Machine Learning	Dr. Mario Fritz and Dr. Björn Andres
Exercise Sheet 2 - 3.11.2014	due: 14.11.2014

Exercise 4 - Basic Concepts (Probability)

This exercise reviews some of the concepts of probability theory needed for the lecture. Reading the first three chapters of the book *All of Statistics* by Larry Wasserman will be helpful if you are not well-versed in these concepts.

a. **(2 points) Probability versus Density:** Let X be a random variable on $\mathbb{R}$ with the following density

$$p(x) = \begin{cases} 2x & 0 \leq x \leq 1/2 \\ -2x+2 & 1/2 \leq x \leq 1 \\ 2x-4 & 2 \leq x \leq 5/2 \\ -2x+6 & 5/2 \leq x \leq 3 \\ 0 & \text{otherwise} \end{cases}$$

1. Sketch the probability density.
2. Find the probabilities: $P(X = \frac{1}{2})$, $P(X \in \{\frac{1}{2}, \frac{5}{2}\})$ and $P(X \in [\frac{1}{2}, \frac{5}{2}])$.
3. Determine the $\frac{1}{4}$-quantile of X.

b. **(2+2 Points) Transformations of Random Variables**:

1. Let (X, Y) be uniformly distributed on the unit disk: $\{(x, y) : x^2 + y^2 \leq 1\}$. Let $R = \sqrt{X^2 + Y^2}$. Determine the cumulative distribution function and the probability density function of R.
2. **(Bonus)** Let $X \sim \mathcal{N}(\mu_1, \sigma^2)$ and $Y \sim \mathcal{N}(\mu_2, \sigma^2)$ be independent Gaussian random variables. Determine the density of $X + Y$.

c. **(2+2 Points) Independence and Expectations**:

1. Let X be a random variable with uniform distribution on $[0, 1]$. Let Y and Z be functions of X defined as

$$Y = \begin{cases} 1 & 0 < X < b \\ 0 & \text{otherwise} \end{cases}$$

and

$$Z = \begin{cases} 1 & a < X < 1 \\ 0 & \text{otherwise} \end{cases}$$

where $0 < a < b < 1$. Are Y and Z independent? Why/Why not? Determine $\mathbb{E}(Y|Z)$.

2. Let $X_1, \dots X_m$ and $Y_1, \dots Y_n$ be random variables and let $a_1, \dots a_m$ and $b_1, \dots b_n$ be constants. Show that

$$\operatorname{Cov}\left(\sum_{i=1}^{m} a_i X_i, \sum_{j=1}^{n} b_j Y_j\right) = \sum_{i=1}^{m} \sum_{j=1}^{n} a_i b_i \operatorname{Cov}(X_i, Y_j)$$

where $\operatorname{Cov}(X, Y)$ is the covariance of X and Y.

d. **(2 Points) Variance and Correlation**: Correlation is a measure of *linear* dependence. Let X and $Y = bX + a$ be random variables, where $b \neq 0$, $a \in \mathbb{R}$, and X takes values in $\mathbb{R}$.

1. Show that X and Y have correlation:

$$\mathrm{Corr}(X, Y) = \frac{b}{|b|}.$$

2. As the absolute value of the correlation is upper bounded by one, this implies that linearly dependent random variables achieve maximal correlation. Now assume that $Y = cX^2 + bX + a$ and $\mathbb{E}[X] = 0$. Compute again the correlation $\mathrm{Corr}(X, Y)$. Do quadratically dependent random variables still achieve maximal correlation?

Exercise 5 - Basic Concepts (Multivariate Analysis)

a. **(2 points)** Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined as $f(x, y) = x^2 + y^2(1-x)^3$. Show that there exists only one critical point ($\nabla f = 0$) and show that is a strict local minimum. Discuss the question if the fact that the strict local minimum is unique is sufficient to guarantee that it is also the unique global minimum.

b. **(2 points)** Show that the following rules hold for functions $f : \mathbb{R}^d \to \mathbb{R}$, where $w \in \mathbb{R}^d$, $A = (a_{ij}) \in \mathbb{R}^{d \times d}$, $B = (b_{ij}) \in \mathbb{R}^{m \times d}$

$$\begin{aligned}
f(x) &= \langle x, w \rangle, && \implies \quad \nabla f(x) = w, \\
g(x) &= \langle x, Ax \rangle = x^T Ax, && \implies \quad \nabla f(x) = Ax + A^T x, \\
h(x) &= \|Bx\|_2^2, && \implies \quad \nabla f(x) = 2B^T Bx, \\
l(x) &= \|Bx - c\|_2^2, && \implies \quad \nabla f(x) = 2B^T (Bx - c).
\end{aligned}$$

c. **(2 points)** We have a set of n points $x_1, \ldots, x_n$ in $\mathbb{R}^d$. Consider the function $\phi : \mathbb{R}^d \to \mathbb{R}$,

$$\phi(c) = \sum_{i=1}^{n} \|x_i - c\|_2^2 .$$

Compute the critical/stationary points of ϕ ($(\nabla\phi)(c) = 0$). Show that the Hessian ($d \times d$-matrix of second derivatives) of ϕ is positive definite everywhere. What does that imply for the computed critical point ?

Hint:

- $\nabla\phi = \left(\frac{\partial\phi}{\partial c_1}, \ldots, \frac{\partial\phi}{\partial c_d}\right)^T$ is the column vector of the partial derivatives of ϕ.

Exercise 6 - Learning Problem: Predicting Gender from Height

In this exercise we will develop a learning rule that predicts the gender of a person based on his/her height. We need only concepts from probability theory to do this exercise.

Let X be a continuous random variable (denoting the height of a person) which takes values in $\mathbb{R}$. Let Y be a discrete random variable (denoting the gender) which takes values from the set:

{male, female}. Let $P_{X,Y}$ be the unknown joint measure. Suppose we only know the following conditional densities (i.e. we know how the heights of male and female are distributed)[1]:

$$p(x|Y = \text{male}) \sim \mathcal{N}(1.75, 0.1) \qquad \text{and} \qquad p(x|Y = \text{female}) \sim \mathcal{N}(1.65, 0.1)$$

a. **(1 points)** Given the height x of a person, to predict the gender, one could determine the following two probabilities and choose the class with higher probability as the prediction:

$$P(Y = \text{male}|x) \qquad \text{and} \qquad P(Y = \text{female}|x)$$

What other information do we need to find these probabilities? Does one really need to calculate these probabilities to make a prediction?

b. **(5 Points)** Suppose $P(Y = \text{male}) = P(Y = \text{female})$. Solve the following for x:

$$P(Y = \text{male}|x) < P(Y = \text{female}|x)$$

(Bayes rule might be helpful here). What would be your predictions for persons whose heights are given by 1.5, 1.6, and 1.7 and why? What is the probability that your prediction is wrong in these three cases. Is there any height x for which your prediction would always be correct?

Exercise 7 - Phenomena in high dimensions

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 and 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 $\text{vol}(B_d(r))$ 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

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

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 $d!! = d(d-2)\dots 1$ where $1!! = 0!! = 1$.

a. **(2 Points)** We consider now the unit ball $B_d(1)$. Show that for any radius $r < 1$ and $0 < \varepsilon < 1$ one can find a dimension d such that the proportion of $\frac{\text{vol}(B_d(1)\setminus B_d(r))}{\text{vol}(B_d(1))}$ is greater than $1 - \varepsilon$.

b. **(2 Points)** Show that the expected distance from the origin to a point Z uniformly drawn from the unit-ball $B_d(1)$ is

$$\mathbb{E}[\|Z\|] = \frac{d}{d+1}.$$

[1] In practice we do not know even this information; we only have a sample (training set) assumed to be generated from $P_{X,Y}$. We have to **estimate** the appropriate quantities from this sample; the statistical estimation problem will be discussed in later exercises.

Hints:

a. For part b) use the fact that the integral of a radial function, $f(x) = f(\|x\|)$, over the unit ball can be written as

$$\int_{B_d(1)} f(\|x\|)dx = \text{Area}(B_d(1)) \int_0^1 f(r)r^{d-1}dr,$$

where $\text{Area}(B_d(1))$ is the surface area of $B_d(1)$, which is

$$\text{Area}(B_d(1)) = d\,\text{vol}(B_d(1)).$$

Exercise 8 - Loss functions and Bayes optimal functions

- **(3 Points)** Let $\mathcal{Y} = \{-1, 1\}$ (binary classification). Show that the Bayes optimal function, $f^*(x) = \arg\min_{c\in\mathbb{R}} \mathbb{E}[L(Y, c)|X = x]$, for the least squares loss, $L(y, f(x)) = (y - f(x))^2$, is $f^*(x) = \mathbb{E}[Y|X = x]$ and deduce that the least squares loss is classification calibrated.

- **(2 Points)** Let $\mathcal{Y} = \mathbb{R}_+ = \{x \in \mathbb{R}|x \geq 0\}$ (regression with output on the positive part of $\mathbb{R}$) and suppose that $\mathbb{E}[Y|X = x] > 0$. Show that the Bayes optimal function, $f^*(x) = \arg\min_{c\in\mathbb{R}} \mathbb{E}[L(Y, c)|X = x]$ for the loss function $L(y, f(x)) = \log(f(x)) + \frac{y}{f(x)}$, is given by $f^*(x) = \mathbb{E}[Y|X = x]$. Discuss the properties of this loss function compared to least squares loss. What kind of noise model do you think this loss function is useful for (note that the target space is the set of non-negative reals)?