Introduction to Machine Learning (WISB 365)
Final Exam

Sjoerd Dirksen

13 April 2022, 13:30-16:30

This exam consists of 4 questions, worth 45 points in total. Please write your name and student number on every sheet of your solutions.

Question 1 [10 points]

In a dystopian future, the lecturer of *Introduction to Machine Learning (WISB365)* is using a linear classifier to determine whether students may participate in the course. He asks each student for a list $x \in \mathbb{R}^d$ of grades for the d first-year courses. A student is admitted to the course if

$$h(x) = \text{sign}(\langle w, x\rangle + b)$$

is equal to 1 and not admitted otherwise. The lecturer has posted $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ on his website, so that any student can check whether he/she is eligible for the course.

(a) Lieke finds that she will not be admitted based on her grade list x_L, but she is really eager to take the course. She wants to make some minimal changes to her grade list in order to get admitted. Explain that she can achieve this by solving

$$\min_{s\in\mathbb{R}^d} \frac{1}{2}\|s\|_2^2 \qquad \text{s.t.} \qquad \langle w, x_L + s\rangle + b = 0. \tag{1}$$

(b) Solve (1).

(c) The solution in (b) turns out to lead to a change of all grades on Lieke's grade list. She is worried that the lecturer may notice these changes. To minimize the risk of discovery, she would prefer it if a minimal number of grades on the list are changed. How could she modify the problem (1) to achieve this?

Question 2 [12 points]

Let $y \in \{-1, 1\}^m$ and let $X \in \mathbb{R}^{m\times d}$ be the data matrix with rows x_i^T, $i = 1, \ldots, m$. Let $f : \mathbb{R}^d \to \mathbb{R}$ denote the objective function of the logistic regression problem, i.e.,

$$f(w) = \sum_{i=1}^{m} h(y_i\langle w, x_i\rangle),$$

where $h : \mathbb{R} \to \mathbb{R}$ given by $h(t) = \log(1 + e^{-t})$.

(a) Show that f is convex.

(b) Prove that

$$\nabla f(w) = X^T p(w)$$

for any $w \in \mathbb{R}^d$, where $p(w) \in \mathbb{R}^m$ is given by

$$p(w)_i = -y_i \frac{1}{1 + \exp(y_i\langle w, x_i\rangle)}, \qquad 1 \le i \le m.$$

(c) Show that f is β-smooth for $\beta = \frac{1}{4}\sigma_{\max}(X^T)\sigma_{\max}(X)(= \frac{1}{4}\sigma_{\max}(X)^2)$, where $\sigma_{\max}$ denotes the largest singular value.

Hint: express $p(w)$ in terms of the logistic function $\sigma(t) = (1 + e^{-t})^{-1}$.

(d) What can you say about the convergence of gradient descent for the logistic regression problem?

Question 3 [10 points]

For any kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, we define its *normalization* $K_n : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ by

$$K_n(x, x') = \begin{cases} 0 & \text{if } K(x,x) = 0 \text{ or } K(x',x') = 0 \\ \frac{K(x,x')}{\sqrt{K(x,x)K(x',x')}} & \text{else.} \end{cases}$$

(a) Show that the Gaussian kernel is the normalization of the exponential kernel.

(b) Show that if K is a positive definite kernel, then K_n is a positive definite kernel as well.

Question 4 [9 points]

Let $\bar{\sigma} : \mathbb{R} \to [0, \infty)$ be the rectified linear unit (ReLU). Prove that for any 1-Lipschitz function $f : [0,1] \to \mathbb{R}$ there is some ReLU neural network of the form

$$\varphi(x) = w_2 \sigma(w_1 x + b_1)$$

with $w_1, w_2^T, b_1 \in \mathbb{R}^{2d}$ such that

$$\sup_{x \in [0,1]} |f(x) - \varphi(x)| \leq \frac{4}{d}.$$

Hint: how can one create a sigmoidal function using ReLU functions?

Question 5 [4 points]

Jim is trying to using gene marker data to learn a predictor that can reliably predict whether a given person will develop Alzheimer's disease. As the number of gene markers for any person runs in the millions, he uses PCA to reduce the dimensionality of his dataset. Afterwards, he makes a 60-30-10 train-validation-test split of his database of test subjects. He trains any given neural network by minimizing

$$\frac{1}{m} \sum_{i=1}^{m} \log(1 + \exp(-y_i \varphi_{W_L,\sigma_L} \circ \cdots \circ \varphi_{W_1,\sigma_1}(x_i)))$$

using stochastic gradient descent with backpropagation and random initialization. Denoting the output of this procedure by $W_L^*, \ldots, W_1^*$, he then uses the classifier

$$h_{W_L^*,\ldots,W_1^*}(x) = \sigma_{L+1}(\varphi_{W_L,\sigma_L} \circ \cdots \circ \varphi_{W_1,\sigma_1}),$$

where $\sigma_{L+1} : \mathbb{R} \to [0,1]$ is the logistic function. He assigns any datum x the label 1 if and only if $h_{W_L^*,\ldots,W_1^*}(x) \geq \frac{1}{2}$. He compares 20 different architectures by computing the fraction of misclassifications on his validation set. He then evaluates the effectiveness of the best performing architecture by computing the fraction of misclassifications on his test set.

Due to a mistake in Jim's setup, the result of his method is not reliable. Which mistake did Jim make?