Machine Learning I: Foundations

Prof. Marius Kloft TA: Rodrigo Alves

Due: 03/05/2019 - 10 am

Exercises Sheet 1

1. In this question we will have you do some exploration of the applications of machine learning. One of the goals of the lecture is that you will be able to identify learning problems on your own. To this end, describe three learning problems that have not been mentioned in the lecture, in particular addressing the following aspects:

   a What is the data? (What are the inputs; what are the labels?)

   b What is the goal? Are humans good at solving this task? Why, why not?

   In addition to finding your own machine learning applications we will have you investigate ongoing work in machine learning. Kaiserslautern is a center of technology in Germany and has been touted as the "Silicon Valley" of Germany. Because of this there are many machine learning and artificial intelligence companies nearby. We would like you to do a bit of your own research into these companies:

   c Find 2-3 AI/ML companies in Kaiserslautern and give a short description of them.

   d Choose one of these companies and visit it. Please describe what you learned during your visit. These visits can be very helpful for getting a deeper understanding of the company's approach to the problems they are trying to solve, provide inspiration for project ideas, and perhaps yield future networking opportunities.

2. Let H be a hyperplane defined by the affine-linear function $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$ ($\mathbf{w} \in \mathbb{R}^n$, $\mathbf{x} \in \mathbb{R}^n$ and $b \in \mathbb{R}$). Show that signed distance[1] between H and $\mathbf{x}^* \in \mathbb{R}^n$ can be defined by:

$$d(\mathbf{x}^*, H) = \frac{\mathbf{w}^\top \mathbf{x}^* + b}{\|\mathbf{w}\|}.$$

3. The Disney Demography Institute made a survey with Disneyland population to check the distribution of weight and height among the citizens genders. The result is disposed in the file "DWH_Training.csv". This dataset is composed of three columns: the first one is the height (in cm), the second is the weight (in kg) and the third one the gender (1 = Male and -1=Female) of the citizens.

   (a) Make a scatter plot where the x-axis is the height of the citizens and the y-axis is the weight of the citizens. The color of the points need to be different for males and females.

   (b) Draw a horizontal line for which you think that it best separates male and female citizens (you don't need to calculate it).

   (c) One of the parents of Mr Duck Tales weighs 62 kg. Would you say that she/he is his mother or his father? Can you guarantee that you are right in your prediction? Why?

   (d) Draw a vertical line for which you think that it best separates male and female citizens (again, you don't need to calculate it).

   (e) Mrs Minnie Mouse has two siblings. One of them is 181 tall. Would you say that she/he is his brother or his sister? Can you guarantee that you are right in your prediction? Why?

   (f) Draw the line (once again, you don't need to calculate it) for which you think that it best separates male and female citizens.

   (g) Donald Duck has a lot of cousins. One of them weights 52kg and is 1.79cm tall. How would you classify his cousin? Would you classify differently if you use the (b) or (d) lines?

[1] Distances are positive by definition. Consider P any point of the hyperplane. In this case, the sign of the distance is related to angle between $\overrightarrow{Px^*}$ and $\overrightarrow{PA}$ (a vector that starts at P and have the same direction as the normal vector)

4. We will implement our first machine learning algorithm! Our aim is to propose a model to classify automaticaly the gender of a Disneyland citizen given her/his weight and height. For this we will use the file "DWH_Training.csv" to train the model and use the file "DWH_Test.csv" to check how accurate this model is. In $\mathbb{R}^2$ a hyperplane can be defined by an affine-linear function $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$ ($\mathbf{w} \in \mathbb{R}^2$, $\mathbf{x} \in \mathbb{R}^2$ and $b \in \mathbb{R}$). Geometrically, it can be seen as a line. To simplify, consider that $\mathbf{w} \in \{(0.576, 0.047)\}$ and $b \in \{-103, -102, -101, -100, -99\}$.

   (a) Repeat the Exercise 2 (a). Add to this scatter plot all nine possible hyperplanes that combine the possible values of $\mathbf{w}$ with the possible values of b $[\{(0.576, 0.047), -103\}, \{(0.576, 0.047), -102\}, \cdots, \{(0.576, 0.047),$

   (b) Using Exercise 1, calculate the distance between each possible hyperplane and each citizen of the training data.

   (c) If the signed distance is negative then classify the citizen as Male. Else, classify as Female.

   (d) Which hyperplane best classifies the training set?

   (e) With the hyperplane selected in the item (d), classify the test set("DWH_Test.csv"). What was the accuracy? Proportionally, is it better or worse than the accuracy found for the training set?

5. In slide 47 (Lecture 1) we present the k-nearest neighbor learning algorithm, a very useful and surprisingly powerful algorithm. Let's implement it. For this task you should not use machine learning libraries.

   (a) Implement the algorithm leaving k as parameter

   (b) Consider $k \in \{3, 5, 20\}$. Using the dataset "DWH_Training.csv", use 10-fold crossvalidation (with $t = 1$) to find the best value of the parameter k for the algorithm implemented in (a). See the slide 49. **Attention:** note that the "k" in the k-nearest neighbor algorithm is different from the "k" in "t-times k-fold cross validation". In the later case, observe that $k = 10$.