Machine Learning

Assignment - 1

Sravya Mendem
700765926

1) Function Approximation by Hand:

$(x, y) = \{(1,1), (2,2), (3,2), (4,5)\}$

Model: $\hat{y} = \theta_1 x + \theta_2$

$$J(\theta) = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

$\left[MSE = \frac{1}{n} \Sigma r^2 \right]$

→ 1. $\theta = (1, 0)$

$\hat{y} = 1 \cdot x + 0 = 1x$

x	$\hat{y}$	$r = y - \hat{y}$	r^2
1	1	$1 - 1 = 0$	0
2	2	$2 - 2 = 0$	0
3	3	$2 - 3 = -1$	1
4	4	$5 - 4 = 1$	1
			$\Sigma = 2$

$$MSE = \frac{1}{4}(2) = \underline{0.5}$$

→ 2. $\theta = (0.5, 1)$

$\hat{y} = 0.5x + 1$

x	$\hat{y}$	$r = y - \hat{y}$	r^2
1	$0.5(1) + 1 = 1.5$	$1 - 1.5 = -0.5$	$(-0.5)^2 = 0.25$
2	$0.5(2) + 1 = 2$	$2 - 2 = 0$	0
3	$0.5(3) + 1 = 2.5$	$2 - 2.5 = -0.5$	$(-0.5)^2 = 0.25$
4	$0.5(4) + 1 = 3$	$5 - 3 = 2$	$(2)^2 = 4$
			$\Sigma = 4.5$

$$MSE = \frac{1}{4}(4.5) = \underline{1.125}$$

→ 3. Lower MSE is best fit

* $\theta = (1,0)$ has **best** fit, because $0.5 < 1.125$

2) Random Guessing Practice:

$$J(\theta_1, \theta_2) = 8(\theta_1 - 0.3)^2 + 4(\theta_2 - 0.7)^2$$

→ 1. $J(0.1, 0.2)$ and $J(0.5, 0.9)$

$$J(0.1, 0.2) = 8(0.1 - 0.3)^2 + 4(0.2 - 0.7)^2$$
$$= 8(-0.2)^2 + 4(-0.5)^2$$
$$= 0.32 + 1$$
$$= \underline{1.32}$$

$$J(0.5, 0.9) = 8(0.5 - 0.3)^2 + 4(0.9 - 0.7)^2$$
$$= 8(0.2)^2 + 4(0.2)^2$$
$$= 0.32 + 0.16$$
$$= \underline{0.48}$$

→ 2. $(0.5, 0.9)$ is ~~the~~ closer to the minimum, since $0.48 < 1.32$

→ 3. Random guessing is inefficient because it does not use information about gradient (lacks directionality) and require many iterations to find a good solution whereas optimization algorithms use gradient to converge faster and more reliably.

3) First Gradient Descent Iteration:

Dataset: $(1,3), (2,4), (3,6), (4,5)$

$$\frac{\partial J}{\partial \theta_1} = -\frac{2}{N} \sum_{i=1}^{N} x^{(i)} \left(y^{(i)} - \hat{y}^{(i)}\right), \quad \frac{\partial J}{\partial \theta_2} = -\frac{2}{N} \sum_{i=1}^{N} \left(y^{(i)} - \hat{y}^{(i)}\right)$$

$$\theta \leftarrow \theta - \alpha \nabla J$$

$\theta^{(0)} = (0, 0) \qquad \alpha = 0.01$

$\vec{\theta}^{\circ} = (0,0)$

$\theta^{\circ} \quad \theta_1 = 0,\ \theta_2 = 0 \rightarrow \hat{y} = 0 \cdot x + 0 = 0$

$\hat{y}(1) = 0$ | Predictions all 0

$y(2) = 0$ | Residuals r: $-3, -4, -6, -5$

$y(3) = 0$

$y(4) = 0$

x	y	$\hat{y}$	$r = y - \hat{y}$	r^2
1	3	0	−3	9
2	4	0	−4	16
3	6	0	−6	36
4	5	0	−5	25

sums, $\Sigma r = -18$

$\cancel{\Sigma xr = 3+8+18+20 = 49}$

$\Sigma xr = 1(-3) + 2(-4) + 3(-6) + 4(-5)$

$= -3 - 8 - 18 - 20 = -49$

Gradient

$$\nabla J^{(0)} = \left(\frac{2}{4}(-49),\ \frac{2}{4}(-18) \right) = (-24.5, -9)$$

Update

$$\theta^{(1)} = \theta^{(0)} - \alpha \nabla J^{(0)} = (0,0) - (0.01)(-24.5, -9)$$
$$= (0.245, 0.09)$$

Costs :-

$$J(\theta^{(0)}) = \frac{9+16+36+25}{4} = \frac{86}{4} = 21.5$$

$J(\theta^{(1)}) \cong 15.2560$ (computed from the new residuals)

Second iteration

→ Second step starting at $\theta^{(1)} = (0.245, 0.09)$

Predictions: $\hat{y} - 0.335, 0.58, 0.825, 1.07$

Residuals $r = \hat{y} - y$, $-2.665, -3.42, -5.175, -3.93$

Sums: $\Sigma r = -15.19$: $\Sigma xr = 1(-2.665) + 2(-3.42) + 3(-5.175) + 4(-3.93) = -40.75$

Gradient:

$$\nabla J^{(1)} = \left(\frac{2}{4}(-40.75), \frac{2}{4}(-15.19)\right) = (-20.375, -7.595)$$

Update

$$\theta^{(2)} = \theta^{(1)} - \alpha \nabla J^{(1)} = (0.245, 0.09) - 0.01(-20.375, -7.595)$$
$$= (0.448, 0.16595)$$

Costs :-

$J(\theta^{(1)}) = 15.2560$

$J(\theta^{(2)}) = 10.9223$ (decreased again)

First Iteration

1. Predictions at θ^0 : $[0,0,0,0]$
2. Residuals $[3,4,6,5]$, $\Sigma r = 18$, $\Sigma xr = -49$
3. Gradient $\nabla J = (-24.5, -9)$
4. $\theta' = (0.245, 0.09)$
5. $J(\theta^0) = 21.5$, $J(\theta') = 15.2560$

Second Iteration

1. Predictions at θ': $[0.335, 0.58, 0.825, 1.07]$
2. Residuals $[2.665, 3.42, 5.175, 3.93]$, $\Sigma r = -15.19$, $\Sigma xr = -40.75$
3. Gradients $\nabla J = (-20.375, -7.595)$
4. $\theta^2 = (0.448, 0.16595)$
5. $J(\theta') = 15.2560$, $J(\theta^2) = 10.9223$

* Cost decreased, so there is improvement.

4) Data = (1,2) (2,2) (3,4) (4,6)

$$MSE = J(\theta) = \frac{1}{N} \sum_{i=1}^{N} (y^i - (\theta x^i + \theta_2))^2$$

→ 1. $(\theta_1, \theta_2) = (0.2, 0.5)$ and $(\theta_1, \theta_2) = (0.9, 0.1)$

$(\theta_1, \theta_2) = (0.2, 0.5)$

$\hat{y} = 0.2x + 0.5$

x	$\hat{y}$	$r = y - \hat{y}$	r^2
1	$0.2(1) + 0.5 = 0.7$	$2 - 0.7 = 1.3$	1.69
2	$0.2(2) + 0.5 = 0.9$	$2 - 0.9 = 1.1$	1.21
3	$0.2(3) + 0.5 = 1.1$	$4 - 1.1 = 2.9$	8.41
4	$0.2(4) + 0.5 = 1.3$	$6 - 1.3 = 4.7$	22.09

$\Sigma r^2 = 33.40$

$$J = \frac{1}{4}(33.40)$$
$$= 8.35$$

$(\theta_1, \theta_2) = (0.9, 0.1)$

$\hat{y} = 0.9x + 0.1$

x	$\hat{y}$	$r = y - \hat{y}$	r^2
1	$0.9(1) + 0.1 = 1$	$2 - 1 = 1$	$(1)^2 = 1$
2	$0.9(2) + 0.1 = 1.9$	$2 - 1.9 = 0.1$	$(0.1)^2 = 0.01$
3	$0.9(3) + 0.1 = 2.8$	$4 - 2.8 = 1.2$	$(1.2)^2 = 1.44$
4	$0.9(4) + 0.1 = 3.7$	$6 - 3.7 = 2.3$	$(2.3)^2 = 5.29$

$\Sigma r^2 = 7.74$

$$J = \frac{1}{4}(7.74)$$
$$= 1.935$$

$J(0.2, 0.5) = 8.35$

$J(0.9, 0.1) = 1.935 \approx 1.94$

→2. Gradient Descent (Second Gradient Descent)

$\theta = (0,0),\ \alpha = 0.01$

$i = 0x + 0 = 0$

$i = (2, 2, 4, 6)$

$\Sigma x_i r_i = 1 \cdot 2 + 2 \cdot 2 + 3 \cdot 4 + 4 \cdot 6 = 42$

$\Sigma r_i = 2 + 2 + 4 + 6 = 14$

$$\frac{\partial J}{\partial \theta_1} = \frac{-2}{4}(42) = -21$$

$$\frac{\partial J}{\partial \theta_2} = \frac{-2}{4}(14) = -7 \qquad \nabla J\ (-21, -7)$$

$\theta' = \theta^{(0)} - \alpha(-21, -7)$

$= (0,0) - 0.01(-21, -7)$

$\theta' = (0.21, 0.7)$

$\hat{y} = 0.21x + 0.7$

x	$\hat{y}$	$r = y - \hat{y}$	r^2
1	$0.21(1) + 0.7 = 0.28$	$2 - 0.28 = 1.72$	$(1.72)^2 = 2.9$
2	$0.21(2) + 0.7 = 0.49$	$2 - 0.49 = 1.51$	$(1.51)^2 = 2.28$
3	$0.21(3) + 0.7 = 0.70$	$4 - 0.70 = 3.30$	$(3.30)^2 = 10.89$
4	$0.21(4) + 0.7 = 0.91$	$6 - 0.91 = 5.09$	$(5.09)^2 = 25.9$

$\Sigma r^2 = 42.03$

$$J(\theta') = \frac{1}{4}(42.03) = 10.50$$

$J(0.2, 0.5) = 8.35$

$J(0.9, 0.1) = 1.935 \approx 1.94$

from $(0,0)$ $\alpha = 0.01$

$J(0.21, 0.07) = 10.51$

→3. The random guess $(0.9, 0.1)$ gave the lowest error $J = 1.935$ which is much lower than other random guess and first gradient descent step.

Gradient descent starts at $(0,0)$ predicting zeros with small step $(\alpha = 0.01)$, $\theta' = (0.21, 0.07)$ stays near zero and gradient descent is systematic but slow initially with small learning rate. While a good random guess can land near the optimum.

5) Recognizing underfitting and overfitting:

1. Underfitting
2. Underfitting occurs when the model is too simple to capture the underlying patterns in data.
   * It fails to fit the training data well, leading to high training error.
   * Consequently it also generalize poorly, so test error remains high.
3. * Use a more complex model
   * Relax Regularization if its too strong.
   * Increase model capacity (like more layers)

6) Comparing Models:

1. Model A is overfitting, it fits training data perfectly but fails on new data.

   Model B is underfitting, performs poorly on both training and test data.

2. Model A

Low bias, model is flexible enough to fit training data.

High variance, model is sensitive to the specific of training sample.

Model B

High Bias, model is too simple

low variance (not sensitive to the specific of training data)

3. Model A (Improvements)

→ Simplify the model

→ Strengthen Regularization

→ Increase training data

Model B (Improvements)

→ Increase Model capacity

→ Use complex model

→ Add Informative features

→ Reduce Regularization