Faculty of Computer Science, Heinrich-Heine-University Düsseldorf

Deep Learning

Winter term 25/26 – Exercise Sheet 1

1. **Differentiation rules**
   a) $\frac{d}{dx}x^n = nx^{n-1}$
   b) $\frac{d}{dx}(f(x)+g(x)) = \frac{d}{dx}f(x) + \frac{d}{dx}g(x) = f'(x) + g'(x)$
   c) $\frac{d}{dx}f(x)g(x) = \frac{df(x)}{dx}g(x) + f(x)\frac{dg(x)}{dx} = f'(x)g(x) + f(x)g'(x)$ (product rule)
   d) $\frac{d}{dx}\frac{f(x)}{g(x)} = \frac{\frac{df(x)}{dx}g(x) - f(x)\frac{dg(x)}{dx}}{(g(x))^2} = \frac{f'(x)g(x) - f(x)g'(x)}{(g(x))^2}$ (quotient rule)
   e) $\frac{d}{dx}f(g(x)) = \frac{df}{dx}(g(x))\frac{dg}{dx}(x) = f'(g(x))g'(x)$ (chain rule)
   f) $\frac{d}{dx}e^x = e^x$
   g) $\frac{d}{dx}\log(x) = \frac{1}{x}$
   h) $\frac{d}{dx}\sin(x) = \cos(x)$
   i) $\frac{d}{dx}\cos(x) = -\sin(x)$
2. **Gradient, Jacobian, Hessian**
   a) Gradient: for a function $f : \mathbb{R}^n \to \mathbb{R}$, the gradient is defined as

$$\nabla_x f(x_1, x_2, \ldots, x_n) = \begin{bmatrix} \frac{\partial}{\partial x_1} f(x) \\ \frac{\partial}{\partial x_2} f(x) \\ \vdots \\ \frac{\partial}{\partial x_n} f(x) \end{bmatrix}$$

b) Jacobian: for a function $f : \mathbb{R}^n \to \mathbb{R}^m$, the Jacobian is defined as

$$\mathbf{J}_f(x_1, x_2, \ldots, x_n) = \begin{bmatrix} \frac{\partial}{\partial x_1} f_1(x) & \ldots & \frac{\partial}{\partial x_n} f_1(x) \\ \frac{\partial}{\partial x_1} f_2(x) & \ldots & \frac{\partial}{\partial x_n} f_2(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial x_1} f_m(x) & \ldots & \frac{\partial}{\partial x_n} f_m(x) \end{bmatrix}$$

c) Hessian: for a function $f : \mathbb{R}^n \to \mathbb{R}$, the Hessian is defined as

$$\mathbf{H}_f(x_1, x_2, \ldots, x_n) = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1 \partial x_1} & \ldots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \ldots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \ldots & \frac{\partial^2 f(x)}{\partial x_n \partial x_n} \end{bmatrix}$$

3. **Differentiation exercises**

a) Compute the derivative of the sigmoid function $\sigma(x) = \dfrac{1}{1 + e^{-x}} = \dfrac{e^x}{1 + e^x}$

b) Compute the gradient of the following function $f : \mathbb{R}^2 \to \mathbb{R}$:

$$f(x, y) = e^{-y} \sin e^x$$

4. **Information theory recap**

a) Given one random variable X with distribution $P : \mathcal{X} \to [0, 1]$, its **entropy** is the average level of "uncertainty" about its possible outcomes and is given by:

$$\begin{aligned} H(X) &= \mathbb{E}_{P_X}[-\log P(X)] && (0.1) \\ &= -\sum_{x \in \mathcal{X}} P(x) \log(P(x)) && (0.2) \end{aligned}$$

b) Given two random variables X and Y on the same set $\mathcal{X}$, with probability distributions $P_X : \mathcal{X} \to [0, 1]$ and $P_Y : \mathcal{X} \to [0, 1]$, the **cross-entropy** between P_X and P_Y measures the average (according to P_X) number of bits needed to describe an event from this set if the coding scheme used is optimized for P_Y. It is given by:

$$\begin{aligned} H(P_X, P_Y) &= \mathbb{E}_{P_X}[-\log P_Y(X)] && (0.3) \\ &= -\sum_{x \in \mathcal{X}} P_X(x) \log(P_Y(x)) && (0.4) \end{aligned}$$

c) Given two random variables X and Y on the same set $\mathcal{X}$, with probability distributions $P_X : \mathcal{X} \to [0,1]$ and $P_Y : \mathcal{X} \to [0,1]$, the **Kullback-Leibler divergence** D_{KL} between P_X and P_Y measures the average (according to P_X) number of bits needed to describe an event from this set if the coding scheme used is optimized for P_Y. It is given by:

$$D_{\text{KL}}(P_X||P_Y) = \sum_{x \in \mathcal{X}} P_X(x) \log\left(\frac{P_X(x)}{P_Y(x)}\right) \tag{0.5}$$

or, if X and Y are continuous random variables:

$$D_{\text{KL}}(P_X||P_Y) = \int_{-\infty}^{+\infty} p_X(x) \log\left(\frac{p_X(x)}{p_Y(x)}\right) \tag{0.6}$$

where p_X and p_Y are probability density functions of their distribution.

We can also express the KL-divergence as follows:

$$D_{\text{KL}}(P_X||P_Y) = H(P_X, P_Y) - H(P_X) \tag{0.7}$$

5. **Information theory exercises**

a) Let the sample space be $\mathcal{X} = \{1, 2, 3\}$. Define two discrete random variables X and Y on $\mathcal{X}$ with probability mass functions

$$P_X(1) = P_X(2) = P_X(3) = \frac{1}{3}, \quad P_Y(1) = \frac{1}{2},\ P_Y(2) = \frac{1}{3},\ P_Y(3) = \frac{1}{6}.$$

Compute the entropy of X and Y.

b) Find an expression for the (continuous) entropy of $\mathbb{U}[a, b]$. (Hint: replace the sum in entropy formula with integral).

c) Compute the cross-entropy and KL divergence between X and Y and between Y and X. What do you see?