

$$\begin{aligned} \frac{\partial l}{\partial w_{(1,2)}^{(1)}} &= \frac{\partial l}{\partial o} \cdot \frac{\partial o}{\partial h^{(2)}} \cdot \frac{\partial h^{(2)}}{\partial w_{(1,2)}^{(1)}} \\ &= \frac{2}{m} \sum_{i=1}^m (o^{(i)} - y^{(i)}) o^{(i)} (1 - o^{(i)}) w_{(2,1)}^{(2)} h^{(2)} (1 - h^{(2)}) x_1^{(i)} \end{aligned}$$

$$\Rightarrow w_{(1,2)}^{(1)+} := w_{(1,2)}^{(1)} - \alpha \cdot \frac{2}{m} \cdot w_{(2,1)}^{(2)} \sum_{i=1}^m (o^{(i)} - y^{(i)}) o^{(i)} (1 - o^{(i)}) h^{(2)} (1 - h^{(2)}) x_1^{(i)}$$

$$\text{where } h^{(2)} = \sigma \left( w_{(1,2)}^{(1)} + x_1^{(i)} w_{(1,2)}^{(1)} + x_2^{(i)} w_{(1,2)}^{(1)} \right)$$

(b) Yes, understand the three hidden layers as three separate classifiers that are lines which ~~complete~~ generate the bounding triangle. We essentially calibrate slopes / intercepts.

(c) No. The resultant model is no better than a linear classifier.

2. (a) We know that

$$D_{KL}(P||Q) = \sum_{n \in X} P(n) \log \frac{P(n)}{Q(n)}$$

$$= E \left[ \log \frac{P(n)}{Q(n)} \right] = - E \left[ -\log \frac{P(n)}{Q(n)} \right]$$

$$\geq -\log E \left[ \frac{Q(n)}{P(n)} \right] = -\log \sum_{n \in X} P(n) \cdot \frac{Q(n)}{P(n)}$$

$$= -\log \sum Q(n) = 0$$

$$\text{If } P=Q, D_{KL}(P||Q) = \sum_n P(n) \log \frac{P(n)}{Q(n)} = \log 1 \sum_n P(n) = 0$$

$$\text{If } D_{KL}(P||Q) = 0, \frac{Q(n)}{P(n)} = 1 \Rightarrow Q(n) = P(n)$$

$$\begin{aligned} (b) D_{KL}(P(X,Y)||Q(X,Y)) &= \sum_n \sum_y P(n,y) \log \frac{P(n,y)}{Q(n,y)} \\ &= \sum_n \sum_y P(n) P(y|n) \log \frac{P(n) P(y|n)}{Q(n) Q(y|n)} \end{aligned}$$



$$\begin{aligned}
 &= \sum_n \sum_y P(n) P(y|n) \left( \log \frac{P(n)}{Q(n)} + \log \frac{P(y|n)}{Q(y|n)} \right) \\
 &= \sum_n \sum_y P(n) P(y|n) \log \frac{P(n)}{Q(n)} + \sum_n \sum_y P(n) P(y|n) \log \frac{P(y|n)}{Q(y|n)} \\
 &= \sum_n P(n) \log \frac{P(n)}{Q(n)} \underbrace{\sum_y P(y|n)}_1 + \sum_n P(n) \sum_y P(y|n) \log \frac{P(y|n)}{Q(y|n)} \\
 &= \sum_n P(n) \log \frac{P(n)}{Q(n)} + \sum_n P(n) \sum_y P(y|n) \log \frac{P(y|n)}{Q(y|n)} \\
 &= D_{KL}(P(X) \| Q(X)) + D_{KL}(P(Y|X) \| Q(Y|X))
 \end{aligned}$$

$$\begin{aligned}
 (c) \argmin_{\hat{P}} D_{KL}(\hat{P} \| P_0) &= \argmin_{\hat{P}} \sum_n \hat{P}(n) \log \frac{\hat{P}(n)}{P_0(n)} \\
 &= \argmin_{\hat{P}} \sum_n \hat{P}(n) \log \hat{P}(n) - \argmin_{\hat{P}} \sum_n \hat{P}(n) \log P_0(n) \\
 &= \argmax_{\hat{P}} \sum_n \hat{P}(n) \log P_0(n) \\
 &= \argmax_{\hat{P}} \sum_n \frac{1}{m} \sum_{i=1}^m 1_{\{n^{(i)}=n\}} \log P_0(n) \\
 &= \argmax_{\hat{P}} \sum_{i=1}^m \sum_n 1_{\{n^{(i)}=n\}} \log P_0(n^{(i)}) \\
 &= \argmax_{\hat{P}} \sum_{i=1}^m \log P_0(n^{(i)})
 \end{aligned}$$

3.4 Soared over my head lol.

5: (b) Entropy, 24 bits were needed  
 Now, 4 are reqd.

Hence, compression has been done by a factor of 6.