CS2109S Tutorial 3

Decision Trees and Linear Models

(AY 25/26 Semester 2)

February 12, 2026

(Prepared by Benson)

Admin Info

- We will do Zoom tutorial next week!
  - The session will be recorded.
  - Still, I encourage you to come! There will be many important discussions.

Contents

Recap: Decision Tree Learning

- **Information content / "surprisal"**: The informational value of communicating that an event happened.

$$I(e) = \log\left(\frac{1}{p}\right) = -\log p \text{ bits}$$

(to bits; frequency of occurrence)

- **Entropy**: The expected amount of information conveyed by identifying the outcome of a random trial.

$$H(X) = \sum_{e \in E} P(e) I(e) = -\sum_{e \in E} P(e) \log P(e)$$

- **Information gain**: Difference between the entropy before the split and the expected entropy (of a sample) after the split.

$$IG(D, A) = H(D) - \sum_{v \in A} \frac{|D_v|}{|D|} H(D_v)$$

Remainder (minimize this)

Q1. Decision Tree

(a) Construct the best decision tree. Calculate the information gain values and remainders at each stage.

$$\begin{aligned}\text{remainder(Experience)} &= \frac{6}{10}H(3,3) + \frac{4}{10}H(2,2) \\ &= \frac{6}{10}(1) + \frac{4}{10}(1) \\ &= 1\end{aligned}$$

$$\begin{aligned}\text{remainder(Edu. Level)} &= \frac{4}{10}H(3,1) + \frac{3}{10}H(2,1) + \frac{3}{10}H(0,3) \\ &= \frac{4}{10}(0.8113) + \frac{3}{10}(0.9183) + \frac{3}{10}(0) \\ &= 0.600\end{aligned}$$

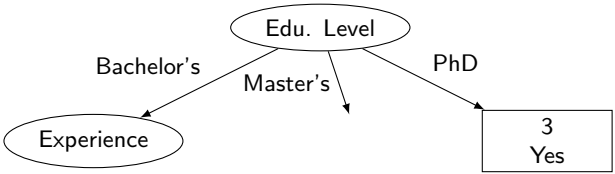
$$\begin{aligned}\text{remainder(Age)} &= \frac{5}{10}H(2,3) + \frac{5}{10}H(3,2) \\ &= \frac{5}{10}(0.9710) + \frac{5}{10}(0.9710) \\ &= 0.971\end{aligned}$$

Experience	Edu. Level	Age	Hire?
Low	Bachelor's	< 30	No
Low	Master's	< 30	Yes
Low	PhD	< 30	Yes
Low	Bachelor's	≥ 30	No
Low	Master's	≥ 30	No
Low	PhD	≥ 30	Yes
High	Bachelor's	< 30	No
High	PhD	< 30	Yes
High	Master's	≥ 30	No
High	Bachelor's	≥ 30	Yes

+ \ −	0	1	2	3	4	5
1	0	1	0.9183	0.8113	0.7219	0.6500
2	0	0.9183	1	0.9710	0.9183	0.8631
3	0	0.8113	0.9710	1	0.9852	0.9544
4	0	0.7219	0.9183	0.9852	1	0.9911
5	0	0.6500	0.8631	0.9544	0.9911	1
6	0	0.5917	0.8113	0.9183	0.9710	0.9940
7	0	0.5436	0.7642	0.8813	0.9457	0.9799

Q1. Decision Tree

(a) Construct the best decision tree. Calculate the information gain values and remainders at each stage.



$$\text{remainder(Experience)} = \frac{2}{4}H(2,0) + \frac{2}{4}H(1,1)$$
$$= \frac{2}{4}(0) + \frac{2}{4}(1) = 0.5$$

$$\text{remainder(Age)} = \frac{2}{4}H(2,0) + \frac{2}{4}H(1,1)$$
$$= \frac{2}{4}(0) + \frac{2}{4}(1) = 0.5$$

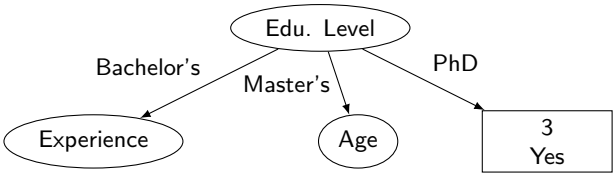
We choose Experience according to the tie-breaking rule.

Experience	Edu. Level	Age	Hire?
Low	Bachelor's	< 30	No
Low	Master's	< 30	Yes
Low	PhD	< 30	Yes
Low	Bachelor's	≥ 30	No
Low	Master's	≥ 30	No
Low	PhD	≥ 30	Yes
High	Bachelor's	< 30	No
High	PhD	< 30	Yes
High	Master's	≥ 30	No
High	Bachelor's	≥ 30	Yes

+ \ −	0	1	2	3	4	5
1	0	1	0.9183	0.8113	0.7219	0.6500
2	0	0.9183	1	0.9710	0.9183	0.8631
3	0	0.8113	0.9710	1	0.9852	0.9544
4	0	0.7219	0.9183	0.9852	1	0.9911
5	0	0.6500	0.8631	0.9544	0.9911	1
6	0	0.5917	0.8113	0.9183	0.9710	0.9940
7	0	0.5436	0.7642	0.8813	0.9457	0.9799

Q1. Decision Tree

(a) Construct the best decision tree. Calculate the information gain values and remainders at each stage.



$$\text{remainder(Experience)} = \frac{2}{3}H(1,1) + \frac{1}{3}H(1,0)$$
$$= \frac{2}{3}(1) + \frac{1}{3}(0) = 0.667$$

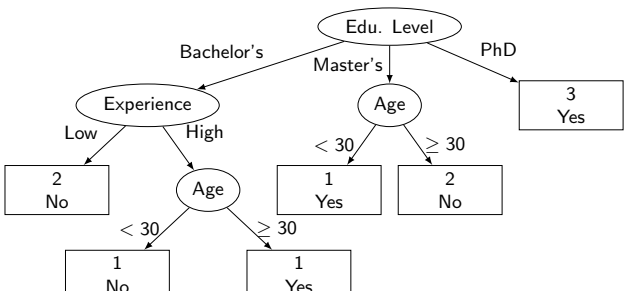
$$\text{remainder(Age)} = \frac{1}{3}H(0,1) + \frac{2}{3}H(2,0)$$
$$= \frac{1}{3}(0) + \frac{2}{3}(0) = 0$$

Experience	Edu. Level	Age	Hire?
Low	Bachelor's	< 30	No
Low	Master's	< 30	Yes
Low	PhD	< 30	Yes
Low	Bachelor's	≥ 30	No
Low	Master's	≥ 30	No
Low	PhD	≥ 30	Yes
High	Bachelor's	< 30	No
High	PhD	< 30	Yes
High	Master's	≥ 30	No
High	Bachelor's	≥ 30	Yes

+ \ −	0	1	2	3	4	5
1	0	1	0.9183	0.8113	0.7219	0.6500
2	0	0.9183	1	0.9710	0.9183	0.8631
3	0	0.8113	0.9710	1	0.9852	0.9544
4	0	0.7219	0.9183	0.9852	1	0.9911
5	0	0.6500	0.8631	0.9544	0.9911	1
6	0	0.5917	0.8113	0.9183	0.9710	0.9940
7	0	0.5436	0.7642	0.8813	0.9457	0.9799

Q1. Decision Tree

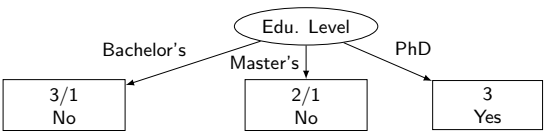
(a) Construct the best decision tree. Calculate the information gain values and remainders at each stage.



Experience	Edu. Level	Age	Hire?
Low	Bachelor's	< 30	No
Low	Master's	< 30	Yes
Low	PhD	< 30	Yes
Low	Bachelor's	≥ 30	No
Low	Master's	≥ 30	No
Low	PhD	≥ 30	Yes
High	Bachelor's	< 30	No
High	PhD	< 30	Yes
High	Master's	≥ 30	No
High	Bachelor's	≥ 30	Yes

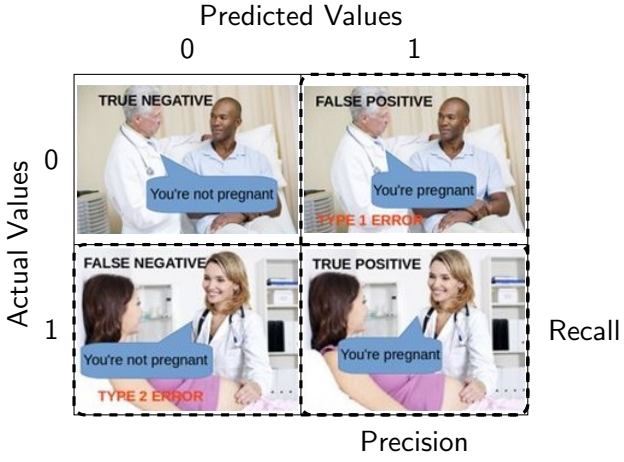
Q1. Decision Tree

(b) Prune your tree so that every leaf node contains at least three samples and outputs the majority decision.



Experience	Edu. Level	Age	Hire?
Low	Bachelor's	< 30	No
Low	Master's	< 30	Yes
Low	PhD	< 30	Yes
Low	Bachelor's	≥ 30	No
Low	Master's	≥ 30	No
Low	PhD	≥ 30	Yes
High	Bachelor's	< 30	No
High	PhD	< 30	Yes
High	Master's	≥ 30	No
High	Bachelor's	≥ 30	Yes

Q1. Decision Tree



Q1. Decision Tree

(c) Compute the Confusion Matrix, Accuracy and F1 Score.

Experience	Edu. Level	Age	Hire?
Low	Bachelor's	< 30	No
Low	PhD	< 30	No
Low	Master's	≥ 30	Yes
Low	PhD	≥ 30	Yes
High	Bachelor's	< 30	Yes
High	Master's	< 30	Yes
High	Bachelor's	≥ 30	Yes

- Confusion Matrix:

Actual Values \ Predicted Values	0	1
0	1	1
1	2	3

- Accuracy $= \dfrac{3+1}{7} = \dfrac{4}{7}$
- Precision $= \dfrac{TP}{TP+FP} = \dfrac{3}{3+1} = \dfrac{3}{4}$
- Recall $= \dfrac{TP}{TP+FN} = \dfrac{3}{3+2} = \dfrac{3}{5}$
- F1 Score $= \dfrac{2}{1/P+1/R} = \dfrac{2}{4/3+5/3} = \dfrac{2}{3}$

Q2. Linear Regression Model Fitting

(a) Apply the Normal Equation formula to obtain a linear regression model that minimizes MSE of the data points.

$$w = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

$$\phi(x_1, x_2) = \begin{bmatrix} 1 & x_1^2 & x_1 x_2 & x_2^2 \end{bmatrix}^\top$$

x_1	x_2	y
3	2	12
5	3	20
7	4	30
9	6	42

Solution.

$$\boldsymbol{X} = \begin{bmatrix} 1 & 9 & 6 & 4 \\ 1 & 25 & 15 & 9 \\ 1 & 49 & 28 & 16 \\ 1 & 81 & 54 & 36 \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} 12 \\ 20 \\ 30 \\ 42 \end{bmatrix}$$

$$\boldsymbol{w} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} = \begin{bmatrix} 8.25 & -6.25 & 22 & -18 \end{bmatrix}^\top$$

$$\therefore \hat{y} = 8.25 - 6.25 x_1^2 + 22 x_1 x_2 - 18 x_2^2.$$

Q2. Linear Regression Model Fitting

Extra Slide

Lemma. $\boldsymbol{X}^\top \boldsymbol{X}$ is invertible if and only if all columns of $\boldsymbol{X}$ are linearly independent.

Proof.

$\boldsymbol{X}^\top \boldsymbol{X}$ is invertible

$\Leftrightarrow \text{rank}(\boldsymbol{X}^\top \boldsymbol{X}) = m$

$\Leftrightarrow \text{rank}(\boldsymbol{X}) = m$

◀ since $\boldsymbol{X}^\top \boldsymbol{X}$ is an $m \times m$ matrix

◀ $\text{rank}(\boldsymbol{X}^\top \boldsymbol{X}) = \text{rank}(\boldsymbol{X})$

MA1522/2001 Refresher: $\text{rank}(\boldsymbol{X}^\top \boldsymbol{X}) = \text{rank}(\boldsymbol{X})$. (Prove this by showing the nullspace of $\boldsymbol{X}^\top \boldsymbol{X}$ is equal to the nullspace of $\boldsymbol{X}$. See tutorial 7 for either course.)

Q2. Linear Regression Model Fitting

Extra Slide

Demo: Ill-conditioned matrices (see HTML version). Numerical stability is important!

Q2. Linear Regression Model Fitting

Extra Slide

Exercise: Can you find the linearly dependent features in this dataset?

Q2. Linear Regression Model Fitting

(b) Show that the matrix is singular after adding a transformed feature $(x_1 + x_2)^2$. State a possible solution.

- $(x_1 + x_2)^2 = \boxed{x_1^2} + 2\boxed{x_1 x_2} + \boxed{x_2^2}$
- $(x_1 + x_2)^2$ is **linearly dependent** on the other transformed features $\Rightarrow$ The matrix is singular.
- There are infinitely many solutions (thus it's impossible to find an "unique optimal solution").
- *Almost* linearly dependent columns create problems too...
  - A slight change in values drastically affects the result.
- **Solution**: Remove linearly dependent features. / Use gradient descent instead.

Q2. Linear Regression Model Fitting

Extra Slide

Ridge regression:

$$J_{reg}(\boldsymbol{w}) = \underbrace{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}^{(i)} - y^{(i)})^2}_{\text{data fit}} + \underbrace{\lambda\|\boldsymbol{w}\|^2}_{\text{complexity penalty}}$$

$$\boldsymbol{w} = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}^\top \boldsymbol{y}$$

Intuition 1:

- A matrix is invertible iff all of its eigenvalues are non-zero.
- $\boldsymbol{X}^\top \boldsymbol{X}$ has non-negative eigenvalues (see "Gram matrix").
- Adding $\lambda \boldsymbol{I}$ to $\boldsymbol{X}^\top \boldsymbol{X}$ shifts all eigenvalues of the matrix by λ.
- $\therefore \boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}$ is always invertible.

Intuition 2:

- Given linearly dependent features, there are infinitely many ways to assign weights to fit the data.
- But there is only one way to minimize the total weight assigned to them.

Q2. Linear Regression Model Fitting

Extra Slide

Pseudoinverse:

- We cannot find $\boldsymbol{A}^{+}$ such that $\boldsymbol{A}\boldsymbol{A}^{+} = \boldsymbol{I}$.
- But we can guarantee $\boldsymbol{A}\boldsymbol{A}^{+}\boldsymbol{A} = \boldsymbol{A}$ and $\boldsymbol{A}^{+}\boldsymbol{A}\boldsymbol{A}^{+} = \boldsymbol{A}^{+}$.
- Computed using singular value decomposition (SVD).

Q2. Linear Regression Model Fitting

(c) Given two invertible square matrices A and B of the same dimension, show that the inverse of their product is the product of their inverses in reverse order:

$$(\boldsymbol{AB})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$$

Goal: To show that $(\boldsymbol{B}^{-1}\boldsymbol{A}^{-1})(\boldsymbol{AB}) = \boldsymbol{I}$ and $(\boldsymbol{AB})(\boldsymbol{B}^{-1}\boldsymbol{A}^{-1}) = \boldsymbol{I}$.

Left Multiplication:

$$\begin{aligned}(\boldsymbol{B}^{-1}\boldsymbol{A}^{-1})(\boldsymbol{AB}) &= \boldsymbol{B}^{-1}(\boldsymbol{A}^{-1}\boldsymbol{A})\boldsymbol{B} \\ &= \boldsymbol{B}^{-1}\boldsymbol{I}\boldsymbol{B} \\ &= \boldsymbol{B}^{-1}\boldsymbol{B} \\ &= \boldsymbol{I}\end{aligned}$$

Right Multiplication:

$$\begin{aligned}(\boldsymbol{AB})(\boldsymbol{B}^{-1}\boldsymbol{A}^{-1}) &= \boldsymbol{A}(\boldsymbol{B}\boldsymbol{B}^{-1})\boldsymbol{A}^{-1} \\ &= \boldsymbol{A}\boldsymbol{I}\boldsymbol{A}^{-1} \\ &= \boldsymbol{A}\boldsymbol{A}^{-1} \\ &= \boldsymbol{I}\end{aligned}$$

Q3. Examining Cost Functions

Mean Squared Error: $L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$

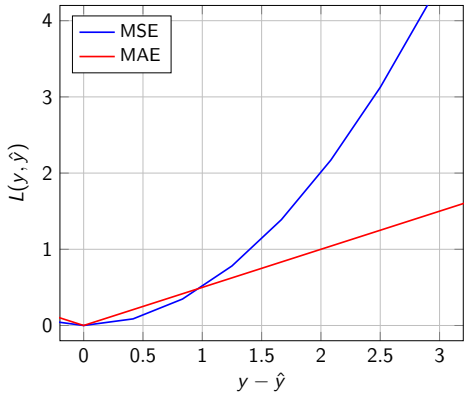
Mean Absolute Error: $L(y, \hat{y}) = \frac{1}{2}|y - \hat{y}|$

(a) Calculate the MSE and MAE of the data points.

- MSE $= \frac{1}{2}(2 - 2.1)^2 = 0.005$

  MAE $= \frac{1}{2}|2 - 2.1| = 0.05$
- MSE $= \frac{1}{2}(4 - 4.9)^2 = 0.405$

  MAE $= \frac{1}{2}|4 - 4.9| = 0.45$

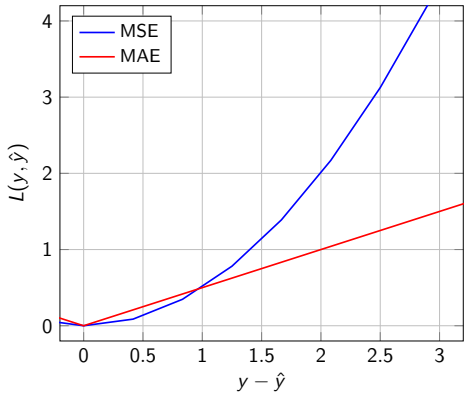


$\hat{y}$	y
2	2.1
4	4.9

Q3. Examining Cost Functions

(b) Discuss how outliers affect MSE and MAE differently.

- For MSE,
$$\frac{L(y_2, \hat{y}_2)}{L(y_1, \hat{y}_1)} = \frac{0.405}{0.005} = 81$$
- For MAE,
$$\frac{L(y_2, \hat{y}_2)}{L(y_1, \hat{y}_1)} = \frac{0.45}{0.05} = 9$$
- $\therefore$ MSE is more sensitive to outliers.

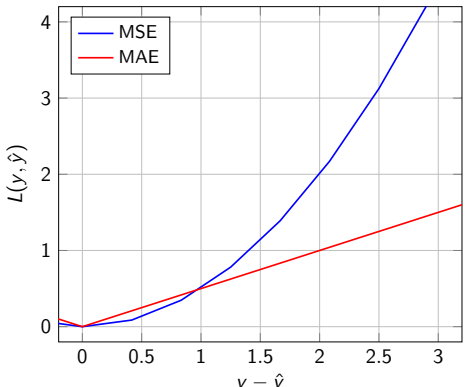


$\hat{y}$	y	
2	2.1	Not an outlier
4	4.9	!!! Outlier

Q3. Examining Cost Functions

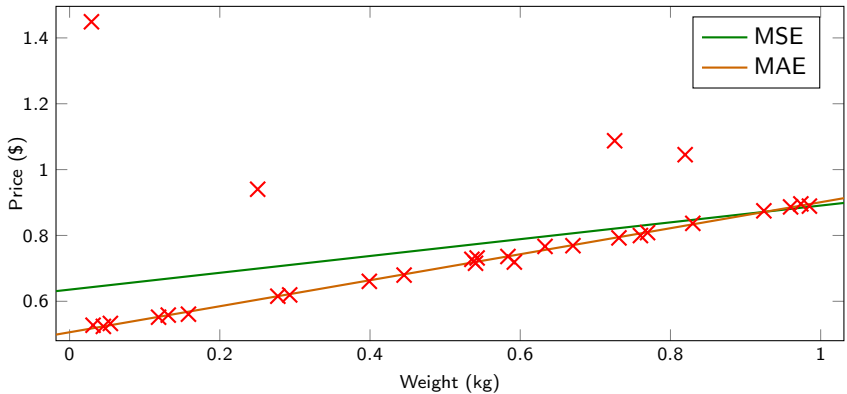
True or False:

1. Consider a dataset where all y values are larger than 1. MSE penalizes the outliers more heavily than MAE.
2. Consider a dataset where all y values are between 0 and 1. MSE penalizes the outliers more heavily than MAE.



Q3. Examining Cost Functions

- Which line corresponds to MSE? Which line corresponds to MAE?



Q3. Examining Cost Functions

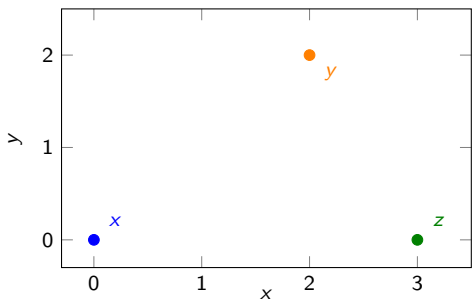
(c) For any two points $a, b \in \mathbb{R}^n$, the L1 distance between a and b is defined as $d_1(a, b) = \sum_{i=1}^{n} |a_i - b_i|$, while the L2 distance is defined as $d_2(a, b) = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}$. Consider three arbitrary points $x, y, z \in \mathbb{R}^2$. If y is the nearest neighbour of x in terms of the L2 distance, show that y may not be the nearest neighbour of x in terms of the L1 distance.

L1 distance:

- x to y: $|2 - 0| + |2 - 0| = 4$.
- x to z: $|3 - 0| + |0 - 0| = 3$.

L2 distance:

- x to y: $\sqrt{(2 - 0)^2 + (2 - 0)^2} \approx 2.828$.
- x to z: $\sqrt{(3 - 0)^2 + (0 - 0)^2} = 3$.



Q3. Examining Cost Functions

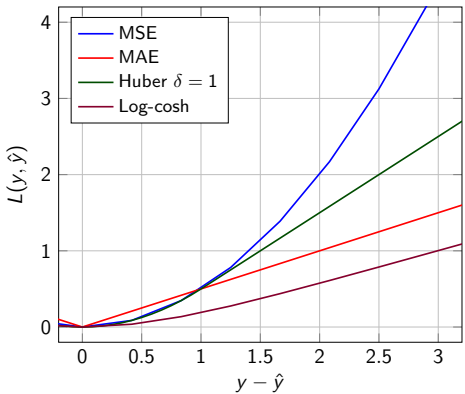
(d) Can you provide examples of cost functions that are better suited to handle outliers more effectively?

Huber loss: MSE $\rightarrow$ MAE.

$$L(y, \hat{y}) = \begin{cases} \frac{1}{2}(y - \hat{y})^2 & \text{for } |y - \hat{y}| \leq \delta \\ \delta \cdot |y - \hat{y}| - \frac{1}{2}\delta^2 & \text{otherwise} \end{cases}$$

Log-cosh loss: $\approx \frac{1}{2}x^2$ for small $x \rightarrow$ $\approx |x| - \log 2$ for large x.

$$L(y, \hat{y}) = \log(\cosh(y - \hat{y}))$$

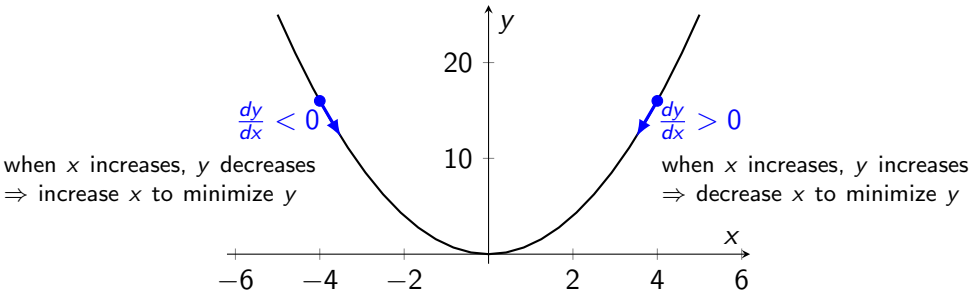


Recap: Gradient Descent

We wish to minimize the loss y by varying x.

- Parameters: Initial value x, Learning rate α. Decides this

How far do we change x?



Theorem. If α is **small enough**, gradient descent would reach a local minima. This would also be the global minima if $f(x)$ is **convex**.

For convex functions, local minima ⇒ global minima.

Recap: Gradient Descent

Assume α is negative. Which of the following is the equation to perform the updates?

A. $x \leftarrow x + \alpha \frac{dy}{dx}$

B. $x \leftarrow x - \alpha \frac{dy}{dx}$

C. $y \leftarrow y + \alpha \frac{dy}{dx}$

D. $y \leftarrow y - \alpha \frac{dy}{dx}$

Solution. When $\frac{dy}{dx} > 0$, we need to decrease x.
Only option A successfully decreases x.

Q4. Gradient Descent

(a) Compute the minimizer $\boldsymbol{x}^*$ of f analytically that gives the global minimum value of f.

$$f(x_1, x_2) = 0.5x_1^2 + x_2^2 + 2x_1 + x_2 + 3$$

$$\frac{df(x_1, x_2)}{dx_1} = 0 \quad \Rightarrow \quad x_1 + 2 = 0$$

$$\frac{df(x_1, x_2)}{dx_2} = 0 \quad \Rightarrow \quad 2x_2 + 1 = 0$$

$$\therefore \boldsymbol{x}^* = \begin{bmatrix} -2 & -1/2 \end{bmatrix}^\top.$$

Q4. Gradient Descent

(b) Perform 3 steps of gradient descent on f starting from the point $\boldsymbol{x}^{(0)} = (0, 0)$, with a constant learning rate $\gamma = 1$.

- $\dfrac{df(x_1, x_2)}{dx_1} = x_1 + 2$ $\qquad$ $\dfrac{df(x_1, x_2)}{dx_2} = 2x_2 + 1$
- Step 1:
  $\left.\dfrac{df(x_1, x_2)}{dx_1}\right|_{\boldsymbol{x}=(0,0)} = 2$ $\qquad$ $\left.\dfrac{df(x_1, x_2)}{dx_2}\right|_{\boldsymbol{x}=(0,0)} = 1$
  $\boldsymbol{x}^{(1)} = \boldsymbol{x}^{(0)} - \gamma(2, 1) = (-2, -1).$
- Step 2:
  $\left.\dfrac{df(x_1, x_2)}{dx_1}\right|_{\boldsymbol{x}=(-2,-1)} = 0$ $\qquad$ $\left.\dfrac{df(x_1, x_2)}{dx_2}\right|_{\boldsymbol{x}=(-2,-1)} = -1$
  $\boldsymbol{x}^{(2)} = \boldsymbol{x}^{(1)} - \gamma(0, -1) = (-2, 0).$
- Step 3:
  $\left.\dfrac{df(x_1, x_2)}{dx_1}\right|_{\boldsymbol{x}=(-2,0)} = 0$ $\qquad$ $\left.\dfrac{df(x_1, x_2)}{dx_2}\right|_{\boldsymbol{x}=(-2,0)} = 1$
  $\boldsymbol{x}^{(1)} = \boldsymbol{x}^{(0)} - \gamma(0, 1) = (-2, -1).$

Q4. Gradient Descent

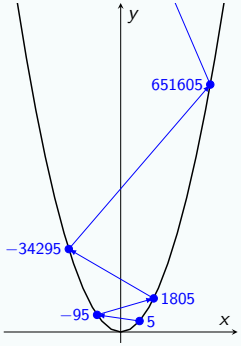
(c) Explain if the gradient descent procedure from Question (b) ever converges to the true minimizer $\boldsymbol{x}^*$? If it does not, how can we fix it?

- It keeps oscillating between $(-2, 0)$ and $(-2, 1)$.
- The learning rate is too large.

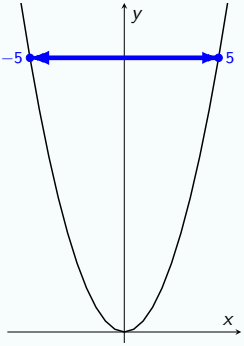
Q4. Gradient Descent

Extra Slide

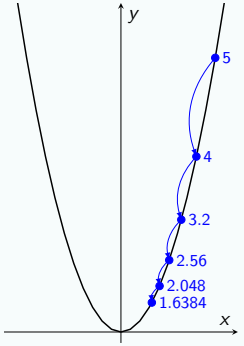
Optimizing $y = x^2$ with different learning rates:



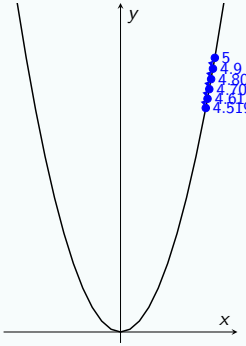
$\alpha = 10$



$\alpha = 1$



$\alpha = 0.1$



$\alpha = 0.01$

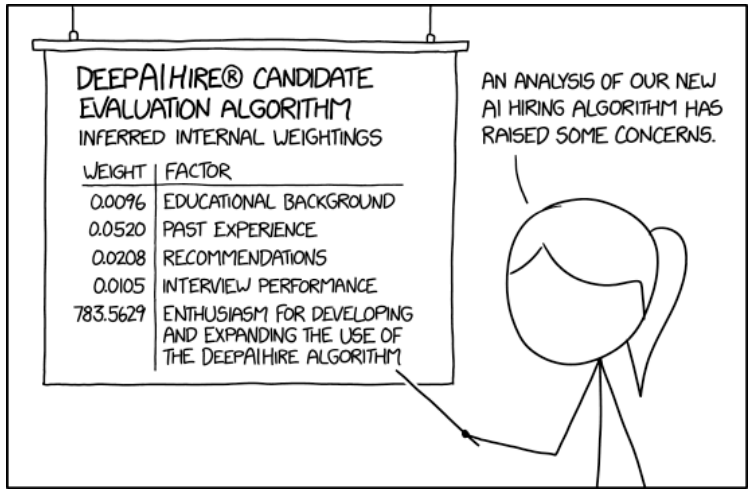
Q4. Gradient Descent

Extra Slide

(d) During the course of training for a large number of epochs/iterations, what can be done to the value of the learning rate α to enable better convergence?

- Large α helps the model to converge faster, but might cause it to overshoot or even diverge.
- Idea: Vary α to help the model "stabilize". But how?
  - **Solution**: Decrease the learning rate α through the course of training.
  - This is the logic behind a **learning rate scheduler**.

That's it!


DEEPAIHIRE® CANDIDATE EVALUATION ALGORITHM
INFERRED INTERNAL WEIGHTINGS
WEIGHT | FACTOR
0.0096 | EDUCATIONAL BACKGROUND
0.0520 | PAST EXPERIENCE
0.0208 | RECOMMENDATIONS
0.0105 | INTERVIEW PERFORMANCE
783.5629 | ENTHUSIASM FOR DEVELOPING AND EXPANDING THE USE OF THE DEEPAIHIRE ALGORITHM
AN ANALYSIS OF OUR NEW AI HIRING ALGORITHM HAS RAISED SOME CONCERNS.