

DEEP LEARNING (SS2020)
SEMINAR 1

Assignment 1. Let X be a real valued random variable with expectation $\mathbb{E}X$ and finite variance $\mathbb{V}X$. The Chebyshev inequality asserts

$$\mathbb{P}(|X - \mathbb{E}X| > \varepsilon) \leq \frac{\mathbb{V}X}{\varepsilon^2}.$$

Let $X_i, i = 1, \dots, m$ be independent, identically distributed random variables with expectation $\mathbb{E}X$ and finite variance $\mathbb{V}X$ and let $Y = \frac{1}{m} \sum_{i=1}^m X_i$ be their empirical mean. Prove the inequality

$$\mathbb{P}(|Y - \mathbb{E}Y| > \varepsilon) \leq \frac{\mathbb{V}X}{m\varepsilon^2}.$$

Assignment 2. Let $X_i, i = 1, \dots, m$ be independent random variables bounded by the interval $[a, b]$, i.e. $a \leq X_i \leq b$. Let $X = \frac{1}{m} \sum_{i=1}^m X_i$ be their empirical mean. The Hoeffding inequality asserts that

$$\mathbb{P}(|X - \mathbb{E}X| > \varepsilon) \leq 2 \exp\left(-\frac{2m\varepsilon^2}{(b-a)^2}\right).$$

Let us now consider a predictor $h: \mathcal{X} \rightarrow \mathcal{Y}$, and a loss $\ell(y, y')$. The risk of the predictor is denoted by $R(h)$ and its empirical risk on a test set $\mathcal{T}^m = \{(x^j, y^j) \mid j = 1, \dots, m\}$ is denoted by $R_{\mathcal{T}^m}(h)$.

a) Prove that the generalisation error of h can be bounded in probability by

$$\mathbb{P}\left(|R(h) - R_{\mathcal{T}^m}(h)| > \varepsilon\right) < 2e^{-\frac{2m\varepsilon^2}{(\Delta\ell)^2}}, \quad (1)$$

where $\Delta\ell = \ell_{max} - \ell_{min}$.

b) Verify the value m given in Example 2. of Lecture 1. for the special case of a binary classifier and the 0/1-loss.

c) Consider an age predictor as described in Example 1. of Lecture 1. but with $\mathcal{Y} = \{0, 1, \dots, 100\}$ and with the loss $\ell(y, y') = |y - y'|$. What is the minimal number of test examples m needed to guarantee that $R(h)$ is in the interval $(R_{\mathcal{T}^m}(h) - 1, R_{\mathcal{T}^m}(h) + 1)$ with probability at least 95%?

Assignment 3. We want to utilise the Hoeffding inequality for choosing the best predictor from a finite set of predictors \mathcal{H} . Denoting the r.h.s. of (1) by δ , we interpret it as follows. Among all possible test sets \mathcal{T}^m of size m there are at most $\delta * 100$ percent “bad” test sets for a given predictor h . We call a test set \mathcal{T}^m bad for the predictor h if $|R(h) - R_{\mathcal{T}^m}(h)| > \varepsilon$.

a) Conclude that the percentage of test sets, which are bad for at least one $h \in \mathcal{H}$ can be bounded by

$$\mathbb{P}\left(\max_{h \in \mathcal{H}} |R(h) - R_{\mathcal{T}^m}(h)| > \varepsilon\right) < 2|\mathcal{H}|e^{-\frac{2m\varepsilon^2}{(\Delta\ell)^2}}$$

b) Suppose, we want to apply this for choosing the best predictor from a set of 10 predictors based on 0/1 loss. What is the minimal size m of a test set \mathcal{T}^m , which ensures that in 95% of cases our choice $h_m = \arg \min_{h \in \mathcal{H}} R_{\mathcal{T}^m}(h)$ will fulfill

$$|R(h_m) - R(h_{\mathcal{H}})| < \varepsilon,$$

where $h_{\mathcal{H}}$ is the best predictor from \mathcal{H} ?

Assignment 4. Suppose that the decision boundary of a binary classifier for points $x \in \mathbb{R}^n$ is given by a convex polyhedron. Show that the classifier can be implemented by a two layer network with binary output neurons.

Show that decision boundaries given by arbitrary polyhedra can be implemented by three layer networks with binary output neurons.

Assignment 5. Consider a neural network with K outputs y_k representing posterior class probabilities. The last layer of this network is a softmax layer with output

$$y_k = \frac{e^{x_k}}{\sum_{\ell} e^{x_{\ell}}},$$

where x_k are the outputs of the penultimate layer and represent class scores. When learning such a network by maximising the log conditional likelihood, we have to consider log-probabilities

$$z_k = \log y_k = x_k - \log \sum_{\ell} e^{x_{\ell}}$$

We will analyse the nonlinear part of the r.h.s

$$f(x) = \log \sum_{\ell} e^{x_{\ell}}$$

a) Prove that its gradient is given by $\nabla f(x) = y$, i.e. by the vector of class probabilities. Conclude that the norm of the gradient is bounded by 1.

b) Compute the second derivative of f and show that it can be expressed as

$$\nabla^2 f(x) = \text{Diag}(y) - yy^T.$$

Prove that this matrix is positive semi-definite and conclude that $f(x)$ is a convex function.