ComS 474
Final Exam

Sean Gordon

Nov 26, 2020

1 Regular Problems

1) $\begin{pmatrix} 1 & 1/2 & 1/2 \\ 1/3 & 1/2 & 1 \end{pmatrix} * \begin{pmatrix} 0.5 & 1 & 6 \\ 3 & -4 & 2 \end{pmatrix} = \begin{pmatrix} 0.5 & 0.5 & 3 \\ 1 & -2 & 2 \end{pmatrix}$

---

2) (a) $\begin{pmatrix} 1 & 1/2 & 1/2 \\ 1/3 & 1/2 & 1 \end{pmatrix} * \begin{pmatrix} 0.5 & 3 \\ 1 & -4 \\ 6 & 2 \end{pmatrix} = \begin{pmatrix} 4 & 2 \\ 6.667 & 1 \end{pmatrix}$

(b) $\begin{pmatrix} 1 & 1/3 \\ 1/2 & 1/2 \\ 1/2 & 1 \end{pmatrix} * \begin{pmatrix} 0.5 & 1 & 6 \\ 3 & -4 & 2 \end{pmatrix} = \begin{pmatrix} 4 & 6.667 \\ 2 & 1 \end{pmatrix}$

---

3) $\left(\begin{pmatrix} 1 & 1/3 \\ 1/2 & 1/2 \\ 1/2 & 1 \end{pmatrix} * \begin{pmatrix} 0.5 & 1 & 6 \\ 3 & -4 & 2 \end{pmatrix} \right) + 1 \Rightarrow \begin{pmatrix} 4 & 6.667 \\ 2 & 1 \end{pmatrix} + 1 = \begin{pmatrix} 5 & 7.667 \\ 3 & 2 \end{pmatrix}$

---

4) $\hat{y} = \phi(w^T x) = ((1/2) * 2)^2 + ((1/3) * 3)^2 + ((1/4) * 4)^2 + ((1/5) * 5)^2 = 4$

---

5) As $\hat{y} = (w^T x)^2$...

(a) $\dfrac{\partial E}{\partial \hat{y}} = \dfrac{\partial (y+\hat{y})}{\partial \hat{y}} = \dfrac{\partial y}{\partial \hat{y}} + \dfrac{\partial \hat{y}}{\partial \hat{y}} = 0+1 = 1$

(b) $\dfrac{\partial \hat{y}}{\partial w^T x} = \dfrac{\partial (w^T x)^2}{\partial w^T x} = \dfrac{\partial (u)^2}{\partial u} = 2u = 2w^T x =$
$2((1/2) * 2) + 2((1/3) * 3) + 2((1/4) * 4) + 2((1/5) * 5) = 8$

(c) $\dfrac{\partial w^T x}{\partial x_1} = \dfrac{\partial (w_0 x_0 + w_1 x_1 + w_2 x_2 + w_3 x_3)}{\partial x_1} = w_1$

(d) $\dfrac{\partial E}{\partial x_1} = \dfrac{\partial E}{\partial \hat{y}} \dfrac{\partial \hat{y}}{\partial w^T x} \dfrac{\partial w^T x}{\partial x_1} = 1 * 8 * w_1 = 1 * 8 * 3 = 24$

---

6)

(a) $\dfrac{\partial E}{\partial x} = \begin{pmatrix} \frac{\partial E}{\partial x_0} \\ \frac{\partial E}{\partial x_1} \\ \frac{\partial E}{\partial x_2} \\ \frac{\partial E}{\partial x_3} \end{pmatrix} = \begin{pmatrix} w_0 \\ w_1 \\ w_2 \\ w_3 \end{pmatrix} = w = \begin{pmatrix} 2 \\ 3 \\ 4 \\ 5 \end{pmatrix}$

(b) $\dfrac{\partial E}{\partial w} = \begin{pmatrix} \frac{\partial E}{\partial w_0} \\ \frac{\partial E}{\partial w_1} \\ \frac{\partial E}{\partial w_2} \\ \frac{\partial E}{\partial w_3} \end{pmatrix} = \begin{pmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \end{pmatrix} = x = \begin{pmatrix} 1/2 \\ 1/3 \\ 1/4 \\ 1/5 \end{pmatrix}$

---

2 Bonus Problems

7) $$x^1 = \phi \left[\begin{pmatrix} 1 & -1 & 0.1 \\ 1 & -1 & 0.1 \\ 1 & -1 & 0.1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \right] = max \left(\begin{pmatrix} 0.1 \\ 0.1 \\ 0.1 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \right) = \begin{pmatrix} 0.1 \\ 0.1 \\ 0.1 \end{pmatrix}$$

$$x^2 = \phi \left[\begin{pmatrix} 0.5 & 0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.5 & 0.5 \end{pmatrix} \begin{pmatrix} 1 \\ 0.1 \\ 0.1 \\ 0.1 \end{pmatrix} \right] = max \left(\begin{pmatrix} 0.65 \\ 0.65 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \end{pmatrix} \right) = \begin{pmatrix} 0.65 \\ 0.65 \end{pmatrix}$$

$$x^3 = \phi \left[\begin{pmatrix} 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 \end{pmatrix} \begin{pmatrix} 1 \\ 0.65 \\ 0.65 \end{pmatrix} \right] = max \left(\begin{pmatrix} 0.575 \\ 0.575 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \end{pmatrix} \right) = \begin{pmatrix} 0.575 \\ 0.575 \end{pmatrix}$$

$$\phi(x) = \begin{cases} 0 & x \leq 0 \\ x & x > 0 \end{cases} \qquad \psi(x) = \phi'(x) = \begin{cases} 0 & x \leq 0 \\ 1 & x > 0 \end{cases}$$

$$\delta^{(l)} = \frac{\partial E}{\partial \mathbb{W}^{(l-1)} x^{(l-1)}} = \frac{\partial (\hat{y} - y)^2}{\partial \mathbb{W}^{(l-1)} x^{(l-1)}} = \frac{\partial (\phi(\mathbb{W}^{(l-1)} x^{(l-1)}) - y)^2}{\partial \mathbb{W}^{(l-1)} x^{(l-1)}} = 2(\phi(\mathbb{W}^{(l-1)} x^{(l-1)}) - y)$$

$$\delta^{(l-1)} = \begin{cases} \psi(x^{(l-1)}) \circ (\mathbb{W}^{(l-1)} \delta^{(l)}_{[1..]}) & \text{if } l \text{ is not output layer} \\ \psi(x^{(l-1)}) \circ (\mathbb{W}^{(l-1)} \delta^{(l)}) & \text{otherwise} \end{cases}$$

---

8) $$\delta^{(3)} = 2 \left(\begin{pmatrix} x_1^{(3)} \\ x_2^{(3)} \end{pmatrix} - \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \right) = \begin{pmatrix} (2x_1^{(3)} - 2y_1) \\ (2x_2^{(3)} - 2y_2) \end{pmatrix} = \begin{pmatrix} (1.15 - 2y_1) \\ (1.15 - 2y_2) \end{pmatrix}$$

$$\delta^{(2)} = \begin{pmatrix} \psi(x_0^{(2)}) \\ \psi(x_1^{(2)}) \\ \psi(x_2^{(2)}) \end{pmatrix} \circ \left(\begin{pmatrix} w_{1,1}^{(2)} & w_{1,2}^{(2)} \\ w_{2,1}^{(2)} & w_{2,2}^{(2)} \\ w_{3,1}^{(2)} & w_{3,2}^{(2)} \end{pmatrix} * \delta^{(3)} \right) = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \circ \left(\begin{pmatrix} 0.25 & 0.25 \\ 0.25 & 0.25 \\ 0.25 & 0.25 \end{pmatrix} * \delta^{(3)} \right)$$

$$\delta^{(1)} = \begin{pmatrix} \psi(x_0^{(1)}) \\ \psi(x_1^{(1)}) \\ \psi(x_2^{(1)}) \\ \psi(x_3^{(1)}) \end{pmatrix} \circ \left(\begin{pmatrix} w_{1,1}^{(1)} & w_{1,2}^{(1)} \\ w_{2,1}^{(1)} & w_{2,2}^{(1)} \\ w_{3,1}^{(1)} & w_{3,2}^{(1)} \\ w_{4,1}^{(1)} & w_{4,2}^{(1)} \end{pmatrix} * \delta^{(2)}_{[1..]} \right) = \begin{pmatrix} 0 \\ 1 \\ 1 \\ 1 \end{pmatrix} \circ \left(\begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \\ 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix} * \delta^{(2)}_{[1..]} \right)$$

$$\delta^{(0)} = \begin{pmatrix}\psi(x_0^{(0)})\\ \psi(x_1^{(0)})\\ \psi(x_2^{(0)})\end{pmatrix} \circ \left(\begin{pmatrix}w_{1,1}^{(0)} & w_{1,2}^{(0)} & w_{1,3}^{(0)}\\ w_{2,1}^{(0)} & w_{2,2}^{(0)} & w_{2,3}^{(0)}\\ w_{3,1}^{(0)} & w_{3,2}^{(0)} & w_{3,3}^{(0)}\end{pmatrix} * \delta_{[1..]}^{(1)}\right) = \begin{pmatrix}1\\1\\1\end{pmatrix} \circ \left(\begin{pmatrix}1 & 1 & 1\\ -1 & -1 & -1\\ 0.1 & 0.1 & 0.1\end{pmatrix} * \delta_{[1..]}^{(1)}\right)$$

---

9) No, a supervised model WITHOUT regularization will usually not perform as well on test data than one WITH regularization.

Supervised models are prone to overfitting, aligning too closely with the training data. This causes the model to be able to predict the training data very well, but not perform well on any other dataset. This can be resolved by regularizing the training data to reduce variance, ensuring the model does not overfit.

---

10) No, as if the dataset is heavily skewed, with - for example - 10,000 samples in class 0 and 10 samples in class 1, a model could simply predict class 0 every time and come out with 99% accuracy.

---

11) K-means pseudocode:

```
K = number of clusters
M = number of data points

for each c in K, place c randomly;

for each p in M:
    c = nearest centroid to p;
    Assign p to c;

for each c in K:
    c = mean of all points p assigned to c
```

Because of the 3 adjacent for loops, this algorithm's time complexity is

O(kmk) = O(2km) ≈ O(km)