Lab Class ML:I, ML:II

By Wednesday, 2017-11-08, solutions for the following exercises have to be submitted: 3, 4, 5a-b, 6.

Exercise 1 : Machine Learning (general)

(a) Define the terms “supervised learning”, “unsupervised learning”, and “reinforcement learning”.

(b) Sketch for each learning paradigm a typical problem together with a description of its technical realization.

Exercise 2 : Machine Learning (general)

(a) Which design decisions are to be made during the development of a learning system?

(b) What is the difference between inductive learning and deductive reasoning (= learning through deduction)?

(c) Name an example of a problem which cannot be solved by learning. Explain your answer.

Exercise 3 : Data

Classify the following attributes as binary, discrete, or continuous. Also classify them as qualitative or quantitative. Some cases may have more than one interpretation, so briefly indicate your reasoning if you think there may be some ambiguity.

(a) Time in terms of AM or PM.

(b) Brightness as measured by a light meter.

(c) Brightness as measured by people’s judgments.

(d) Angles as measured in degrees between $0\,^\circ$ and $360\,^\circ$.

(e) ISBN numbers for books.

(f) Bronze, Silver, and Gold medals as awarded at the Olympics.

(g) Height above sea level.

Exercise 4 : Data

You are approached by the marketing director of a local company, who believes that he has devised a foolproof way to measure customer satisfaction. He explains his scheme as follows: ”It’s so simple that I can’t believe that no one has thought of it before. I just keep track of the number of customer complaints for each product. I read in a data mining book that counts are ratio attributes, and so, my measure of product satisfaction must be a ratio attribute. But when I rated the products based on my new customer satisfaction measure and showed them to my boss, he told me that I had overlooked the obvious, and that my measure was worthless. I think that he was just mad because our best-selling product had the worst satisfaction since it had the most complaints. Could you help me set him straight?”

(a) Who is right, the marketing director or his boss? If you answered, his boss, what would you do to fix the measure of satisfaction?

(b) What can you say about the attribute type of the original product satisfaction attribute?

Exercise 5 : Linear Regression

The table below describes four cars by their age and stopping distance for a full braking at 100km/h till stop:

Car	Wartburg	Moskvich	Lada	Trabi
Age (year)	5	7	15	28
Mileage (km)	30 530	90 000	159 899	270 564
Stopping distance (meter)	50	79	124	300

(a) Determine the weights w_i for the linear regression for the age variable.

(b) Extrapolate the expected average stopping distance for a 15-year old car. Note: use the model from (a).

(c) Consider the mileage of the cars as an additional variable and repeat (a) and (b) under this setting.

(d) Draw a scatter plot of the data and the linear regression for a variable of your choice.

(e) Discuss the problems and pitfalls of extrapolation.

Exercise 6 : P Basic Data Analysis and Linear Regression

In this course, we will use the Python programming language, version 3, to work with data sets and implement fundamental machine learning algorithms. The following exercises will help you prepare for subsequent programming assignments.

(a) Read Sections 1.1. through 1.4. and 1.6. from www.scipy-lectures.org. If you are new to Python, consult other sources, such as docs.python.org/3/tutorial and www.diveintopython3.net, to learn the basics.

(b) Download Fisher's *Iris* data set from www.math.uah.edu/stat/data/Fisher.html. Write a Python program that reads the data set into memory and computes the mean, minimum and maximum of the *petal width*, *petal length*, *sepal width* and *sepal length* attributes for each of the three species of flower. Which of the species will be easy to distinguish, and which will be hard?

(c) Using the `matplotlib` library, draw a scatter plot that shows the *petal length* attribute on the x-axis, and the *sepal length* attribute on the y-axis. Use different colors for the three different species and label the axes.

(d) Create a subset of the Iris data that contains only the *sepal length* attribute, and only the *setosa* and *virginica* classes. Draw a scatterplot showing the attribute on the x-axis and the class on the y-axis. Using the LMS algorithm given in the lecture, compute the weight vector (w_0, w_1), and add the line of best fit to your plot. What is the residual sum of squares (RSS) for your weight vector?