DEEP LEARNING (SS2022)
SEMINAR 1

Assignment 1. Consider a neuron

$$y = f\Big(\sum_{i=1}^{n} w_i x_i + b\Big)$$

for inputs $x \in \mathbb{R}^n$ with weights $w \in \mathbb{R}^n$, bias $b \in \mathbb{R}$ and activation function f. Show that the affine mapping given by (w, b) can be replaced by a linear mapping if we extend the input space by one dimension. Give a geometric interpretation.

Assignment 2 (Softmax)**.** Show the following properties of the function

$$\mathrm{softmax}\colon \mathbb{R}^n \to \mathbb{R}^n_+ \colon s \mapsto \frac{e^{s_i}}{\sum_j e^{s_j}}$$

a) $\mathrm{softmax}$ is invariant to adding the same number to all scores s

b) $\arg\max_i \mathrm{softmax}(s)_i = \arg\max_i (s_i)$

c) Consider a classification problem with features $x \in \mathbb{R}^n$ and two classes $\{+1, -1\}$. A neural network for this problem may be designed to output the vector of class probabilities $(p, 1-p)$ for classes $+1, -1$, respectively, as

$$\mathrm{softmax}(Wx + b), \tag{1}$$

using weights $W \in \mathbb{R}^{2\times n}$, $b \in \mathbb{R}^2$. Another network for the same problem is designed to output real-valued scores

$$a = v^{\mathsf{T}} x + c, \tag{2}$$

using weights $v \in \mathbb{R}^n$, $c \in \mathbb{R}$. It defines the predictive probability as $p(y{=}1 \,|\, x) = \mathrm{sigmoid}(a)$. Given such a network, find a network of the form (1) that outputs equivalent predictive probabilities.

Assignment 3. Consider a two-layer network

$$x^2 = F(x^0) = f \circ A^2 \circ f \circ A^1 x^0$$

with affine mappings $A^k x^{k-1} = W^k x^{k-1} + b^k$, $k = 1, 2$ and element-wise activation function f.

a) Assume that the activation function f is the identity mapping $f\colon x \to x$. Show that the network is equivalent to a network with only one affine layer.

b) Assume that the activation function is ReLU, i.e. $f(x) = \max(0, x)$. Show that re-scaling $(W^1, b^1) \to (\lambda W^1, \lambda b^1)$ and $(W^2, b^2) \to (\lambda^{-1} W^2, b^2)$ with some positive λ keeps the network mapping F unchanged.

Assignment 4. Let us consider the logistic regression model

$$p(y \mid x; w) = \mathrm{S}(y w^T x),$$

where $y = \pm 1$ is the class, $x \in \mathbb{R}^n$ is the feature vector, $w \in \mathbb{R}^n$ is a parameter vector and S denotes the logistic sigmoid function. Given training data $\mathcal{T}^m = \{(x_j, y_j) \mid j = 1 \ldots m\}$, we want to estimate w by maximising the (conditional) log-likelihood $\mathbb{E}_{\mathcal{T}^m} \log p(y \mid x; w)$.

a) Let us assume that the training data are linearly separable. Show that in this case the logistic regression problem has no finite optimal solution.

Hint: show that for any w that achieves a correct classification taking $w' = \alpha w$ with $\alpha > 1$ achieves a higher likelihood.

b) Show that adding the regularizer on the weight norm $\lambda \|w\|^2$ with some $\lambda > 0$ fixes this problem.

Assignment 5 (Hopfield network)**.** Let us consider a fully connected recurrent network with n binary neurons. Denoting their outputs by $x_i = \pm 1$, the corresponding dynamical system reads

$$x_i(t+1) = \operatorname{sign}\Big(\sum_{j \neq i} w_{ij} x_j(t)\Big),$$

We will assume sequential updates in some fixed order over the neurons. Let us further assume that the weight matrix is symmetric, *i.e.* $w_{ij} = w_{ji}$.

Prove that the function

$$\mathcal{H}(x) = -\frac{1}{2} \sum_{i,j} x_i w_{ij} x_j = -\frac{1}{2} x^T W x$$

is a Ljapunov function for this dynamical system, *i.e.* it can decrease only:

$$\mathcal{H}(x(t+1)) \leq \mathcal{H}(x(t)).$$

Conclude that the network will eventually reach a fixpoint configuration.

Hint: express the change of $\mathcal{H}(x)$ for an update of a single neuron x_i.