Introduction to Machine Learning Course

Short HW5 – Generative models and Deep learning

Submitted individually by Wednesday, 25.01, at 23:59.

You may answer in Hebrew or English and write on a computer or by hand (but be clear).

Please submit a PDF file named like your ID number, e.g., *123456789.pdf*.

Bonus (maximal grade is 100): Writing on a computer (using LyX/LaTeX, Word + Equation tool, etc.) = 3 pts.

1. Let $X_1, \dots, X_{10}$ be random variables sampled i.i.d. from a continuous uniform distribution on the interval $[0, \theta]$ for an unknown $\theta > 0$. We denote the instantiations of these variables by $x_1, \dots, x_{10}$.
   1.1. Write the pdf of an observation $f(X_i = x)$.
   1.2. Derive (פַּתֵּחוּ) the likelihood function $L(x_1, \dots, x_{10}; \theta) = \Pr(x_1, \dots, x_{10}; \theta)$.
   1.3. Differentiate the likelihood function. What $\hat{\theta}$ holds $\frac{\mathrm{d}L(x_1,\dots,x_{10};\theta)}{\mathrm{d}\theta} = 0$?
   1.4. Using 1.2, plot the log-likelihood function $\log L(x_1, \dots, x_{10}; \theta)$ in the interval $(0{,}100)$ and attach the plot to your report (you can use https://www.desmos.com/calculator for quick plotting).
   1.5. Using all the above, conclude what is the MLE estimator $\hat{\theta}_{\mathrm{MLE}}$, i.e., the estimator that maximizes the likelihood. Explain your answer briefly.

2. Consider a trained fully connected neural network with L linear layers.

   Denote the function of the network by $F_\Theta: \mathbb{R}^d \to \mathbb{R}$, where $\Theta = \left(\underbrace{\mathbf{W}^{(1)}}_{\in\mathbb{R}^{d\times p}}, \underbrace{\mathbf{W}^{(2)}}_{\in\mathbb{R}^{p\times p}}, \dots, \underbrace{\mathbf{W}^{(L-1)}}_{\in\mathbb{R}^{p\times p}}, \underbrace{\boldsymbol{w}^{(L)}}_{\in\mathbb{R}^{p}}\right)$ is the set of all weights and no biases (we follow the notations from Tutorial 12).

   As an activation function, we use the ReLU function $\sigma(z) = \max\{0, z\}$.

   The network's output is given by: $F_\Theta(x) = \boldsymbol{w}^{(L)\top} \boldsymbol{h}^{(L-1)}(x)$,

   where we recursively define the hidden layers: $\boldsymbol{h}^{(1)}(x) = \sigma\left(\mathbf{W}^{(1)\top} x\right),\ \boldsymbol{h}^{(\ell)}(x) = \sigma\left(\mathbf{W}^{(\ell)\top} \boldsymbol{h}^{(\ell-1)}(x)\right)$.

   We now scale all weights in Θ by a factor of $\alpha \in \mathbb{R}_{>0}$.

   Notice: The ReLU function is positive-homogeneous in the sense that $\sigma(\alpha \cdot z) = \alpha \cdot \sigma(z)$.

   2.1. Show that the new output function holds $F_{\alpha\cdot\Theta}(x) = c \cdot F_\Theta(x)$ for some scalar $c \in \mathbb{R}$.
   You need to prove your answer briefly. No need to be rigorous (don't use induction).
   In your answer, find an appropriate c value for which this statement holds.

   We wish the model to output a probability. As seen in Lecture 09 (for logistic regression), we can apply the sigmoid function to F_Θ. That is, the model's output will be: $\frac{1}{1+\exp\{-F_{\alpha\cdot\Theta}(x)\}}$.

   2.2. For $\alpha \to 0$, to which probability does the output converge?

   Think (don't include in your answers): For $\alpha \to \infty$, to which probability does the output converge?