Florian Rösel, Prof. Dr. Frauke Liers Winter semester 2021/2022

Mathematics of Learning – Worksheet 5

Basics [Big-O (Landau) Notation.]
a) Express the relationships of the functions n^{1000}, 2^n, e^n, e^{n^2}, $n!$ and n^n with the help of Landau's notation (i.e., prove, if $f \in \mathcal{O}(g), f \in \Omega(g)$ or $f \in \Theta(g)$ for every pair of the functions above). Prove your statements.

b) An equivalence relation is a homogeneous relation over some set M[1], which is

1. Reflexive: $m_1 \simeq m_1$,
2. Symmetric: $m_1 \simeq m_2$ implies $m_2 \simeq m_1$,
3. Transitive: If $m_1 \simeq m_2$ and $m_2 \simeq m_3$ then $m_1 \simeq m_3$

for all $m_1, m_2, m_3 \in M$. For example, the a relation $\simeq$ on $\mathbb{R}^n$, $v \simeq w$ if and only if $v_1 = w_1$ is an equivalence relation.
Prove or disprove, that $\simeq_L$ which we define as

$$f \simeq_L g \text{ if and only if } f \in \Theta(g)$$

is an equivalence relation on the set of mappings from $\mathbb{N}$ to $\mathbb{R}_{>0}$.

Exercise 1 [Reading assignment: Supervised learning].
Read chapter 2 of the Hastie Book. It gives you a good overview of supervised learning, what will be the content of the course in the next weeks. Discuss the contents of the chapter with a fellow student for at least half an hour.

Exercise 2 [Regression].
The regression problem takes as input data N data points of *explanatory* or *independent* vectors $X_i \in \mathbb{R}^p$, and some *response* or *dependent* reals $Y_i \in \mathbb{R}$. The goal of regression is now, to express the response variables as good as possible as a function of the explanatory variables, i.e., to find a function $g\colon \mathbb{R}^p \to \mathbb{R}$ such that $g(X_i) \approx Y_i$. Since

$$\tilde{g}\colon x \to \begin{cases} Y_i, if x = X_i \\ 0 \text{ otherwise} \end{cases}$$

would perfectly do it, but presumably has terrible out of sample behavior, we cannot allow the whole space of functions as candidates (type "overfitting" in a search engine of your choice). Rather, we'd make use of a set of pre-specified predictor functions, i.e., $f_j\colon \mathbb{R}^p \to \mathbb{R}$ which are simple or natural in some sense, e.g., $f_1(x) = x_1$, $f_2(x) = x_1 \cdot x_2$ or $f_3(x) = e^{x_1}$, and try to linearly combine a function of these to predict the response variables as good as possible. No matter which set of predictor funcitons we choose, this leads to a minimization problem with linear constraints, and, depending on how we define "as good as possible fit the response variables" to different objective

[1] i.e. $\simeq\colon M \times M \to \{0,1\}$; one would rather write $m_1 \simeq m_2$ instead of $\simeq (m_1, m_2) = 1$ and $m_1 \not\simeq m_2$ instead of $\simeq (m_1, m_2) = 0$.

functions, but classically the quality of approximation is defined as minimizing the euclidean norm of approximation error vector, leading to a convex-quadratic objective functions (what is going to be the standard in the following).

Hence, confronted with $N \in \mathbb{N}$ data vectors $X_i \in \mathbb{R}^p$ and $Y_i \in \mathbb{R}$, to be fitted, the procedure is the following:

1. Choose (guess; depending on what kind of relation you expect between X and Y) functions $f_1, ..., f_m$, mapping from $\mathbb{R}^p$ to $\mathbb{R}$
2. Calculate the matrix A, $A_{i,j} := f_j(X_i)$.
3. Solve the minimization problem

$$\min_{\beta \in \mathbb{R}^m} ||A\beta - Y||_2. \tag{1}$$

The solution $\tilde{\beta}$ of the optimization problem can be interpreted as the coefficients of the predictor functions: Read it as "Y can be expressed optimally as $\beta_1 f_1(X) + \cdots + \beta_m f_m(X)$".

a) Prove: A vector $\tilde{\beta} \in \mathbb{R}^m$ solves the optimization problem 1, if and only if it solves the linear system

$$A^T A\beta = A^T Y$$

b) Derive, that the solution is unique, if and only if A has full column rank.

Exercise 3 [Implementing regression].
a) Implement the regression algorithm described in the previous exercise.
Find out the function generating the points which are presented in the file "regression.csv". To generate the response points from the exploratory points, I used a few (but not all) of the functions $1, x, x^2, x^3, e^x, ln(|x| + 1), sqrt(|x|), sin(x), cos(x), tan(x)$ and a normal distributed perturbation to linearly combine the response values. Use half of the data points as training set and half of the data points as validation set. Hint: It could be necessary to first invest a little bit of thinking which of the prediction functions can be excluded beforehand.

Exercise 4 [Apply techniques learned so far to (more) realistic data sets]. Download the yale faces data set at the StudOn Platform. Interpret them as grayscale vectors and apply everything what you learned so far (apply k-means-clustering and EM-clustering with known and unknown k on the data set with full and (linearly and kernel) reduced dimension). Visualize, interpret and discuss your results.