CSC413: Homework 1

Tianyu Du (1003801647)

January 21, 2020

1 Hard-Coding Networks

1.1 Verify Sort

Soln. The first layer performs pairwise comparison to construct indicators $\mathbb{1}\{x_1 \leq x_2\}$, $\mathbb{1}\{x_2 \leq x_3\}$, and $\mathbb{1}\{x_3 \leq x_4\}$. The second layer performs an `all()` operation on indicators from the previous layer.

$$\mathbf{W}^{(1)} = \begin{pmatrix} -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{pmatrix} \tag{1.1}$$

$$\mathbf{b}^{(1)} = \begin{pmatrix} 0 & 0 & 0 \end{pmatrix} \tag{1.2}$$

So that

$$\varphi(\mathbf{h}) = \varphi(\mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)}) = \varphi \begin{pmatrix} x_2 - x_1 \\ x_3 - x_2 \\ x_4 - x_3 \end{pmatrix} = \begin{pmatrix} \mathbb{1}\{x_2 \geq x_1\} \\ \mathbb{1}\{x_3 \geq x_2\} \\ \mathbb{1}\{x_4 \geq x_3\} \end{pmatrix} \tag{1.3}$$

$$\mathbf{w}^{(2)} = \begin{pmatrix} 1 & 1 & 1 \end{pmatrix} \tag{1.4}$$

$$b^{(2)} = -0.5 \tag{1.5}$$

Such that $y = 1$ if and only if all components of $\mathbf{h}$ are ones, i.e., the list is sorted. ■

1.2 Perform Sort

Soln. Algorithm:

1. Let $\ell := ((x_{i1}, x_{i2}, x_{i3}, x_{i4}))_{i=1}^{4P4}$ denote the collection of all permutations of the input;
2. Let $\mathbf{y} := (\texttt{network}(x_{i1}, x_{i2}, x_{i3}, x_{i4}))_{i=1}^{4P4}$ denote variables indicating whether each permutation is sorted or not;
3. Return $\hat{f}(x_1, x_2, x_3, x_4)$ as the ℓ`[y==1]`.

■

1.3 Universal Approximation Theorem

1.3.1

Soln. To avoid over-using of notations, let $\varphi(y) := \mathbb{1}\{y > 0\}$ denote the activation function.

$$\begin{aligned} n &= 2 && (1.6) \\ \mathbf{W}_0 &= (1, -1) && (1.7) \\ \mathbf{b}_0 &= (-a, b) && (1.8) \\ \mathbf{W}_1 &= (1, 1) && (1.9) \\ \mathbf{b}_1 &= -0.5 && (1.10) \end{aligned}$$

Justification:

$$\begin{aligned} \varphi(\mathbf{h}) &= \varphi((x-a, b-x)) && (1.11) \\ &= (\mathbb{1}\{x-a>0\}, \mathbb{1}\{b-x>0\}) && (1.12) \\ &= (\mathbb{1}\{x>a\}, \mathbb{1}\{x<b\}) && (1.13) \\ \varphi(\mathbf{W}_1\varphi(\mathbf{h}) + \mathbf{b}_1) &= \mathbb{1}\{\mathbb{1}\{x>a\} + \mathbb{1}\{x<b\} - 0.5\} && (1.14) \\ &= \mathbb{1}\{x>a\} \wedge \mathbb{1}\{x<b\} && (1.15) \\ &= \mathbb{1}\{a<x<b\} && (1.16) \end{aligned}$$

■

1.3.2

Soln.

$$\begin{aligned} \hat{f}_1(x) &= \hat{f}_0(x) + g(h_1, a_1, b_1, x) && (1.17) \\ &= 0 + g\left(\right) && (1.18) \end{aligned}$$

■

1.3.3

Soln. ■

1.3.4

Soln. Not required. ■

2 Backprop

2.1 Computational Graph

2.1.1

Soln. TODO: *Add graph* ■

2.1.2

Soln.

$$\begin{aligned}
\overline{\mathbf{x}} &= \overline{\mathbf{z}}\frac{\partial \mathbf{z}}{\partial \mathbf{x}} && (2.1)\\
&= \overline{\mathbf{z}}\mathbf{W}^{(1)} && (2.2)\\
&= \overline{\mathbf{h}}\frac{\partial \mathbf{h}}{\partial \mathbf{z}}\mathbf{W}^{(1)} && (2.3)\\
&= \overline{\mathbf{h}}\mathbb{1}\{\mathbf{z} \geq 0\}\mathbf{W}^{(1)} && (2.4)\\
&= \left(\overline{\mathcal{R}}\frac{\partial \mathcal{R}}{\partial \mathbf{h}} + \overline{\mathbf{y}}\frac{\partial \mathbf{y}}{\partial \mathbf{h}}\right)\mathbb{1}\{\mathbf{z} \geq 0\}\mathbf{W}^{(1)} && (2.5)\\
&= \left(\overline{\mathcal{R}}\mathbf{r}^T + \overline{\mathbf{y}}\mathbf{W}^{(2)}\right)\mathbb{1}\{\mathbf{z} \geq 0\}\mathbf{W}^{(1)} && (2.6)\\
&= \left(\mathbf{r}^T + \overline{\mathbf{y}'}\frac{\partial \mathbf{y}'}{\partial \mathbf{y}}\mathbf{W}^{(2)}\right)\mathbb{1}\{\mathbf{z} \geq 0\}\mathbf{W}^{(1)} && (2.7)\\
&= \left(\mathbf{r}^T + \overline{\mathbf{y}'}\texttt{softmax}'(\mathbf{y})\mathbf{W}^{(2)}\right)\mathbb{1}\{\mathbf{z} \geq 0\}\mathbf{W}^{(1)} && (2.8)\\
&= \left(\mathbf{r}^T + \overline{\mathcal{S}}\frac{\partial \mathcal{S}}{\partial \mathbf{y}'}\texttt{softmax}'(\mathbf{y})\mathbf{W}^{(2)}\right)\mathbb{1}\{\mathbf{z} \geq 0\}\mathbf{W}^{(1)} && (2.9)\\
&= \left(\mathbf{r}^T + \mathbf{e}_k\ \texttt{softmax}'(\mathbf{y})\mathbf{W}^{(2)}\right)\mathbb{1}\{\mathbf{z} \geq 0\}\mathbf{W}^{(1)} && (2.10)
\end{aligned}$$

where $\mathbf{e}_k$ denotes the one-hot vector in $\mathbb{R}^M$ in which the k^{th} element is one. ■

2.2 Vector-Jacobean Product (VJPs)

2.2.1

2.2.2

2.2.3

3 Linear Regression

3.1 Driving the Gradient

Soln.

$$\begin{aligned}
\frac{d}{d\hat{\mathbf{w}}}\frac{1}{n}(X\hat{\mathbf{w}} - \mathbf{t})^2 &= \frac{d}{d\hat{\mathbf{w}}}\frac{1}{n}||X\hat{\mathbf{w}} - \mathbf{t}||_2^2 && (3.1)\\
&= \frac{2}{n}(X\hat{\mathbf{w}} - \mathbf{t})^T X && (3.2)
\end{aligned}$$

■

3.2 Under-parameterized Model

3.2.1

Soln. Assume $d < n$ so that $X^T X$ is invertible. The gradient descent algorithm converges when the gradient equals zero:

$$\frac{2}{n}(X\hat{\mathbf{w}} - \mathbf{t})^T X = 0 \tag{3.3}$$
$$\implies (X\hat{\mathbf{w}} - \mathbf{t})^T X = 0 \tag{3.4}$$
$$\implies X^T(X\hat{\mathbf{w}} - \mathbf{t}) = 0^T \tag{3.5}$$
$$\implies X^T X\hat{\mathbf{w}} - X^T\mathbf{t} = 0^T \tag{3.6}$$
$$\implies X^T X\hat{\mathbf{w}} = X^T\mathbf{t} \tag{3.7}$$
$$\implies \hat{\mathbf{w}} = (X^T X)^{-1} X^T\mathbf{t} \tag{3.8}$$

■

3.2.2

Soln. Let $\mathbf{x} \in \mathbb{R}^d$, note that $(X^T X)^{-1}$ is symmetric. Assuming target $\mathbf{t}$ is generated by a linear process, then $\mathbf{t} = X\mathbf{w}^*$. Immediately, $\mathbf{t}^T = \mathbf{w}^{*T} X^T$.

$$(\mathbf{w}^{*T}\mathbf{x} - \hat{\mathbf{w}}^T\mathbf{x})^2 = (\mathbf{w}^{*T}\mathbf{x} - [(X^T X)^{-1} X^T\mathbf{t}]^T\mathbf{x})^2 \tag{3.9}$$
$$= (\mathbf{w}^{*T}\mathbf{x} - \mathbf{t}^T X(X^T X)^{-1}\mathbf{x})^2 \tag{3.10}$$
$$= (\mathbf{w}^{*T}\mathbf{x} - \mathbf{w}^{*T} X^T X(X^T X)^{-1}\mathbf{x})^2 \tag{3.11}$$
$$= (\mathbf{w}^{*T}\mathbf{x} - \mathbf{w}^{*T}\mathbf{x})^2 \tag{3.12}$$
$$= 0 \tag{3.13}$$

■

3.3 Over-parameterized Model: 2D Example

3.3.1

Soln. To minimize the empirical risk minimizer,

$$\min_{w_1, w_2} (w_1 x_1 + w_2 x_2 - t_1)^2 \tag{3.14}$$
$$\text{equivalently,} \quad \min_{w_1, w_2} (2w_1 + w_2 - 2)^2 \tag{3.15}$$

Any pair of (w_1, w_2) satisfying

$$2w_1 + w_2 - 2 = 0 \quad (\dagger) \tag{3.16}$$

attains the minimum level of empirical risk (zero). Equivalently, any $\hat{\mathbf{w}}$ on the line

$$\hat{\mathbf{w}} = \begin{pmatrix} 0 \\ 2 \end{pmatrix} + t \begin{pmatrix} 1 \\ -2 \end{pmatrix} \quad \text{for } t \in \mathbb{R} \tag{3.17}$$

satisfies (†). Therefore, there are infinitely many empirical risk minimizers. ■

3.3.2

Soln. ■