be transferred to higher dimensional spaces. However, often one encounters learning problems where one has a lot of features und thus one is working in a high-dimensional space. Therefore it is important to get some idea what can go wrong in high dimensions.

The volume $\mathrm{vol}(B_d(r))$ of the d-dimensional ball $B_d(r)$ ($B_d(r) = \{x \in \mathbb{R}^d \mid \|x\|_2 \leq r\}$) of radius r in $\mathbb{R}^d$ is given as

$$\mathrm{vol}(B_d(r)) = \frac{\pi^{\frac{d}{2}} r^d}{\Gamma(\frac{d}{2} + 1)},$$

where Γ is the Gamma function and one has

$$\Gamma\left(\frac{d}{2} + 1\right) = \begin{cases} \left(\frac{d}{2}\right)! & \text{if } d \text{ even,} \\ \sqrt{\pi} \frac{d!!}{2^{\frac{d+1}{2}}} & \text{if } d \text{ is odd} \end{cases}.$$

Note, that $d!!$ is the **double factorial** defined as $d!! = d(d-2)\ldots 1$ where $1!! = 0!! = 1$.

a. **(1 Point)** Derive the limit of the volume of $B_d(1)$ as $d \to \infty$.

b. **(3 Points)** Suppose you want to sample from the uniform measure on $B_d(1)$ (not just the surface - the whole volume).

- What is the density of the uniform probability measure on $B_d(1)$?
- Somebody suggests *rejection sampling* to get samples from the uniform measure on $B_d(1)$. First one draws a point X uniformly from the d-dimensional cube $[-1,1]^d$ ($\mathtt{X = 2 * rand(d, 1) - 1}$ in Matlab). Then one selects X as a sample if $\|X\|_2 \leq 1$ otherwise one discards it. What is the probability that a sample is discarded ? How many samples do you have to draw on average in the first step in order to get 1000 samples from the d-dimensional unit ball $B_d(1)$ for $d = 3$ and $d = 20$?

c. **(2 Points)** Clearly, the method based on rejection sampling does not work in high dimensions. Describe a more efficient way of sampling from the d-dimensional unit ball using Gaussian samples where no sample is discarded ? Prove that the method you suggest generates samples from the uniform distribution on $B_d(1)$.

Hints:

a. The probability in b) should be calculated and not be determined by numerical experiments (for $d = 20$ it will anyway not work :)).

b. There might be several ways in part c). The hint points in one particular direction.

- In spherical coordinates $(r, \theta_1, \ldots, \theta_{d-1})$ on $\mathbb{R}^d$ one has for every $S \subset \mathbb{R}^d$,

$$\int_S f(x_1, \ldots, x_d)\, dx_1 \ \ldots \ dx_d = \int_S f(r, \theta_1, \ldots, \theta_d)\, r^{d-1}\, dr\, d\Omega,$$

where $d\Omega$ is the surface element on the sphere (up to a factor this is the uniform measure on the surface of $B_d(1)$). You may use that the surface S_{d-1} of the unit sphere in $\mathbb{R}^d$ has $(d-1)$-dimensional volume $\mathrm{vol}(S_{d-1}) = d\,\mathrm{vol}(B_d(1))$.

- This suggests a two-step procedure. First one has to sample a point uniformly on the surface of $B_d(1)$. This will yield the directional part and second one has to modify the norm of this point in a way that it behaves as the radial part of the uniform measure on $B_d(1)$.

Solution:

a. We use the fact that the factorial grows larger than the exponential function. This can be seen from the following upper and lower bound derived by Robbins

$$\sqrt{(2\pi)} n^{n+\frac{1}{2}} e^{-n+\frac{1}{12n+1}} < n! < \sqrt{(2\pi)} n^{n+\frac{1}{2}} e^{-n+\frac{1}{12n}}$$

Moreover, one can check that $d!! > \left(\frac{d}{2}\right)!$ (every term in the product is larger than the corresponding one in the other product). Lower bounding also the term for d even using $\left(\frac{d}{2}\right)! > \frac{\left(\frac{d}{2}\right)!}{2^{\frac{d+1}{2}}}$, we get in total

$$\mathrm{vol}(B_d(1)) = \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)} < \frac{(2\pi)^{\frac{d}{2}}\sqrt{2}}{\left(\frac{d}{2}\right)!}.$$

And since the upper bound converges to zero since the factorial grows faster than any exponential function, we have

$$\lim_{d\to\infty} \mathrm{vol}(B_d(1)) = 0.$$

b.
- The density of any uniform distribution on a set A is given as

$$p(x) = \frac{1}{\mathrm{vol}(A)}, \quad \forall\, x \in A.$$

and thus the density of the uniform measure on $B_d(1)$ is

$$p(x) = \frac{1}{\mathrm{vol}(B_d(1))} = \frac{\Gamma(\frac{d}{2}+1)}{\pi^{\frac{d}{2}}}, \quad \forall\, x \in B_d(1).$$

- We discard the samples lying in $[-1,1]^d \backslash B_d(1)$. Since we sample uniformly from $[-1,1]^d$, the probability that a sample X is discarded is

$$\mathrm{P}(X \text{ is discarded}) = \mathrm{P}(X \in [-1,1]^d \backslash B_d(1)) = \frac{\mathrm{vol}\left([-1,1]^d \backslash B_d(1)\right)}{\mathrm{vol}\left([-1,1]^d\right)} = 1 - \frac{\pi^{\frac{d}{2}}}{2^d\,\Gamma(\frac{d}{2}+1)}.$$

This is just a Bernoulli-random variable and thus repeated sampling yields the Binomial distribution. The expected value of a binomial random variable Z with parameters (n,p) is given by $\mathbb{E}[Z] = n\,p$. The number of accepted samples is $n\,(1 - \mathrm{P}(X \text{ is discarded}))$ and thus we can resolve for our requested 1000 samples. Thus the number of samples we have to draw on average is given by

$$n = \frac{1000}{1 - \mathrm{P}(X \text{ is discarded})} = \frac{1000}{\frac{\pi^{\frac{d}{2}}}{2^d\,\Gamma(\frac{d}{2}+1)}} = \frac{1000\,2^d\,\Gamma(\frac{d}{2}+1)}{\pi^{\frac{d}{2}}}.$$

For $d = 3$, $\mathrm{P}(X \text{ is discarded})$ is equal to ≈ 0.4764 and thus we have to sample on average 1910 samples in order to get 1000 samples from the uniform measure on $B_d(1)$ for $d = 3$. For $d = 20$ the probability of discarding a sample is equal to $\approx 1 - 2.46 * 10^{-8}$ and thus we have to sample on average $4.06 * 10^{10}$ samples in order to get 1000 samples from

the uniform measure on $B_d(1)$. (for $d = 40$, roughly $3.1 * 10^{23}$ samples - Sampling one million samples from the uniform measure on $[-1, 1]^{40}$ takes 0.8 seconds - so the total time to get 1000 samples from $B_d(1)$ for $d = 40$ would be 7.5 billion years !!! - for $d = 20$ only eight hours :)).

c. Using the hint with $S = B_d(1)$ and $f(x) = \frac{1}{\text{vol}(B_d(1))}$ we see that in spherical coordinates $(r, \theta_1, \ldots, \theta_{d-1})$ the density is given as

$$p(r, \theta_1, \ldots, \theta_{d-1}) = \frac{1}{\text{vol}(B_d(1))} r^{d-1} g(\theta_1, \ldots, \theta_{d-1}),$$

where g is the function such that $d\Omega = g(\theta_1, \ldots, \theta_{d-1}) d\theta_1 \ldots d\theta_{d-1}$. In particular, we observe that in spherical coordinates the r-component is independent of the angles $\theta_1 \ldots \theta_{d-1}$ since the density factorizes into a product of an angular and radial part. We use this independence to decompose the generation of uniform samples from $B_d(1)$ into two independent procedures one for the angular part and one for the radial part. As stated in the hint $d\Omega$ is just the uniform measure on the surface of $B_d(1)$. We can construct a procedure for sampling from the surface of the unit sphere by taking any spherically symmetric probability density, that is

$$p(x_1, \ldots, x_d) = p(\|x\|).$$

Since the density is the same for any direction this yields the uniform measure on the sphere. The radial part is eliminated by normalizing the vector to have norm 1. One possibility is to use a Gaussian density $p(x) = \frac{1}{(2\pi)^{d/2}} e^{-\frac{\|x\|^2}{2}}$. Now it remains to get the radial part correct. Denoting by Ω the surface of $B_d(1)$ we note that the marginal density in r of the uniform measure on $B_d(1)$ is given as

$$\begin{aligned} p(r) &= \int_\Omega p(r, \theta_1, \ldots, \theta_{d-1}) d\Omega = \frac{1}{\text{vol}(B_d(1))} r^{d-1} \int_\Omega g(\theta_1, \ldots, \theta_{d-1}) d\theta_1 \ \ldots \ d\theta_{d-1} \\ &= \frac{1}{\text{vol}(B_d(1))} r^{d-1} \int_\Omega d\Omega = \frac{d \ \text{vol}(B_d(1))}{\text{vol}(B_d(1))} r^{d-1} = d\, r^{d-1}, \end{aligned}$$

where we have used in the last step that the surface S_{d-1} of the d-dimensional unit ball has $(d-1)$-dimensional volume $S_{d-1} = d \ \text{vol}(B_d(1))$. We can also make the sanity check in order to see if the marginal integrates to 1. Note, that the radius of points in the unit-ball varies between 0 and 1,

$$\int_0^1 p(r)\, dr = d \int_0^1 r^{d-1}\, dr = \left[r^d\right]_0^1 = 1.$$

The remaining task is to generate a density on $[0, 1]$ with density $p(r) = d\, r^{d-1}$. Suppose we know X has uniform distribution on $[0, 1]$ (Matlab: `rand`). Then

$$z = x^{\frac{1}{d}} \implies x = z^d \implies \frac{\partial x}{\partial z} = d\, z^{d-1}.$$

Thus $p(z) = \left|\frac{\partial x}{\partial z}\right| p(x) = d\, z^{d-1}$ is distributed as we want. We can summarize the sampling as follows

1. Generate sample Y from a centered Gaussian distribution with covariance $C = \sigma^2 \mathbb{1}$ (for some σ - the actual value does not matter as the distribution is spherically symmetric for any σ) and normalize it $Y' = \frac{1}{\|Y\|} Y$.
2. Draw a sample X uniformly from $[0, 1]$ and apply the transformation $R = X^{\frac{1}{d}}$
3. The final sample Z from the uniform distribution on $B_d(1)$ is then $Z = RY'$.

For 1000 samples in 40 dimensions it takes 0.3 seconds. Compared to 7.5 billion years using rejection sampling this it quite some speed up.