AI1072: Machine learning, exercise sheet 4

1 Preparations (on paper)

Compute the following expression symbolically! Remember the derivative of the softmax function $\partial_{x_i} S_j = \delta_{ij} - S_i S_j$ and the properties of the Kronecker delta: $\delta ij = \begin{cases} 1 & i = j \\ 0 & \text{otherwise} \end{cases}$

a) $\frac{\partial x_a}{\partial x_b}$
b) $\frac{\partial X_{ij}}{\partial X_{ab}}$
c) $\frac{\partial}{\partial A_{23}} \sum_{i,j} A_{ij}^2$
d) $\frac{\partial}{\partial A_{23}} \sum_{i,j} \log A_{ij} T_{ij}$
e) $\frac{\partial}{\partial y_1} S_2(\vec{y})$
f) $\frac{\partial}{\partial y_1} S_1(\vec{y})$
g) $\sum_{i,j} A_{ij}^2 \delta_{i1} \delta_{j5}$
h) $\sum_{i,j} A_{ij}^2 \delta_{i1}$

2 Vector-vector chain rule

Given the matrix-matrix function $A(B) = BW$ and $\mathcal{L}(A) = -\sum_{l,m} A_{lm}$: compute the partial derivative $\frac{\partial \mathcal{L}}{\partial B_{nk}}$!

3 Linear regression

Assume a machine learning model given by $Y = f(X, W, \vec{b}) = XW + \vec{b}^T$. The loss is $\mathcal{L}(Y) = N^{-1} \sum_n \sum_k (T_{nk} - Y_{nk})^2$. This is the mean-squared-error loss for linear regression, by the way. Since this is a totally flat model, back-propagation is not required but we can still practise gradient computations. Compute:

a) The derivative $\frac{\partial \mathcal{L}}{\partial Y_{ij}}$
b) The derivative $\frac{\partial \mathcal{L}}{\partial W_{ab}}$
c) The derivative $\frac{\partial \mathcal{L}}{\partial b_a}$

4 A back-propagation step with numbers

Assuming we are faced with a transfer function layer X, $g(x) = x^2$, and we know that $\frac{\partial \mathcal{L}}{\partial A^{(X)}} = \begin{pmatrix} 4 & 1 \\ 0 & 4 \end{pmatrix}$, as well as $A^{(X-1)} = \begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix}$.

Compute all entries of the matrix $\frac{\partial \mathcal{L}}{\partial A_{ij}^{(X-1)}}$!

5 A back-propagation step with numbers, part 2

Repeat the last exercise for an affine layer with $W^{(X)} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$ and $\vec{b}^{(X)} = 0$ (that is, no bias vector)!