Mathematics of machine learning (winter term 2023/24)

Exercise sheet II

November 12th, 2024

1. Task. (Convexity of logistic regression)
A function $f\colon \mathbb{R}^d \to \mathbb{R}$ is called *convex*, if for any $\lambda \in [0,1]$

$$f(\lambda x + (1-\lambda)\tilde{x}) \leq \lambda f(x) + (1-\lambda) f(\tilde{x}), \qquad \forall x, \tilde{x} \in \mathbb{R}^d.$$

Convexity is very advantageous for minimization.

If $f\colon \mathbb{R}^d \to \mathbb{R}$ is differentiable, then it is convex iff

$$f(x) \geq f(\tilde{x}) + \nabla f(\tilde{x})(x - \tilde{x}), \qquad \forall x, \tilde{x} \in \mathbb{R}^d.$$

If f is twice differentiable, then it is convex iff its second derivative (its *Hessian* matrix) is positive semidefinite:

$$\nabla^2 f(x) \geq 0 \qquad \forall x \in \mathbb{R}^d.$$

a) Show that

$$g(t) := \ln(1 + \exp(-t)), \qquad t \in \mathbb{R}$$

is convex.

b) Argue, that the log-loss

$$\ell(h_{\mathbf{w},b}, (\mathbf{x}, y)) = \ln(1 + \exp(-y(\mathbf{w} \cdot \mathbf{x} + b)))$$

is convex w. r. t. $\mathbf{w}$ and b. Use that $(\mathbf{w}, b) \mapsto y(\mathbf{w} \cdot \mathbf{x} + b)$ is linear.

Then derive that the empirical risk $\mathcal{R}_s(h_{\mathbf{w},b})$ based on the log-loss is convex w. r. t. $\mathbf{w}$ and b. To this end, use and prove the fact, that the sum of convex functions is again convex.

c) Show that a convex function can have only global minima. Give an example of a convex function which has *no* minima.

Draw conclusions for the training of logistic regression.

d) Show that the empirical risk for logistic regression has *no* minimum if the training data is *linearly separable*.

2. Task. (Credit default (incl. programming task))
We consider the problem of credit assessment: Based on the monthly gross income x of a customer, we want to predict if a loan of 200,000 euros will be paid back within 15 years time ($y = 1$) or whether it will be default ($y = 0$). As hypotheses class we simply use

$$\mathcal{H} = \left\{ h\colon \mathbb{R} \to \{0,1\} \mid h(x) = \mathbf{1}_{[a,+\infty)}(x),\ a \in \{a_1, \ldots, a_n\} \right\}$$

with predefined income tresholds $a_1, \ldots, a_n > 0$.

We assume that the gross income of our customers is uniformly distributed between 3000 and 7000 Euros. Furthermore, we assume that there is a true hypothesis $h^\dagger \in \mathcal{H}$ describing the deterministic relation between credit default $y = 0$ and gross income x being below a treshold $a^\dagger$. We aim to learn this true hypothesis using empirical risk minimization (ERM) based on finitely many training data $(x_1, y_1), \ldots, (x_m, y_m)$ and want to answer the following question:

"How many data do we need in order to learn with a probability of at least 90% a hypothesis h_s which will predict future credit defaults correctly in at least 95 out of 100 cases?"

a) Use a result from the lecture to answer this question.

b) Assume now that $a_1 = 4400$, $a_2 = 4600$, $\ldots$, $a_9 = 6000$, $a_{10} = 6200$ as well as $a^\dagger = 5400$. Compute the risks $\mathcal{R}_\mu(h_i)$ of $h_i = \mathbf{1}_{[a_i,+\infty)}$.

c) Given $a_1, \ldots, a_{10}$ and $a^\dagger$ as before compute the probability that $\mathcal{R}_S(h_i) = 0$ for all $i = 1, \ldots, 10$.

d) Based on the results in b) and c) derive a new bound for the number of data m answering the quesion above.

e) Use the provided Jupyter Notebook `credit_sim_teach` in order to check the two bounds from a) and d) empirically by performing the supervised learning based on m data points for $M = 1,000$ repetitions.

f) Which of the two bounds from a) and d) would still be valid if the distribution of the gross income x changes?

Homework:
Apply logistic regression to the *heart* data set for the classification of patients with heart diseases.

a) Download the data set from http://archive.ics.uci.edu/ml/datasets/Statlog+%28Heart%29 and extract the six real-valued characteristics (see data set description) and the labels (last column) from the data set.

b) Divide the datas et into a *training dataset* (70% of the data) and a *test dataset* (30% of the data).

c) Use the training data to learn a hypothesis from $\mathcal{L}_{d,\text{sig}}$ using *logistic regression*.

d) How many of the training data are misclassified? Using the test data, also estimate the resulting expected risk with respect to the 0-1 loss.

e) Is the training sample linearly separable or not? How can you tell?