Introduction to Machine Learning Course

Short HW5 – Logistic regression and Deep learning

Submitted individually by Wednesday, 05.07, 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.) = 2 pts.

1. Recap (Lecture 09): In logistic regression, we solve a binary classification problem by assuming a probabilistic model, according to which, given an input $\mathbf{x} \in \mathbb{R}^d$, the label is a binomial random variable with probability $\hat{p} = \sigma(\mathbf{w}^\top \mathbf{x})$, where $\sigma(z) = \frac{1}{1+e^{-z}}$ is the sigmoid function.

   More formally, the model assumes that:

$$y \mid \mathbf{x}, \mathbf{w} \sim \text{Binomial}(\hat{p}).$$

   Notice: in this entire question we treat y as $\{0,1\}$ and not as $\{-1,1\}$.

   1.1. Prove that $1 - \sigma(z) = \sigma(-z)$.

   1.2. Prove that $\sigma'(z) = \sigma(z)\big(1 - \sigma(z)\big)$.

   1.3. Derive (פַּתְּחוּ) $\nabla_{\mathbf{w}}\hat{p}$ (remember that it should be of the same size as $\boldsymbol{w}$).
   Hint: if $h(\mathbf{w}) \triangleq f\big(g(\mathbf{w})\big)$ for $f: \mathbb{R} \to \mathbb{R}, g: \mathbb{R}^d \to \mathbb{R}$, then the chain rule dictates that $\nabla h(\mathbf{w}) = f'\big(g(\mathbf{w})\big)\nabla g(\mathbf{w})$.

   We defined the cross-entropy loss $\ell^{CE}(y, \hat{p}) = -y \ln \hat{p} - (1-y)\ln(1-\hat{p})$ which is convex and differentiable w.r.t. $\mathbf{w}$. We also presented an interpretation of logistic regression as solving $\min_{\mathbf{w}\in\mathbb{R}^d} \sum_{i=1}^{m} \ell^{CE}\big(y_i, \sigma(\mathbf{w}^\top \mathbf{x}_i)\big)$, which we denote by (P).

   1.4. Prove that $\nabla_{\mathbf{w}}\ell^{CE}(y, \hat{p}) = (-y + \hat{p})\mathbf{x}$.

   1.5. When $\{\mathbf{x}_i\}_{i=1}^{m}$ are linearly independent, what can be said about the solution of problem (P)? Justify your answer mathematically. There might be several valid answers to this question.
   Hint: Is (P) convex? How can its minimum be attained? When does $-y_i + \hat{p}_i = -y_i + \sigma(\mathbf{w}^\top \mathbf{x}_i)$ become zero?

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 $\Theta = \left(\underbrace{\mathbf{W}^{(1)}}_{\in\mathbb{R}^{d\times p}}, \underbrace{\mathbf{W}^{(2)}}_{\in\mathbb{R}^{p\times p}}, \dots, \underbrace{\mathbf{W}^{(L-1)}}_{\in\mathbb{R}^{p\times p}}, \underbrace{\mathbf{w}^{(L)}}_{\in\mathbb{R}^{p}} \right)$ 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: $F_\Theta(\mathbf{x}) = \mathbf{w}^{(L)^\top} \boldsymbol{h}^{(L-1)}(\mathbf{x})$,

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

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 $F_{\alpha\cdot\Theta}(\mathbf{x}) = c \cdot F_\Theta(\mathbf{x})$ 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: $\frac{1}{1+\exp\{-F_{\alpha\cdot\Theta}(\mathbf{x})\}}$.

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

Think (don't include in your answers): For $\alpha \to 0$, to which probability does the output converge?