# Exercise for Machine Learning (SS 20)

## Assignment 0: Introduction (Theory)

Prof. Dr. Steffen Staab, steffen.staab@ipvs.uni-stuttgart.de

Alex Baier, alex.baier@ipvs.uni-stuttgart.de

Janik Hager, janik-manel.hager@ipvs.uni-stuttgart.de

Analytic Computing, IPVS, University of Stuttgart

Submit your solution in Ilias as either PDF for theory assignments or Jupyter notebook for practical assignments.
Mention the names of all group members and their Email addresses in the file.
**Submission is possible until the following Monday, 27.04.2020, at 14:00.**

This assignment does not count towards the admission. Nonetheless you should try to upload your solution as PDF to Ilias. This will help you and us in testing the platform and identifying potential problems.
The first two tasks were taken from “Exercise 1” of the previous semester held by Prof. Dr. Marc Toussaint.

## 1 Matrix equations

1. Let $X$, Y$A$ be arbitrary matrices, A invertible. Solve for $X$:

   $$XA + A^\top = \mathbf{I}$$

   **Solution:**

   $$\begin{aligned} XA + A^\top &= \mathbf{I} && \vert{-A^\top} \\ XA &= \mathbf{I} - A^\top && \vert{\cdot A^{-1}} \\ X &= (\mathbf{I} - A^\top)A^{-1} \end{aligned}$$

   Note: **I** is the square identity matrix.

2. Let $X$, $A$, $B$ be arbitrary matrices, $(C - 2A^\top)$ invertible. Solve for $X$:

   $$X^\top C = [2A(X + B)]^\top$$

   **Solution:**

   $$\begin{aligned} X^\top C &= [2A(X + B)]^\top \\ X^\top C &= (X + B)^\top 2A^\top \\ X^\top C &= X^\top 2A^\top + B^\top 2A^\top && \vert{-X^\top 2A^\top} \\ X^\top C - X^\top 2A^\top &= B^\top 2A^\top \\ X^\top (C - 2A^\top) &= B^\top 2A^\top && \vert{\cdot(C - 2A^\top)^{-1}} \\ X^\top &= B^\top 2A^\top (C - 2A^\top)^{-1} \end{aligned}$$

3. Let $x \in \mathbb{R}^n$, $y \in \mathbb{R}^d$, $A \in \mathbb{R}^{d\times n}$, $A^\top A$ invertible. Solve for $x$:

   $$(Ax - y)^\top A = \mathbf{0}_n^\top$$

   **Solution:**

   $$\begin{aligned}
   (Ax - y)^\top A &= 0_n^\top \\
   (x^\top A^\top - y^\top)A &= \mathbf{0}_n^\top \\
   x^\top A^\top A - y^\top A &= \mathbf{0}_n^\top && |{+y^\top A} \\
   x^\top A^\top A &= y^\top A && |{\cdot(A^\top A)^{-1}} \\
   x^\top &= y^\top A(A^\top A)^{-1} && |^\top \\
   x &= (A^\top A)^{-1}A^\top y
   \end{aligned}$$

4. As above, additionally $B \in \mathbb{R}^{n\times n}$, $B$ positive-definite. Solve for $x$:

   $$(Ax - y)^\top A + x^\top B = \mathbf{0}_n^\top$$

   **Solution:**

   $$\begin{aligned}
   (Ax - y)^\top A + x^\top B &= \mathbf{0}_n^\top \\
   (x^\top A^\top - y^\top)A + x^\top B &= \mathbf{0}_n^\top \\
   x^\top A^\top A - y^\top A + x^\top B &= \mathbf{0}_n^\top && |{+y^\top A} \\
   x^\top A^\top A + x^\top B &= y^\top A \\
   x^\top(A^\top A + B) &= y^\top A && |(A^\top A + B)^{-1} \\
   x^\top &= y^\top A(A^\top A + B)^{-1} && |^\top \\
   x &= (A^\top A + B^\top)^{-1}A^\top y
   \end{aligned}$$

## 2 Vector derivatives

Let $x \in \mathbb{R}^n$, $y \in \mathbb{R}^d$, $A \in \mathbb{R}^{d\times n}$.

1. What is $\frac{\partial}{\partial x}x$? (Of what type/dimension is this thing?)

   **Solution:**

   The derivative of a vector w.r.t. a vector is a Jacobian matrix:

   $$\begin{aligned}
   \frac{\partial}{\partial x}x &= \left(\frac{\partial x_i}{\partial x_j}\right)_{(i=1,\dots n, j=1,\dots n)} \\
   &= \begin{bmatrix} \frac{\partial x_1}{\partial x_1} & \cdots & \frac{\partial x_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial x_n}{\partial x_1} & \cdots & \frac{\partial x_n}{\partial x_n} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & \cdots & 1 \end{bmatrix} = \mathbf{I}_{n\times n}
   \end{aligned}$$

2. What is $\frac{\partial}{\partial x} x^\top x$?

   **Solution:**

   The derivative of scalar w.r.t. vector is a row vector.

   $$\frac{\partial}{\partial x} x^\top x = [\frac{\partial}{\partial x_i} x_1^2 + \cdots + x_n^2]_{(i=1,\ldots n)} = [2x_i^2]_{(i=1,\ldots n)} = 2x^\top$$

3. Let $B$ be symmetric and positive definite. What is the minimum of $(Ax-y)^\top(Ax-y) + x^\top Bx$ w.r.t. $x$?

   **Solution:**

   The function represents the objective function for linear regression with regularization. Important to note is that the function is convex, which implies that an unique optimum exists. We can identify the $x$, which minimizes the function by deriving the function w.r.t. $x$ and setting to $\mathbf{0}_n$.

   First, we multiply out the function:

   $$\begin{aligned}
   &(Ax-y)^\top(Ax-y) + x^\top Bx \\
   =&(x^\top A^\top - y^\top)(Ax-y) + x^\top Bx \\
   =&x^\top A^\top Ax - x^\top A^\top y - y^\top Ax + y^\top y + x^\top Bx \\
   =&x^\top A^\top Ax - x^\top A^\top y - x^\top A^\top y + y^\top y + x^\top Bx \\
   =&x^\top A^\top Ax - 2x^\top A^\top y + y^\top y + x^\top Bx \\
   =&x^\top(A^\top A + B)x - 2x^\top A^\top y + y^\top y
   \end{aligned}$$

   Then we computes its derivative:

   $$\begin{aligned}
   &\frac{\partial}{\partial x} x^\top(A^\top A + B)x - 2x^\top A^\top y + y^\top y \\
   =&\frac{\partial}{\partial x} x^\top(A^\top A + B)x - \frac{\partial}{\partial x} 2x^\top A^\top y + \frac{\partial}{\partial x} y^\top y \\
   =&\frac{\partial}{\partial x} x^\top A^\top Ax - \frac{\partial}{\partial x} 2x^\top A^\top y + 0 \\
   =&2x^\top(A^\top A + B) - 2y^\top A
   \end{aligned}$$

   Set the derivative to 0 and solve for $x$:

   $$\begin{aligned}
   2x^\top(A^\top A + B) - 2y^\top A &= \mathbf{0}_n && |{+2y^\top A}, \cdot\frac{1}{2} \\
   x^\top(A^\top A + B) &= y^\top A && |\cdot(A^\top A + B)^{-1} \\
   x^\top &= y^\top A(A^\top A + B)^{-1} && |^\top \\
   x &= (A^\top A + B)^{-1} A^\top y
   \end{aligned}$$

## 3 Error Measures

Let $y, \hat{y} \in \mathbb{R}^n$ be $n$ true and predicted values of a regression problem.

1. Formally define the error measures *Mean Squared Error* (MSE) and *Mean Absolute Error* between $y$ and $\hat{y}$.

   **Solution:**

   Let $r_i := y_i - \hat{y}_i$ be the error residual of prediction $i$ for $i = 1, \dots n$. The MSE is defined as:

   $$\text{MSE}(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n} r_i^2$$

   The MAE is defined as:

   $$\text{MAE}(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n} |r_i|$$

2. How would choosing MAE or MSE as objective function for a regression problem impact the resulting prediction model? How do MSE and MAE differ?

   **Solution:**

   For an increasing magnitude in the residual $r_i$, the absolute error grows proportionally with $r_i$. On the other hand, the squared error grows quadratically. In the context of training a model, MSE therefore disproportionally punishes large errors. If large errors are undesirable, which is usually the case, then MSE is preferable to MAE. However, MSE can be sensitive to outliers in the data. Additionally, the MSE is continuously differentiable, which makes it usage with gradient-based optimization algorithms feasible.

3. Let $y, \hat{y} \in \{0, 1\}^n$ be $n$ true and predicted labels of a binary classification problem. What would the MSE and MAE calculate in this case?

   **Solution:**

   From $y, \hat{y} \in \{0, 1\}^n$ follows $r_i \in \{-1, 0, 1\}$ and therefore $r_i^2 = |r_i|$. Consequently, the MSE and MAE are equivalent. $|r_i|$ is 0 if the true and predicted labels $y_i$ and $\hat{y}_i$ match, otherwise it is 0. This corresponds to the 0-1 loss function. Accordingly, we can give an alternative definition of the MSE and MAE for binary class labels:

   $$\begin{aligned} \text{MSE}(y, \hat{y}) &= \frac{1}{n}\sum_{i=1}^{n} [y_i \neq \hat{y}_i] \\ &= \frac{\#(\text{incorrect predictions})}{n} \end{aligned}$$

   where [.] denotes the Iverson bracket, which returns 1 if the bracketed term is true and 0 otherwise. The MSE and MAE therefore compute the ratio of incorrect predictions over all predictions.

   This error is directly related to an important metric for binary classification called accuracy. Accuracy computes the ratio of correct predictions over all predictions. We can directly relate our error measure to accuracy:

   $$\begin{aligned} \text{Accuracy}(y, \hat{y}) &= \frac{\#(\text{correct predictions})}{n} \\ &= \frac{1}{n}(n - \#(\text{incorrect predictions})) \\ &= 1 - \text{MSE}(y, \hat{y}) \end{aligned}$$