LISTA III - AAP

THIAGO MALTA CUTINHO - 2014123335

(1) Primeira imagem → HIGH BIAS. A complexidade do modelo pode não ser suficiente, ou os pesos não foram ajustados de maneira correta. Podemos aumentar a complexidade do modelo, treinar por mais tempo ou mudar o algoritmo de otimização.

Terceira imagem → HIGH VARIANCE. O modelo pode estar sobreajustado, podemos introduzir uma regularização L1, L2, dropout ou early-stopping.

(2) λ é a variável que regula o trade-off entre minimizar o erro no conjunto de treinamento e diminuir a norma do vetor de pesos. Um lambda pequeno implica em um maior ajuste da reta aos dados. Um lambda maior implica em menor ajuste da reta aos dados.

(3) a) Sem Regularização: $W^{(n+1)} = W^{(n)} - \alpha \nabla J(\theta)$; $\alpha = 0,1$

$$W^{[1]} = \begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix} - \begin{bmatrix} -1 & 0,5 \\ 0,1 & 0,25 \end{bmatrix} = \begin{bmatrix} 2 & 1,5 \\ 1,9 & 2,75 \end{bmatrix} \quad ; \quad b^{[1]} = \begin{bmatrix} -2 \\ 4 \end{bmatrix} - \begin{bmatrix} 0,3 \\ -0,3 \end{bmatrix} = \begin{bmatrix} -2,3 \\ 4,3 \end{bmatrix}$$

$$W^{[2]} = \begin{bmatrix} 1 \\ 5 \end{bmatrix} - \begin{bmatrix} -0,1 \\ 0,2 \end{bmatrix} = \begin{bmatrix} 1,1 \\ 4,8 \end{bmatrix} \quad ; \quad b^{[2]} = [-4] - [0,2] = [-3,8]$$

Com Regularização: $W^{(n+1)} = W^{(n)} - \nabla J(\theta) - \underbrace{2\alpha\lambda}_{0,1}\theta$

$$W^{[1]} = \begin{bmatrix} 2 & 1,5 \\ 1,9 & 2,75 \end{bmatrix} - \begin{bmatrix} 0,1 & 0,2 \\ 0,2 & 0,3 \end{bmatrix} = \begin{bmatrix} 1,9 & 1,3 \\ 1,7 & 2,45 \end{bmatrix}$$

$$b^{[1]} = \begin{bmatrix} -2,3 \\ 4,3 \end{bmatrix} - \begin{bmatrix} -0,2 \\ 0,4 \end{bmatrix} = \begin{bmatrix} -2,1 \\ 3,9 \end{bmatrix} \quad \Big| \quad W^{[2]} = \begin{bmatrix} 1,1 \\ 4,8 \end{bmatrix} - \begin{bmatrix} 0,1 \\ 0,5 \end{bmatrix} = \begin{bmatrix} 1,0 \\ 4,3 \end{bmatrix}$$

$$b^{[2]} = [-3,8] - [-0,4] = [-3,4]$$

---

b) O valor de cada peso, em absoluto, é menor ao utilizar a regularização.

c) A rede sem regularização irá ajustar os pesos aos dados dando maior importância às features que importam no treino, isso pode causar sobre-ajuste. A rede regularizada terá a norma do vetor de pesos menores, diminuindo a complexidade e o ajuste do modelo aos dados, possivelmente generalizando melhor.

---

(4) Ao normalizar as entradas, a magnitude dos dados será a mesma, o que evita discrepância em magnitudes nos pesos da rede. Ao otimizar a função de custo, que é função dos pesos da rede, a superfície tenderá a ser mais "circular" e menos achatada. Isso facilita que os algoritmos de otimização convergam devido ao vetor gradiente que é ortogonal aos níveis da função de custo.

---

(5) a) O efeito de regularização aumenta, pois teremos menos neurônios no treinamento, resultando em menor complexidade final.

b) A tendência é que o erro de treino aumente, dado que a rede é menos complexa e se ajusta menos ao conjunto de treino.

c) Não, os neurônios são modificados apenas no treino. Em teste todos os neurônios são utilizados.

(6) a) Em torno de 50, pois a curva passou por um "cotovelo".

b) Em torno de 40.

c) O conjunto de validação permite verificar se o modelo realmente está generalizando ou apenas aprendendo os dados de treinamento. Verificar se o erro de validação continua caindo, se a curva começa a subir é melhor parar de treinar para evitar o overfitting.

---

(7) Se o batch for de tamanho 1, não podemos aproveitar as técnicas de vetorização; e se o batch for m, teremos que percorrer todo o conjunto de dados para atualizar os pesos da rede. Batchs intermediários são ideais, porque permitem atualizar os pesos de forma eficiente e com melhor qualidade.

---

(8) a) Épocas em batch e mini-batch são diferentes.

b) Uma iteração em mini-batch é mais rápida, pois usa menos dados.

---

(9) Algoritmo em mini-batch, pois como o batch é menor, os dados podem não representar constantemente a função de perda da rede. Assim, hora sim, hora não, o erro sobe e desce. Mas a tendência final é de decrescimento.

(10) $\theta_1 = 10$; $\theta_2 = 10$; $\beta = 0,5$; $V_0 = 0$; $V_t = \beta V_{t-1} + (1-\beta)\theta_t$

$V_t^{\text{corrigido}} = \frac{V_t}{1-\beta^t}$

$V_1 = \beta V_0 + (1-\beta)\theta_1$

$V_1 = 0 + 0,5 \cdot 10 = 5$

$V_2 = \beta V_1 + (1-\beta)\theta_2$

$V_2 = 2,5 + 5 = \boxed{7,5}$

$V_2^{\text{corrigido}} = \frac{V_2}{1-(0,5)^2} = \frac{7,5}{0,75} = \boxed{10}$