CS303L/CS519L, Test 1, Dr. Y. Kalidas, 2023/Sep/05, 15:00-16:00, 100 marks

Q1) Provide gradient descent weight update equation for 10 degree uni-variate polynomial fitting for mean squared error loss function and explain each term of the equation. Clearly indicate correct matrix dimensionalities. (10 marks)

(Revised) Q1 Ans.

Expected answer:

Define $X_{N\times 11}$ matrix and $y_{N\times 1}$ vector (1 marks)

$w_{11\times 1}$: vector of parameters (1 m);

$L = (Xw - y)^T(Xw - y)$ (1 marks)

$\nabla L_{11\times 1}$: vector of functions (1 m);

For squared error, the vector is given by, $\nabla L = [\sum_{i=1}^{i=N}(2 \times (y_i - \sum_{j=0}^{j=10} a_j x_{i,j}) \times x_{i,j})]_{j=[0,...,10]}$

Or alternatively

$\nabla L = 2X^T(Xw - y)$

$w^{(new)} = w^{(old)} - \alpha \times \nabla L|_{w = w^{(old)}}$ (3 marks) //expand gradient

α: learning rate (1m)

$\alpha > 0$ (1m)

and is small (1m)

Common mistakes:

1) Incorrect dimensions
2) α not defined or not explained
3) Missing weight update equation
4) Incorrect loss function formula
5) X or y are not defined
6) ∇L is mentioned but $\frac{\partial L}{\partial w_i}$ is not expanded

Q2) Provide gradient descent weight update equation for 10 degree 12-variate polynomial fitting for mean squared error loss function and explain each term of the equation. Clearly indicate correct matrix dimensionalities. (10 marks)

(Revised) Q2 Ans.

Define $X_{N\times 11}$ matrix and $Y_{N\times 12}$ matrix (1 marks)

$W_{11\times12}$: matrix of parameters (1 marks);

$L = tr\,(Xw - Y)^T(Xw - Y)$ (1 marks)

$\nabla L_{11\times12}$: matrix of functions (1 marks);

For squared error, the matrix is given by,

$$\nabla L = [\sum_{i=1}^{i=N}(2 \times (y_i - \sum_{j=0}^{j=10} a_{j,k}x_{i,j}) \times x_{i,j})]_{j=[0,\dots,10],k=[1,\dots,12]}$$

Or alternatively

$\nabla L = 2X^T(Xw - Y)$

$W^{(new)} = W^{(old)} - \alpha \times \nabla L|_{W = W^{(old)}}$ (3 marks)

α: learning rate (1m)

$\alpha > 0$ (1m)

and is small (1m)

Common mistakes:

1) Incorrect dimensions
2) α: not defined or not explained
3) Missing weight update equation or incorrect weight update equation
4) Missing loss function
5) trace is missing in loss function

Q3) Show the derivation of gradient descent learning law for $y = mx$. Clearly explain using diagrammatic representation based on coordinate geometry, points, lines, and axes. (10 marks)

(Revised) Q3 Ans.

$L(m + h) = L(m) + h \times L'(m)$ (1 mark)

$Let\ \exists h = -\alpha \times L'(m)$ (1 mark)

$\rightarrow L(m + h) = L(m) + -\alpha \times L'(m) \times L'(m)$ (2 mark)

$\rightarrow L(m + h) < L(m)$ (1 mark)

(Diagram below: 5 marks for 4 coordinates, axes labels, slope indication)

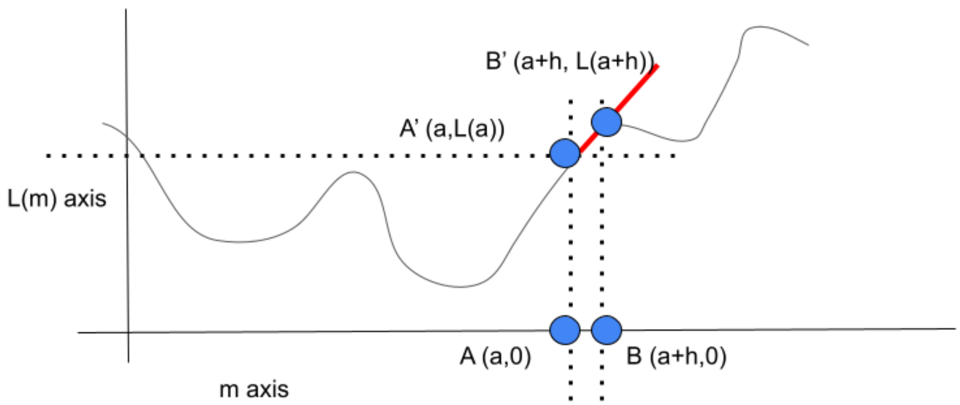
Or alternatively,

$f(x + h)$ or $f(m + h)$ notation is also considered.

Common mistakes:

1) The variable should have been m but x is chosen as variable
2) The function should have been $L(m)$ but $f(x)$ is chosen as the function
3) In place of m being updated, $L(a)$ or $f(a)$ were used and a was updated

4) Diagram is missing
5) Coordinate are not indicated
6) Axes labels are not shown
7) Derivation has errors
8) Directly formula is shown without derivation



Q4) Show the derivation of Newton-Raphson for <u>root finding</u> for $y = m\,x$. Use coordinate geometry notation. (10 marks)

(Revised) Ans Q4.
Alternative variations that are also ok are

1) $tan(\theta)$
2) Not using α is ok for root-finding alone

Common mistakes:

1. The variable should have been m but x is chosen as variable
2. The function should have been $L(m)$ but $f(x)$ is chosen as the function
3. In place of m being updated, $L(a)$ or $f(a)$ were used and a was updated
4. Derivation has errors or incomplete
5. Directly formula is shown without derivation

$$L'(a) = Lt_{h\to 0} \frac{L(a+h) - L(a)}{h} \text{ (2 mark)}$$

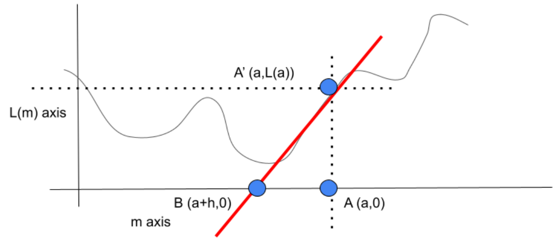
$$h = \frac{L(a+h) - L(a)}{L'(a)} \text{ (2 mark)}$$

$$h = \frac{0 - L(a)}{L'(a)} = \frac{-L(a)}{L'(a)} \text{ (2 mark)}$$

$$a^{(new)} = a + h \text{ (1 marks)}$$

$$a^{(old)} = a \text{ (1 marks)}$$

$$a^{(new)} = a^{(old)} - \frac{L(a)}{L'(a)} \text{ (2 mark)}$$



Q5) Show the derivation of Newton-Raphson for <u>minimization</u> for $y = m\,x$ based on Newton-Raphson for root finding. (10 marks)

(Revised) Ans Q5.

$Let,\ g(m) = L'(m)$ (2 marks)

At minimum value, $(\exists m):\ L'(m) = 0$ (2 marks)

Apply root finding for $g(m)$ (2 marks)

$$m^{(new)} = m^{(old)} - \alpha \times \frac{g(m)}{g'(m)}\Big|_{m=m^{(old)}} \text{ (2 marks)}$$

$$m^{(new)} = m^{(old)} - \alpha \times \frac{L'(m)}{L''(m)}\Big|_{m=m^{(old)}} \text{ (2 marks)}$$

Common mistakes:

1. The variable should have been m but x is chosen as variable
2. The function should have been $L(m)$ but $f(x)$ is chosen as the function
3. In place of m being updated, $L(a)$ or $f(a)$ were used and a was updated
4. Diagram is missing (but ok), Coordinate are not indicated, Axes labels are not shown
5. Incomplete derivation
6. Directly formula is shown without derivation
7. α is missing or not described

Q6) What are the five elements of machine learning? Provide proper order of the steps and required details. (10 marks)

(Revised) Ans Q6. (i) x vector, (ii) y vector, (iii) mapping function $\hat{y} = f(x)$, (iv) loss function $l(y, \hat{y})$ and (v) data set $D = \{(x_1, y_1), ..., (x_N, y_N)\}$.

Common mistakes:

1) Input word should be used before x (but ok)
2) Output word should be used before y (but ok)

Q7) Derive condition for the bowl shape of squared error loss function for $y = m\,x + c$ over a data set of $D = \{(x_1, y_1), ..., (x_N, y_N)\}$. (10 marks)

(Revised) Ans Q7).

$$L(m, c) = \sum_{i=1}^{i=N} (y_i - (m\,x_i + c))^2$$

Construct Hessian, $H = [$

$\frac{\partial^2 L}{\partial m^2}, \frac{\partial^2 L}{\partial m\,\partial c}$

$\frac{\partial^2 L}{\partial c\,\partial m}, \frac{\partial^2 L}{\partial c^2}]$

To show that Hessian is full of constant values

It is positive definite $p = [a, b]^T$

$p^T Hp > 0\ \forall p$

Common mistakes:

1) Students have determined minimization condition for m separately and for c separately but not considered $\frac{\partial^2 L}{\partial c\,\partial m}$
2) Very few students have gone till the Hessian, but some mistakes in proving positive definiteness
3) Some students tried to determine minimum m and c values, but incorrectly equated $\frac{\partial L}{\partial m}$ and $\frac{\partial L}{\partial c}$ to zero by missing out dependent terms

Q8) Consider the problem of mapping an English sentence to a Hindi sentence. How would you apply machine learning methodology? Use simple terminology introduced in the class. (Do not complicate your answer through LSTM etc it's unnecessary to explain now). Very clearly indicate your thought process by steps, number each step and be unambiguous. (10 marks)

Ans Q8.
Step 1: Prepare a vocabulary set for English.
Step 2: Prepare x vector based on presence or absence of each English word.
Step 3: Prepare a vocabulary set for Hindi.
Step 4: Prepare a y vector based on presence or absence of each Hindi word.
Step 5: Prepare a mapping function from x to y. The mapping function can have a pipeline of transformations.
Step 6: Prepare a loss function.
Step 7: Prepare a data set of (x,y) tuples for pairs of English and Hindi sentences.

Common mistakes:

1) Feature engineering details are missing for x and y vectors. Only vague statements are presented. Just the word one-hot encoding is not sufficient.
2) Missing mapping function, loss function
3) Missing data set creation details