Machine Learning Assignment # 2
Universität Bern

Due date: 11/10/2017

Late submissions will incur a penalty. Submit your answers in ILIAS (as a pdf or as a picture of your written notes if the handwriting is very clear). Submission instructions will be provided via email. You are not allowed to work with others.

Calculus review **[Total 100 points]**

Recall that the Jacobian of a function $f : \mathbb{R}^n \to \mathbb{R}^m$ is an $m \times n$ matrix of partial derivatives

$$Df(x) = \begin{bmatrix} \frac{\partial f_1(x)}{\partial x_1} & \frac{\partial f_1(x)}{\partial x_2} & \dots & \frac{\partial f_1(x)}{\partial x_n} \\ \frac{\partial f_2(x)}{\partial x_1} & \frac{\partial f_2(x)}{\partial x_2} & \dots & \frac{\partial f_2(x)}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m(x)}{\partial x_1} & \frac{\partial f_m(x)}{\partial x_2} & \dots & \frac{\partial f_m(x)}{\partial x_n} \end{bmatrix}$$

where $x = [x_1 \; x_2 \; \dots \; x_n]^\top$, $f(x) = [f_1(x) \; f_2(x) \; \dots \; f_m(x)]^\top$ and $\frac{\partial f_i(x)}{\partial x_j}$ is the partial derivative of the i-th output with respect to the j-th input. When f is a scalar-valued function (*i.e.*, when $f : \mathbb{R}^n \to \mathbb{R}$), the Jacobian $Df(x)$ is a $1 \times n$ matrix, *i.e.,* it is a row vector. Its transpose is called the *gradient* of the function

$$\nabla f(x) = Df(x)^\top = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix}$$

Also, recall that the **chain rule** is a tool to calculate gradients of function compositions. Suppose $f : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at x and $g : \mathbb{R}^m \to \mathbb{R}^p$ is differentiable at $f(x)$. Define the composition $h : \mathbb{R}^n \to \mathbb{R}^p$ by $h(z) = g(f(z))$. Then h is differentiable at x, with Jacobian

$$Dh(x) = Dg(z)\Big|_{z=f(x)} Df(x).$$

1. Answer the following questions by using the above definitions (show all the steps of your working)
   (a) Consider the function $f : \mathbb{R}^n \to \mathbb{R}^m$ and $f(x) = Ax$ where $A \in \mathbb{R}^{m\times n}$. Show that $Df(x) = A$. [10 points]
   (b) Consider the function $g : \mathbb{R}^n \to \mathbb{R}$ and $g(x) = x^\top x$. Calculate $\nabla g(x)$. [10 points]
   (c) Consider the function $h : \mathbb{R}^n \to \mathbb{R}$ and $h(x) = x^\top Q x$, where $Q \in \mathbb{R}^{n\times n}$ is an orthogonal matrix. Calculate $\nabla h(x)$ by using the chain rule, the Jacobian in (a), and without calculating any partial derivative. [20 points]
   (d) Consider the function $f : \mathbb{R}^n \to \mathbb{R}$, where $f(x) = \|Ax - b\|^2$, $A \in \mathbb{R}^{m\times n}$, and $b \in \mathbb{R}^m$. Calculate $\nabla h(x)$ by using the chain rule, the Jacobian in (a), and without calculating any partial derivative. [10 points]
2. Show that $tr(A^2) \le tr(A^T A)$ for any matrix $A \in \mathbb{R}^{m\times m}$. [10 points]
3. Assume $A \in \mathbb{R}^{m\times n}$, $X \in \mathbb{R}^{m\times n}$, and $B \in \mathbb{R}^{m\times m}$. Show that
   - $\nabla_X tr(AX^T) = A$ [10 points]
   - $\nabla_X tr(X^T A) = A$ [10 points]
   - $\nabla_X tr(AX^T B) = BA.$ [10 points]
4. Find the vector $\hat{x} = \arg\min_x \frac{x^T A x}{x^T x}$, where $A \in \mathbb{R}^{m\times m}$ is a real symmetric matrix. [10 points]