IOWA STATE UNIVERSITY

Department of Computer Science

COM S 573: Machine Learning

Lecture 3: Linear Regression

Data is lin

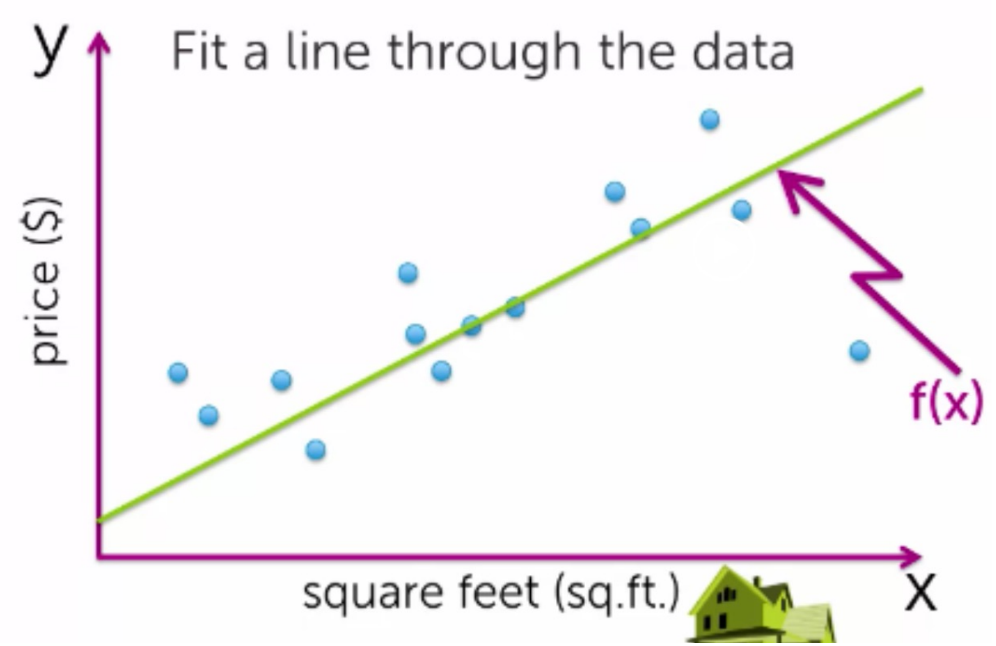
model is lin

Linear Models

- Linear regression: predict a scalar
  - House price
  - Weight of a planet
- Linear perceptron: classifier
  - Predict an animal is a dog or not
  - Predict an image contains a square or not
- Logistic regression: classifier based on a probability
  - Predict how likely a team win
  - Predict how likely tomorrow is sunny

Linear Regression

- Predict a scalar based on input features



What does "**LINEAR**" mean?

Linear Regression: Intuitions

- Given an input like house
- We extract some features from it:
  - For example: [size, #bedrooms, #floors, …]

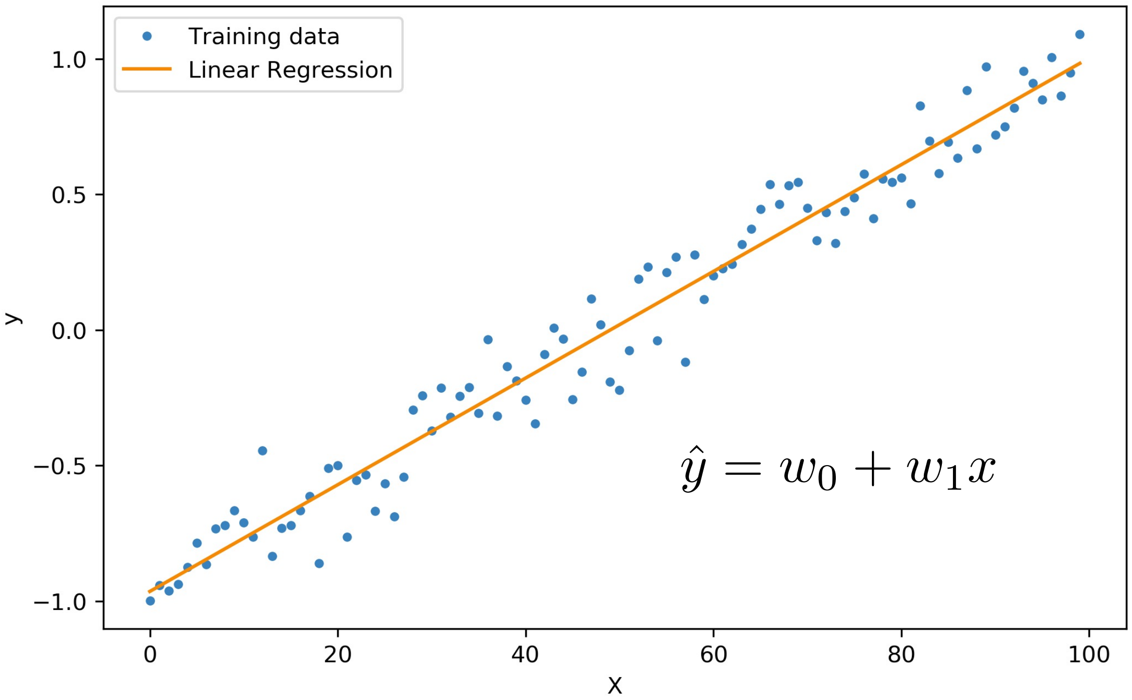
$$\boldsymbol{x} = [x_0, x_1, \cdots, x_d]^T \in \mathbb{R}^d$$

- We want to get an aggregation result by giving these features different weights.

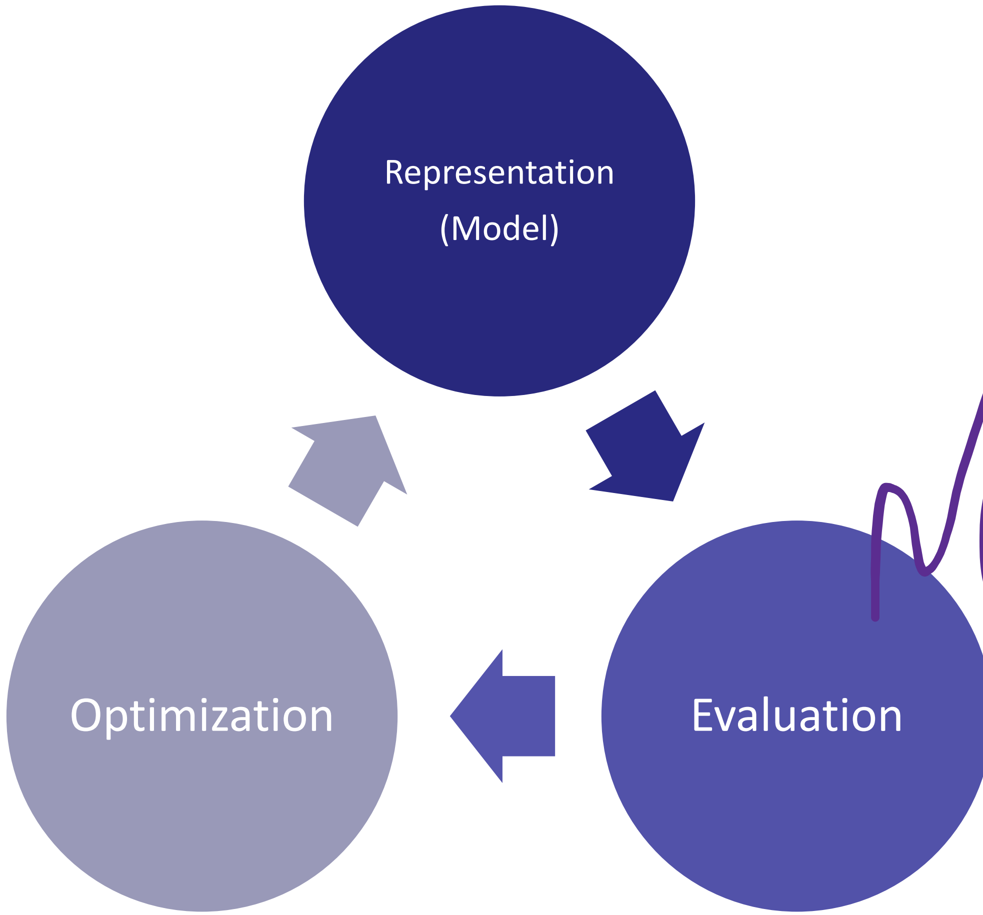
$$y = 1 + 3 * size + 4 * \#bedrooms + 5 * \#floors + ...$$

Linear Regression

- Linear regression is a linear approach to modeling the relationship between a scalar response and one or more independent variables



Three Components of Learning



$$\frac{|y_2 - \hat{y}_2|}{\sqrt{\left((y_1 - \hat{y}_1\right)}}$$

Linear Regression

- Given a training example $< \boldsymbol{x}, y >$
- Each $\boldsymbol{x}$ has d features $x_1, \cdots, x_d$
- The prediction is computed

$$\hat{y}_i = w_0 + w_1 x_{i,1} + w_2 x_{i,2} + \cdots + w_d x_{i,d}$$

Linear Regression: Example

- Predict house price
  - $\boldsymbol{x} = [size, distance]$
  - $y = price$

Training Sample	Size (sq.ft.)	Distance (miles)	Price ($)
1	498	10	600
2	267	9	455
3	399	7.8	546
…	…	…	…

Linear Regression: Representation

- For each training sample $< \boldsymbol{x}_i, y_i >$

- $\hat{y}_i = w_0 + w_1 x_{i,1} + w_2 x_{i,2} + \cdots + w_d x_{i,d}$

$x_i = [1, x_{i1}, x_{i2}, x_{i3} \cdots x_{id}]$

- Suppose $\boldsymbol{w} = [w_0, w_1, \cdots, w_d]^T$

- $\hat{y}_i = \boldsymbol{w}^T \boldsymbol{x}_i$

What is in $\boldsymbol{x}_i$?

$d+1$

$w : d+1$

Linear Regression: Representation

- For each training sample $< \boldsymbol{x}_i, y_i >$
- $\hat{y}_i = w_0 + w_1 x_{i,1} + w_2 x_{i,2} + \cdots + w_d x_{i,d}$
- Suppose $\boldsymbol{w} = [w_0, w_1, \cdots, w_d]^T$
- $\hat{y}_i = \boldsymbol{w}^T \boldsymbol{x}_i$

What is in $\boldsymbol{x}_i$? $\boldsymbol{x_i} = [1, x_1, \cdots, x_d]^T$

Linear Regression: Representation

- For all training sample $< \boldsymbol{x}_1, y_1 >, \cdots, < \boldsymbol{x}_n, y_n >$

$y_i = w^T x$

$= x_i^T w$

- $\hat{y}_i = w_0 + w_1 x_{i,1} + w_2 x_{i,2} + \cdots + w_d x_{i,d}$

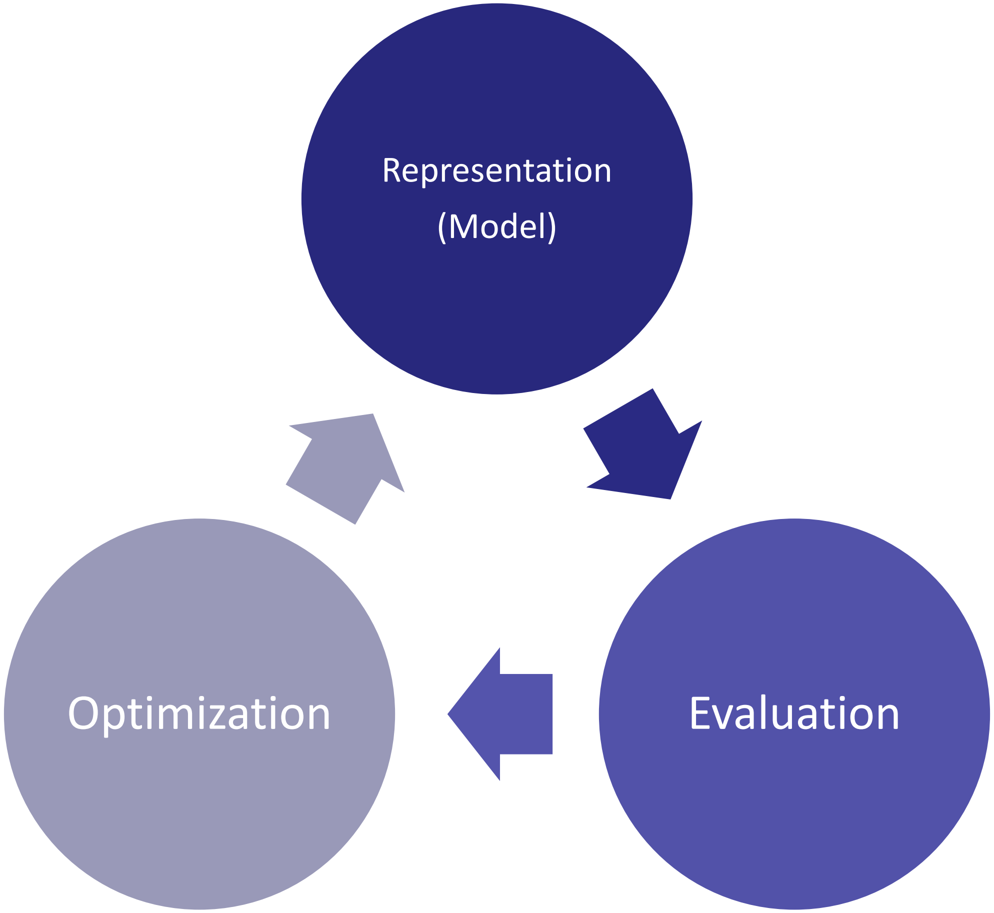
- $$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \cdots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,d} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,d} \\ \cdots & & & & \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,d} \end{bmatrix} \boldsymbol{w}$$

- $\hat{\boldsymbol{y}} = \boldsymbol{X}\underline{\boldsymbol{w}}$ $= n \times 1$

$n \times (d+1)$ $(d \times 1)$

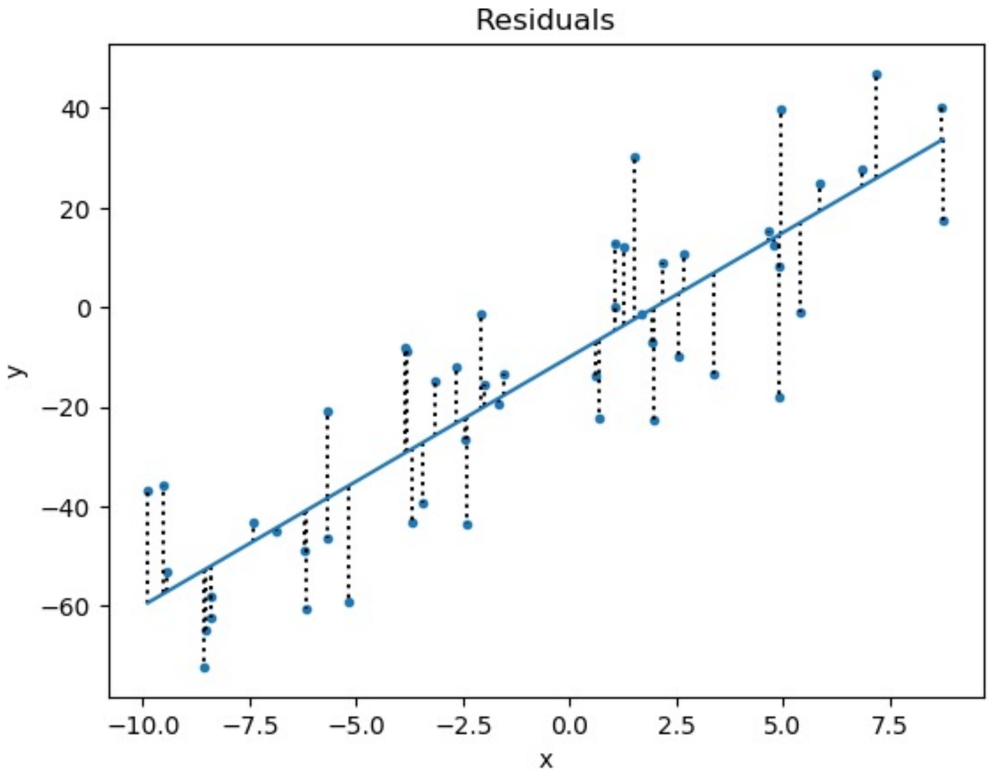
What's the dimension of X?

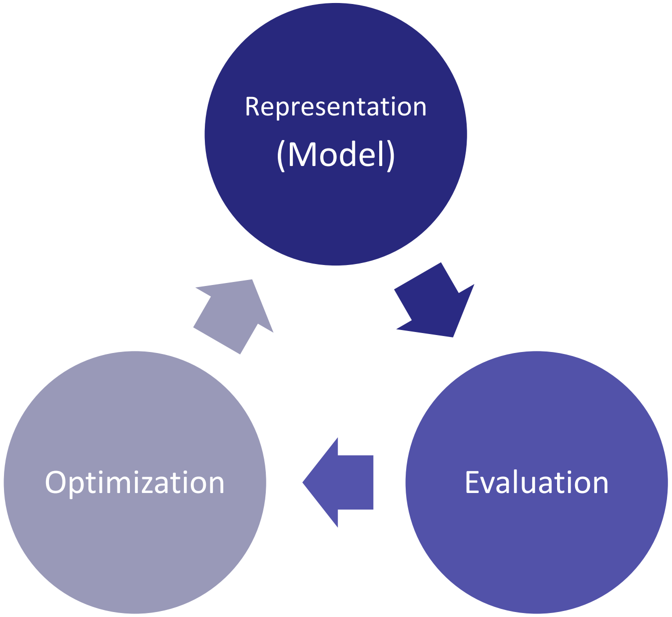
Three Components of Learning



Linear Regression: Evaluation

- Residual Squares $(y_i - \boldsymbol{w}_i^T \boldsymbol{x}_i)^2$





Linear Regression: Evaluation

- Residual Sum of Squares (RSS)

$$RSS(\boldsymbol{w}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- Equivalently

$$RSS(\boldsymbol{w}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})$$

$r_i = y_i - \hat{y}_i$

$r \begin{bmatrix} r_1 \\ \vdots \\ r_n \end{bmatrix}$

$\| y \|_2$

$y^T \cdot y$

$r^T r$

Why?

Linear Regression: Optimization

- $RSS(\boldsymbol{w}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})$
- Find the minimal $RSS(\boldsymbol{w})$

$$\frac{\partial RSS(w)}{\partial w} = 0 \Rightarrow w^*$$

How to find it?

Linear Regression: Optimization

- $RSS(\boldsymbol{w}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})$

- Find the minimal $RSS(\boldsymbol{w})$

- When the first derivative of a function equals zero, the minimum of a function is achieved.

$$\frac{\partial RSS(\boldsymbol{w})}{\partial \boldsymbol{w}} = 0$$

- The optimal $\boldsymbol{w}$ is obtained by solving this equation.

Linear Regression: Optimization

- $$\begin{aligned} RSS(\boldsymbol{w}) &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) \\ &= \boldsymbol{y}^T\boldsymbol{y} - \underline{\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{w}} + \boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} \\ &= \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} \end{aligned}$$

$$1 \times \underbrace{(d+1) \quad (d+1)}\times \underbrace{N \quad \times N} \times 1$$

$$= 1$$

Linear Regression: Optimization

- $$\begin{aligned} RSS(\boldsymbol{w}) &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) \\ &= \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{w} + \boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\boldsymbol{w} \\ &= \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y} + \boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\boldsymbol{w} \end{aligned}$$

- $$\frac{\partial RSS(\boldsymbol{w})}{\partial \boldsymbol{w}} = \ ?$$

Linear Regression: Optimization

- $$\begin{aligned} RSS(\boldsymbol{w}) &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) \\ &= \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{w} + \boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\boldsymbol{w} \\ &= \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y} + \boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\boldsymbol{w} \end{aligned}$$

- $$\frac{\partial RSS(\boldsymbol{w})}{\partial \boldsymbol{w}} = \; ?$$

$\boxed{?}$

Linear Regression: Optimization

- $$\begin{aligned} RSS(\boldsymbol{w}) &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) \\ &= \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{w} + \boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\boldsymbol{w} \\ &= \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y} + \boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\boldsymbol{w} \end{aligned}$$
- $$\frac{\partial RSS(\boldsymbol{w})}{\partial \boldsymbol{w}} = \ ?$$

$\boxed{0}$ $\boxed{?}$

Linear Regression: Optimization

- $$\begin{aligned} RSS(\boldsymbol{w}) &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) \\ &= \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{w} + \boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} \\ &= \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} \end{aligned}$$

- $$\frac{\partial RSS(\boldsymbol{w})}{\partial \boldsymbol{w}} = \; ?$$

$\boxed{0}$ $\qquad$ $\boxed{-2\boldsymbol{X}^T\boldsymbol{y}}$ $\qquad$ $\boxed{?}$ $= X^T X w$

$X^T X w$

$+ (w^T X^T X)^T$

Linear Regression: Optimization

$(X^T X)^{-1} X^T y$

- $$RSS(\boldsymbol{w}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})$$
$$= \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{w} + \boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w}$$
$$= \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w}$$

1

- $$\frac{\partial RSS(\boldsymbol{w})}{\partial \boldsymbol{w}} = \ ?$$

$w^* = \frac{y}{x}$

$$\boxed{0} + \boxed{-2\boldsymbol{X}^T\boldsymbol{y}} + \boxed{2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w}} = 0$$

2

$X^T X w - X^T y = 0$

$\frac{X^T y}{X^T X}$

$$X^T X w - X^T y = 0$$

$$(X^T X)^{-1} X^T X w = X^T y$$

$$w = (X^T X)^{-1} X^T y$$

Linear Regression: Optimization

- $$RSS(\boldsymbol{w}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})$$
$$= \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{w} + \boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w}$$
$$= \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w}$$

- $$\frac{\partial RSS(\boldsymbol{w})}{\partial \boldsymbol{w}} = -2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} = 0$$

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

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

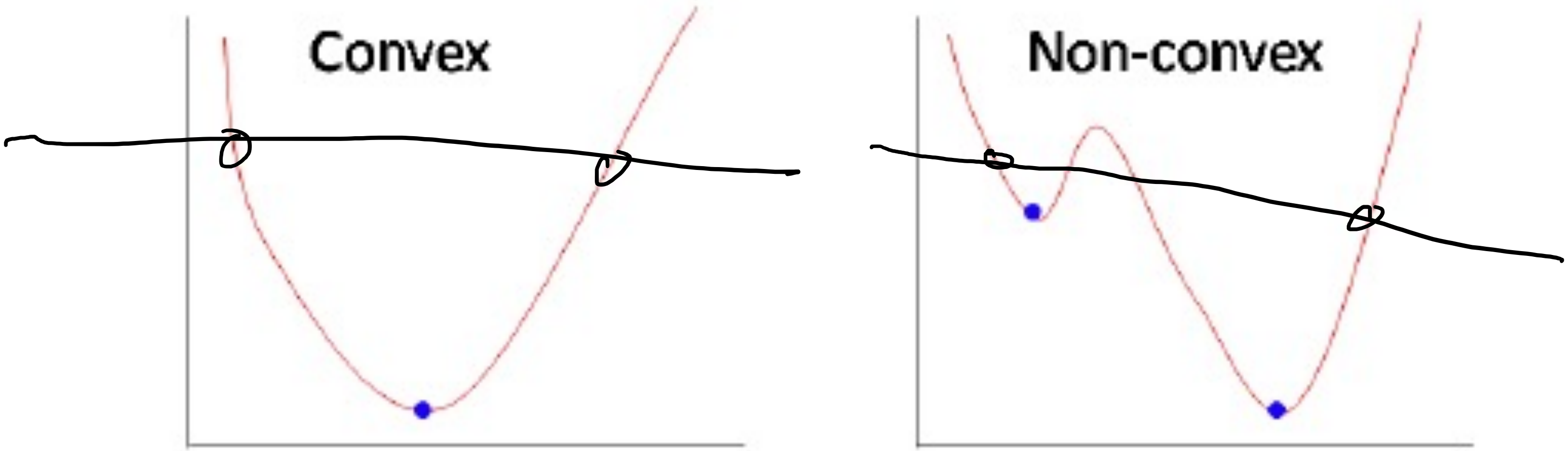
Any protentional issue?

Linear Regression: Questions

- $\boldsymbol{w}^*$ are global optima?

Linear Regression: Questions

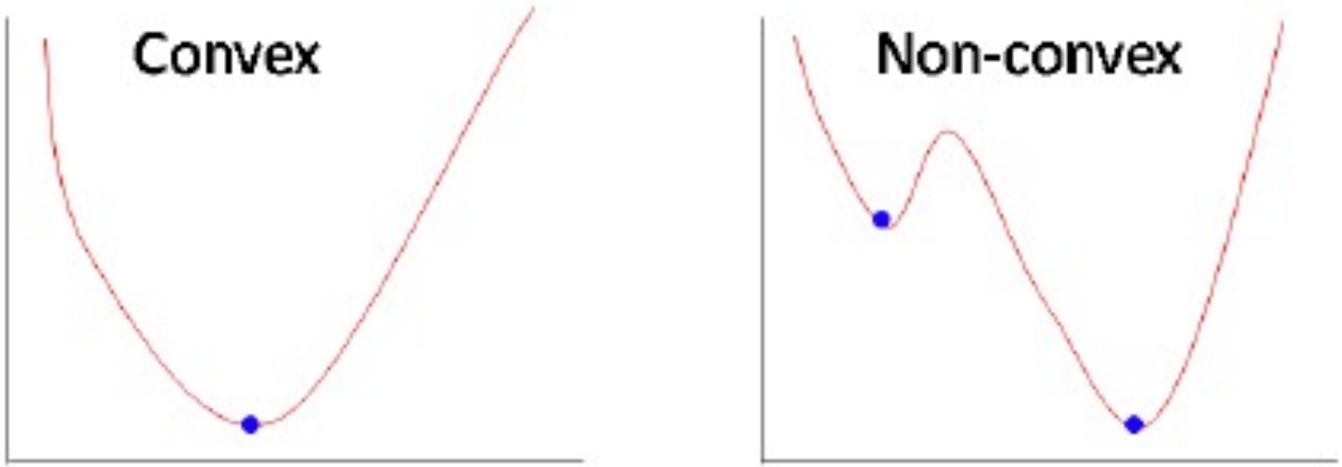
- $\boldsymbol{w}^*$ are global optima? Yes
- $RSS(\boldsymbol{w})$ is a convex function



- What is convex?

Linear Regression: Questions

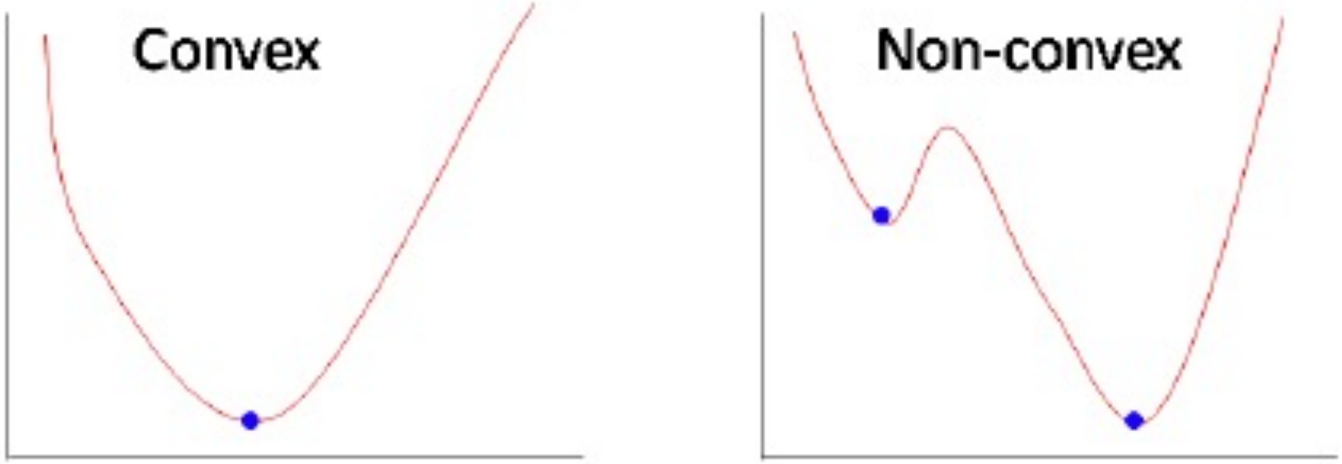
- $\boldsymbol{w}^*$ are global optima? Yes
- $RSS(\boldsymbol{w})$ is a convex function



- What is convex?
- Convex is a property that a line joining any two points on its graph lies on or above the graph.

Linear Regression: Questions

- $\boldsymbol{w}^*$ are global optima? Yes
- $RSS(\boldsymbol{w})$ is a convex function



- What is convex?
- How to prove a function is convex?

Linear Regression: Questions

- $\boldsymbol{w}^*$ are global optima? Yes
- $RSS(\boldsymbol{w})$ is a convex function

$$H(\boldsymbol{w}) = \frac{\partial^2 RSS(\boldsymbol{w})}{\partial \boldsymbol{w}^2}$$

- For every $\boldsymbol{u} \in \mathbb{R}^d$, we have

$$\boldsymbol{u}^T H(\boldsymbol{w}) \boldsymbol{u} >= 0$$

Linear Regression: Summary

- Representation

$$\hat{y}_i = \boldsymbol{w}\boldsymbol{x}_i^T$$

Predict a continuous scalar.

- Evaluation

$$RSS(\boldsymbol{w}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})$$

- Optimization

$$\frac{\partial RSS(\boldsymbol{w})}{\partial \boldsymbol{w}} = 0 \rightarrow \boldsymbol{w}^* = (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y}$$

Linear Models: Next

- Linear regression: predict a scalar
  - House price
  - Weight of a planet
- Linear perceptron: classifier of discrete prediction
  - Predict an animal is a dog or not
  - Predict an image contains a square or not
- Logistic regression: classifier based on a probability
  - Predict how likely a team win
  - Predict how likely tomorrow is sunny