Problem 1

Wednesday, 8 January 2020 11:16

Problem 1: You are trying to solve a regression task and you want to choose between two approaches:

1. A simple linear regression model.
2. A feed forward neural network $f_{\boldsymbol{W}}(\boldsymbol{x})$ with L hidden layers, where each hidden layer $l \in \{1, ..., L\}$ has a weight matrix $\boldsymbol{W}_l \in \mathbb{R}^{D\times D}$ and a ReLU activation function. The output layer has a weight matrix $\boldsymbol{W}_{L+1} \in \mathbb{R}^{\cancel{D\times 1}\ 1\times D}$ and no activation function.

In both models, there are no bias terms.

Your dataset $\mathcal{D}$ contains data points with nonnegative features $\boldsymbol{x}_i$ and the target y_i is continuous:

$$\mathcal{D} = \{\boldsymbol{x}_i, y_i\}_{i=1}^N, \qquad \boldsymbol{x}_i \in \mathbb{R}^D_{\geq 0}, \qquad y_i \in \mathbb{R}$$

Let $\boldsymbol{w}^*_{LS} \in \mathbb{R}^D$ be the optimal weights for the linear regression model corresponding to a *global* minimum of the following least squares optimization problem:

$$\boldsymbol{w}^*_{LS} = \arg\min_{\boldsymbol{w}\in\mathbb{R}^D} \mathcal{L}_{LS}(\boldsymbol{w}) = \arg\min_{\boldsymbol{w}\in\mathbb{R}^D} \frac{1}{2}\sum_{i=1}^N (\boldsymbol{w}^T\boldsymbol{x}_i - y_i)^2, \quad \mathcal{L}^*_{LS} = \mathcal{L}_{LS}(w^*_{LS})$$

Let $\boldsymbol{W}^*_{NN} = \{\boldsymbol{W}^*_1, \dots, \boldsymbol{W}^*_{L+1}\}$ be the optimal weights for the neural network corresponding to a *global* minimum of the following optimization problem:

$$\boldsymbol{W}^*_{NN} = \arg\min_{\boldsymbol{W}} \mathcal{L}_{NN}(\boldsymbol{W}) = \arg\min_{\boldsymbol{W}} \frac{1}{2}\sum_{i=1}^N (f_{\boldsymbol{W}}(\boldsymbol{x}_i) - y_i)^2, \quad \mathcal{L}^*_{NN} = \mathcal{L}_{NN}(W^*_{NN})$$

$\star$ The following holds in our case with $x_i \in \mathbb{R}^D_{\geq 0}$:

For arbitrary $w \in \mathbb{R}^D$ there exists a set $W_{NN}(w)$ such that

$$f_{W_{NN}(w)}(x) \overset{\star}{=} w^T x \quad \text{for all } x \in \mathbb{R}^D_{\geq 0}$$

$$\rightarrow \quad f_W(x) = \underbrace{W_{L+1}}_{w^T} \underbrace{\mathrm{ReLU}(W_L\, \mathrm{ReLU}(\dots \mathrm{ReLU}(W_1 x)))}_{x}$$

Define $W_{NN}(w) = (I, \dots, I, w^T)$ $\Rightarrow$ $\star$ since

identity matrix in $\mathbb{R}^{D\times D}$

$\mathrm{ReLU}(Wx) = Wx$ if $x \geq 0$ and $W_{ij} \geq 0$

a) Assume that the optimal $\boldsymbol{W}^*_{NN}$ you obtain are non-negative.
What will the relation $(<, \leq, =, \geq, >)$ between the neural network loss $\mathcal{L}_{NN}(\boldsymbol{W}^*_{NN})$ and the linear regression loss $\mathcal{L}_{LS}(\boldsymbol{w}^*_{LS})$ be? Provide a mathematical argument to justify your answer.

From $\star$ we get that

$$\mathcal{L}^*_{LS} = \mathcal{L}^*(w^*_{LS}) = \mathcal{L}_{NN}(W_{NN}(w^*_{LS})) \geq \mathcal{L}^*_{NN}$$

$$\mathcal{L}^*_{LS} = \mathcal{L}^*_{LS}(w^*_{LS}) = \mathcal{L}_{NN}(W_{NN}(w^*_{LS})) \geq \mathcal{L}^*_{NN}$$

From $W^*_{NN} \geq 0$ we get that

$$f_{W^*_{NN}}(x_i) = W^*_{L+1}\, \mathrm{ReLU}(\underbrace{W^*_L}_{\geq 0}\, \mathrm{ReLU}(\dots \underbrace{\mathrm{ReLU}(\underbrace{W^*_1}_{\geq 0}\, \underbrace{x_i}_{\geq 0})}_{= W^*_1 x_i})))$$

$$= \overbrace{\underbrace{W^*_{L+1}}_{1\times D}\, \underbrace{W^*_L}_{D\times D} \cdot \dots \cdot \underbrace{W^*_1}_{D\times 1}}^{1\times D}\, x_i$$

For non-negative W^*_{NN} function $f_{W^*_{NN}}$ is linear

$$\rightarrow \quad \mathcal{L}^*_{LS} \leq \mathcal{L}_{LS}(W^*_{L+1} \cdot \dots \cdot W^*_1 x_i) = \mathcal{L}_{NN}(W^*_{NN}) = \mathcal{L}^*_{NN}$$

$$\Rightarrow \quad \mathcal{L}^*_{LS} = \mathcal{L}^*_{NN}$$

b) In contrast to (a), now assume that the optimal weights $\boldsymbol{w}^*_{LS}$ you obtain are non-negative. What will the relation ($<, \leq, =, \geq, >$) between the linear regression loss $\mathcal{L}_{LS}(\boldsymbol{w}^*_{LS})$ and the neural network loss $\mathcal{L}_{NN}(\boldsymbol{W}^*_{NN})$ be? Provide a mathematical argument to justify your answer.

See the sample solution on Piazza.