Johannes Gutenberg-Universität Mainz
Institut für Informatik
Data Mining
Prof. Dr. Stefan Kramer

Machine Learning - Sheet 7

02.07.2020
Deadline: 12.07.2020 - 12:00

Task 1: MLP Bipartite Connections, Biases, Sigmoid Layers *(9 Points)*

In this exercise, we focus on the most common components of artificial neural networks. In the following, let n, m, N be some natural numbers.

(1) (*3 points*) **Full bipartite connection layer:** Let $D \in \mathbb{R}^{N\times n}$ and $W \in \mathbb{R}^{n\times m}$ be two matrices. Suppose that in the forward pass, the matrix multiplication node gets D and W as inputs, and outputs their matrix product $A := DW$, that is, $A \in \mathbb{R}^{N\times m}$ with $A_{ij} := \sum_{q=1}^{n} D_{iq}W_{qj}$. In the backward pass, the node receives the backpropagated error $B \in \mathbb{R}^{N\times m}$ with $B_{ij} = \frac{\partial E}{\partial A_{ij}}$. Prove[1]:

$$\frac{\partial A_{ij}}{\partial W_{kl}} = D_{ik}\delta_{jl}, \qquad \frac{\partial E}{\partial W_{kl}} = (D^\top B)_{kl}, \qquad \frac{\partial A_{ij}}{\partial D_{kl}} = \delta_{ik}W_{lj}, \qquad \frac{\partial E}{\partial D_{kl}} = (BW^\top)_{kl}.$$

(2) (*3 points*) **Bias layer:** Suppose that $D \in \mathbb{R}^{N\times n}$ and $b \in \mathbb{R}^n$. Suppose that $A \in \mathbb{R}^{N\times n}$ with $A_{ij} = D_{ij} + b_j$. Let $B \in \mathbb{R}^{N\times n}$ be the backpropagated error, that is: $B_{ij} = \frac{\partial E}{\partial A_{ij}}$. Show:

$$\frac{\partial A_{ij}}{\partial D_{kl}} = \delta_{ik}\delta_{jl}, \qquad \frac{\partial E}{\partial D_{kl}} = B_{kl}, \qquad \frac{\partial A_{ij}}{\partial b_k} = \delta_{kj}, \qquad \frac{\partial E}{\partial b_k} = \sum_i B_{ik}.$$

(3) (*3 points*) **Sigmoid layer:** The input is $D \in \mathbb{R}^{N\times n}$. The activation is $A \in \mathbb{R}^{N\times n}$ with $A_{ij} := \sigma(D_{ij})$, where $\sigma(t) := 1/(1+e^{-t})$ is the *sigmoid function*. Let $B \in \mathbb{R}^{N\times n}$ be the backpropagated error, i.e., $B_{ij} = \frac{\partial E}{\partial A_{ij}}$. Show that $\sigma'(t) = \sigma(t)(1-\sigma(t))$, and $\frac{\partial E}{\partial D_{kl}} = B_{kl}A_{kl}(1-A_{kl})$.

Task 2: Maximum Margin Hyperplane *(2 Points)*

Show that the value ρ of the margin for the maximum-margin hyperplane is given by

$$\frac{1}{\rho^2} = \sum_{n=1}^{N} a_n,$$

where $\{a_n\}$ are given by maximizing

$$\tilde{L}(\mathbf{a}) = \sum_{n=1}^{N} a_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m y_n y_m k(\mathbf{x}_n, \mathbf{x}_m)$$

subject to the constraints $a_n \geq 0$, $n = 1, \ldots, N$, and $\sum_{n=1}^{N} a_n y_n = 0$.

[1] Here, δ_{ij} stands for the Kronecker Delta. This means: δ_{ij} is 1 iff $i = j$, and 0 otherwise. Hint: if $\xi(i)$ is some expression that depends on the index i, then $\sum_i \xi(i)\delta_{ij} = \xi(j)$.

Task 3: Support Vector Machines: Kernels *(9 Points)*

For any non-empty set $\mathcal{X}$ a kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is said to be *positive semi-definite*, if the following conditions hold for all $x_1, \ldots, x_m \in \mathcal{X}$:

- $k(x_i, x_j) = k(x_j, x_i)$ for all x_i, x_j (symmetry)
- $\forall c_1, \ldots, c_m \in \mathbb{R}$: $\sum_{i,j=1}^{m} c_i c_j k(x_i, x_j) \geq 0$

Every positive semi-definite kernel can be represented as a dot product in a linear space, thus allowing for the *kernel trick*.

(a) (*2 points*) Show that the dot product $k : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}, k(x, y) := \sum_{i=1}^{n} x_i y_i$ is a positive semi-definite kernel.

(b) (*2 points*) Show that the polynomial kernel $k : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}, k(x, y) := (\sum_{i=1}^{n} x_i y_i)^2$ is a positive semi-definite kernel.

Now, we want to build more complex kernels from simpler ones. Suppose that *we already know* that if $k, k_1, k_2 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ are kernels, then

- $k_1 + k_2$ (i.e, $(x, y) \mapsto k_1(x, y) + k_2(x, y)$)
- $k_1 \cdot k_2$ (i.e, $(x, y) \mapsto k_1(x, y) \cdot k_2(x, y)$)
- $\exp \circ k$ (i.e, $(x, y) \mapsto \exp(k(x, y))$)

are also kernels. Relying only on the definition of the kernel and these three "rules", show that the following functions are also kernels:

(c) (*1 point*) Let $d \in \mathbb{N}$ be some exponent, and k a kernel. Show that k^d, i.e. $(x, y) \mapsto (k(x, y))^d$ is a kernel.

(d) (*2 points*) Let $n \in \mathbb{N}$, $c_0, \ldots, c_n \in \mathbb{R}_{\geq 0}$, let k be a kernel. Show that $\sum_{i=0}^{n} c_i \cdot k^i$ is also a kernel.

(e) (*2 points*) Let $f : \mathcal{X} \to \mathbb{R}$ an arbitrary function. Show that $(x, y) \mapsto f(x)k(x, y)f(y)$ is a kernel.