23K-0019

ASSIGNMENT 2

Date

M	T	W	T	F	S	S

Question 1:

(a) As M increases the model becomes more flexible and can fit more complex relationships. For high M, the polynomial can follow noise in the training data and produce large oscillations, this is overfitting. Overfitting occurs as high-degree polynomials have many parameters and can fit random fluctuations in the training set instead of the underlying pattern.

This problem can be addressed by including a penalty term to smoothen the curve, known as regularization. Increasing training data and using cross-validation also addresses the overfitting problem.

(b) Split the data into training and validation.

Train polynomial models for different values of M.

Evaluate performance on validation set.

Select M that gives the best validation metric.

Signature

Date____

M	T	W	T	F	S	S

(c) $$\hat{y}(x_i, w) = \sum_{j=0}^{M} w_j x_i^j$$

MSE:

$$J(w) = \frac{1}{2N} \sum_{i=1}^{N} (x_i w - y_i)^2 \Rightarrow \underbrace{\frac{1}{2N} \| xw - y \|^2}_{\text{without regularization.}}$$

→ gradient descent can be used to reduce the cost function:

$$\frac{\partial J}{\partial w} = \frac{1}{N} x^T (xw - y) + \lambda w \quad , \quad w_j = w_i - \alpha \frac{\partial J}{\partial w_j}$$

Question 2:

(a) Stochastic gradient descent is preferred over batch gradient descent when the dataset is large because stochastic is faster that batch gradient descent, it also improves generalization.

(b) The learning rate (α) scales how far we move along the gradient in each update. A large value of α may lead to divergence of the algorithm. A small α value leads to extremely slow convergence. Moderate α with decay; large steps for fast progress, smaller steps later for convergence.

Date ____

M	T	W	T	F	S	S

(c) The randomness in ~~SGD~~ stochastic gradient descent helps avoid getting stuck at saddle points and acts as regularization, however it needs tuning for stable convergence.

Q3:

(a) $P(y=m|x) = h_{\theta_m}(x) = \dfrac{e^{\theta_m^T x}}{\sum_{k=0}^{m-1} e^{\theta_k^T x}}$

$$\Rightarrow L(\theta) = -\sum_{i=1}^{N}\sum_{k=1}^{K} y_{(i,k)} \log\left(P_{(i,k)}\right)$$

(b) Gradient:

$$\frac{\partial L}{\partial \theta_k} = \sum_{i=1}^{n}\left(h_\theta(x_i) - y_{(i,k)}\right)x_i$$

$$\Rightarrow \theta_k = \theta_k - \frac{\alpha\,\partial L}{\partial \theta_k}$$

(c) Yes, cross entropy is exactly the negative log-likelihood used in logistic regression.

$$L_i = -\left[y_i \log p_i + (1-y_i)\log(1-p_i)\right]$$

(d) Cross entropy handles multiple classes by comparing the true one-hot distribution to the predicted probability distribution, it uses the softmax output and checks for loss.

$$-\sum_{k=1}^{K} y_{(i,k)} \log\left(h_\theta(x_i)_k\right)$$

 RC Signature ____

Date
M	T	W	T	F	S	S

Question 4:

(a) $\beta_0 = -3, \quad -2, \quad -1$

$\beta_1 = 0.4, \quad 0.6, \quad 0.9$

$\beta_2 = 0.05, \quad 0.07, \quad 0.1$

$\Rightarrow$ Hours of study = 5

$\Rightarrow$ Previous score = 70.

$\Rightarrow z_0 = -3 + 0.4(5) + 0.05(70) = 2.5$

$\Rightarrow z_1 = -2 + 0.6(5) + 0.07(70) = 5.9$

$\Rightarrow z_3 = -1 + 0.9(5) + 0.1(70) = 10.5$

$\Rightarrow e^{z_0} = e^{2.5} = 12.18$

$\Rightarrow e^{z_1} = e^{5.9} = 365.04.$

$\Rightarrow e^{z_2} = e^{10.5} = 36316.503$

$\Rightarrow \sum_{k=0}^{m-1} e^{\theta_k T_x} = 36692.72$

$\Rightarrow P_0 = \frac{12.18}{36692.72} = 0.00033$

$P_1 = \frac{365.037}{36692.72} = 0.00995$

$P_2 = \frac{36315.503}{36692.727} = 0.9897$

(b)

$t_0 = t_1 = 0$

$t_2 = 1$

$\Rightarrow L = -\log(0.9897) = 0.0104.$

Page No.

Signature