EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

MATH. - NATURWISS. FAKULTÄT
FACHBEREICH INFORMATIK
KOGNITIVE SYSTEME · PROF. A. ZELL

Deep Neural Networks

Assignment 3

Assignment due by: 29.05.2018, Discussions on: 05.06.2018

Question 1 Information Theory (1+1+1+2+2=7 Points)

(a) What is the self information in a drawing of the 7 out 49 lottery (no replacement, order does not matter)?

(b) Look up the frequency distribution of letters in English, call it A. Compute the entropy $H(A)$. What does this value mean?

(c) Take the same alphabet you used in (b), but assume a uniform distribution $B \sim \mathrm{Uniform}(\mathrm{size}(A))$. What is the entropy $H(B)$? Is it larger or smaller than in $H(A)$? Why is that?

(d) Compute the cross-entropy between your two distributions $H(A, B)$ and $H(B, A)$. Comment on the results, what are the differences between the two?

(e) Set the probability of one letter in B to zero (keeping the rest of the probabilities uniform) and call the new distribution C. What changes when you compute the entropy $H(C)$ and cross-entropy $H(A, C)$?

Question 2 Over- and Underflow (3+2=5 points)

(a) The $\log \mathrm{softmax}(\boldsymbol{x})$ is a common quantity in neural networks for classification tasks, however it can easily suffer from over- or underflow problems. Using the same approach that was used in the lecture to stabilize the pure $\mathrm{softmax}(\boldsymbol{x})$, find a numerically stable expression for

$$\log\left(\mathrm{softmax}(\boldsymbol{x})_i\right)$$

and explain why it should be stable against underflow and overflow.

(b) Make yourself familiar with the IEEE 754 standard for floating point numbers. Due to the hardware available (typically consumer graphics cards), training of deep neural networks is usually carried out with single-precision (32 bit) floating point numbers, which have a fairly limited range.
Given the biggest and smallest (positive) representable IEEE 754 single precision floating point numbers, what inputs would cause an overflow or underflow in the exponential function?

Question 3 Convergence of gradient descent (2+1+2+1+2=8 points)

Consider the quadratic function $f(x) = \frac{1}{2}x^T Ax + b^T x$, where A is a symmetric positive definite matrix. Let $\{x_k\}_{k\geq 0}$ be the obtained sequence of points, applying the gradient descent algorithm for an arbitrary starting point x_0.

The error is bounded by

$$f(x^{k+1}) - f(x^*) \leq \left(\frac{\lambda_{max}(A) - \lambda_{min}(A)}{\lambda_{max}(A) + \lambda_{min}(A)}\right)^2 (f(x^k) - f(x^*)) \tag{1}$$

where x^* is a global minimum of f (i.e $Ax^* + b = 0$). Let κ be the condition number $\frac{\lambda_{max}}{\lambda_{min}}$ of the matrix A. In this assignment, we will study the convergence of the gradient descent algorithm.

(a) Show that for steepest descent direction $d = Ax + b$,

$$f(x + \epsilon d) = \frac{1}{2}\epsilon^2 d^T Ad + \epsilon d^T d + f(x)$$

and that the optimal learning rate ϵ^* is given by

$$\epsilon^* = \frac{-d^T d}{d^T Ad}$$

(b) Rewrite the factor $\frac{\lambda_{max}(A)-\lambda_{min}(A)}{\lambda_{max}(A)+\lambda_{min}(A)}$ in terms of κ i.e. compute $c(\kappa) = \frac{\lambda_{max}(A)-\lambda_{min}(A)}{\lambda_{max}(A)+\lambda_{min}(A)}$

(c) Using eqn. (1) and (b) show that $\forall k \geq 0$:

$$f(x^{k+1}) - f(x^*) \leq c(\kappa)^{2(k+1)}(f(x^0) - f(x^*))$$

(d) Show that $\forall x \in \mathbb{R}^n$:

$$f(x) - f(x^*) = \frac{1}{2}(x - x^*)^T A(x - x^*)$$

(e) Using the spectral theorem we can derive the fact that $\forall x \in \mathbb{R}^n, x \neq 0$:

$$\lambda_{min}(A) \leq \frac{x^T Ax}{\|x\|_2^2} \leq \lambda_{max}(A) \tag{2}$$

Using this and the results from (c) and (d) show that $\forall k \geq 0$:

$$\left\|x^{k+1} - x^*\right\|_2 \leq c(\kappa)^{(k+1)}\sqrt{\kappa}\left\|x^0 - x^*\right\|_2$$