MDL - TD1

09/01/2024

Exercise 1: Properties of MLPs

Q1: The number of parameters of a model g_θ is the number of coordinates d of $\theta \in \mathbb{R}^d$. How many parameters has an MLP of depth L and widths $d^{(l)}$ for $l \in [\![0, L]\!]$?

Q2: We now assume that all widths are equal $d^{(l)} = d$. How does the number of parameters scale with the width and depth?

Q3: Show that, for two differentable functions f, g, we have $J_{f \circ g}(x) = J_f(g(x))J_g(x)$.

Q4: Show that the Jacobian of an MLP is

$$J_{g_\theta}(x) = W^{(L)}D^{(L-1)}W^{(L-1)}D^{(L-2)} \dots W^{(2)}D^{(1)}W^{(1)}$$

where $D^{(l)} = \mathrm{diag}(\sigma'(g_\theta^{(2l-1)}(x)))$ and $g_\theta^{(2l-1)}(x)$ is the input of the l^{th} activation.

Exercise 2: Algebraic strucutre of ReLU networks

Q5: Show that, for any $L \geq 1$, the identity $g(x) = x$ is a ReLU networks of depth L.

Q6: Show that for $f, g : \mathbb{R}^d \to \mathbb{R}^D$ two MLPs, $f + g$ is still a ReLU network.

Q7: Show that for $f : \mathbb{R}^p \to \mathbb{R}^q$ and $g : \mathbb{R}^q \to \mathbb{R}^r$ two ReLU networks, $g \odot f$ is still a ReLU network.

Q8: Are these properties specific to ReLU activations?

Exercise 3: Concentration inequalities

Q9: *(Markov's inequality)* Let $X \geq 0$ be a positive random variable. Show that, $\forall a > 0$, $\mathbb{P}(X \geq a) \leq \mathbb{E}(X)/a$.

Q10: *(Chernoff's bound)* Let X be a real random variable. Using Markov's inequality, show that, $\forall a \in \mathbb{R}$ and $t > 0$, $\mathbb{P}(X \geq a) \leq \mathbb{E}(e^{tX})e^{-ta}$. Which probability distribution has exponential probability tails?

Q11: *(Hoeffding's inequality)* Let $X_1, \dots, X_n$ be n i.i.d. bounded random variable such that $\forall i_i n[\![1, n]\!]$, $\mathbb{E}(X_i) = 0$ and $|X_i| \leq B$. Using Chernoff's inequality and *Hoeffding's lemma* (i.e. that $\forall t \in \mathbb{R}$, $\mathbb{E}(e^{tX_1}) \leq e^{t^2B^2/2}$), show that, $\forall a > 0$, $\mathbb{P}(\sum_i X_i \geq a) \leq e^{\frac{-a^2}{2nB^2}}$. Which distribution has doubly exponential tails? How is this result related to the CLT?

Q12: Let $\theta \in \mathbb{R}^d$ be a fixed parameter, g_θ a machine learning model, $(X_i, Y_i)_{i \in [\![1,n]\!]}$ i.i.d. random data samples drawn according to a data distribution $\mathcal{D}$, and ℓ a loss function bounded uniformly by a constant $B > 0$. Then, let $\mathcal{L}(\theta) = \mathbb{E}(\ell(g_\theta(X), Y))$ be the risk and $\hat{\mathcal{L}}_n(\theta) = \frac{1}{n}\sum_i \ell(g_\theta(X_i), Y_i)$ the empirical risk of the model g_θ on the data distribution $\mathcal{D}$. Using Hoeffding's inequality, show that, $\forall \delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\left|\hat{\mathcal{L}}_n(\theta) - \mathcal{L}(\theta)\right| \leq B\sqrt{2n \ln \frac{1}{\delta}}$$

How can this result be useful for learning purposes?