Exercise 1:

Identify which type of machine learning (supervised or unsupervised? What type of task?) could be used in these cases:

a) When crossing the alps using the Brenner Autobahn, there is the option to pay electronically in advance. When approaching the toll station, the barrier automatically opens when the number plate was recognised. The recognition happens automatically by a digital camera system.

b) Diagnose whether a patient suffers from cancer or not.

c) The owner of an internet site wants to protect his system against various violations of the terms of service (bot programs, manipulation of timestamps, etc.)

d) An online shopping portal wants to determine products that are automatically offered to registered customers upon login.

e) We want to sort our news into different groups.

f) We want to sort our Email into Spam/Non-Spam.

g) In a supermarket, products that are often bought together are said to be placed side by side on a shelf to increase the sales.

h) We want to extract a list of skills from XING.

i) We want to know our top customers (i. e. highest sales, logistics, etc.).

Exercise 2:

Suppose we observe 6 data pairs and want to describe the underlying relationship between y and x.

x	0.56	0.22	1.7	0.63	0.36	1.2
y	160	150	175	185	165	170

Calculate the β coefficients manually (+ calculator):

a) Assuming a standard linear relationship:
$$y_i = \beta_0 + \beta_1 x_i$$

b) Assuming a non-linear relationship (polynomial degree 2):
$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$$

Exercise 3:

Imagine you work at a bank and have the job to develop a credit scoring model. This means, your model should predict whether a customer applying for a credit will be able to pay it back in the end.

a) Is this a supervised or unsupervised learning problem? Justify your answer.

b) How would you set up your data? Which is the target variable, what feature variables could you think of? Do you need labeled or unlabeled data? Justify all answers.

c) Is this a regression or classification task? Justify your answer.

d) Is this "learning to predict" or "learning to explain"? Justify your answer.

e) In classical statistics, you could use e.g. the logit model for this task. This means we assume that the targets are conditionally independent given the features, so $y^{(i)}|\mathbf{x}^{(i)} \perp y^{(j)}|\mathbf{x}^{(j)}$ for all $i, j = 1, \ldots, n$, where n is the sample size. We further assume that $y^{(i)}|\mathbf{x}^{(i)} \sim Bin(\pi^{(i)})$, where $\pi^{(i)} = \frac{\exp(\boldsymbol{\theta}^\top \mathbf{x}^{(i)})}{1+\exp(\boldsymbol{\theta}^\top \mathbf{x}^{(i)})}$. Looking at this from a Machine Learning perspective, write down the hypothesis space for this model. State explicitly which parameters have to be learned.

f) In classical statistics, you would estimate the parameters via Maximum Likelihood estimation. (The log-Likelihood of the Logit-Model is: $\sum_{i=1}^{n} y^{(i)} \log(\pi^{(i)}) + (1 - y^{(i)})(\log(1 - \pi^{(i)})))$. How could you use the model assumptions to define a reasonable loss function? Write it down explicitly.

g) Now you have to optimize this risk function to find the best parameters and hence the best model. Describe with a few sentences, how you could do this.

Congratulations, you just designed your first Machine Learning project!