L6: Linear Regression

Prof. Xun Jiao

Before class

- HW1 due by this Wed class
- HW2 will be posted on Wed

Outline

- Build a supervised learning model: a high-level perspective
- Linear Regression

Review

```
                                                    Linear
                                                    Regression
                              regression
           supervised
           learning
                              classification
Machine
learning

           unsupervised
           learning
```

Linear Regression

……

regression

supervised learning

classification

Machine learning

unsupervised learning

Review on Classification Evaluation

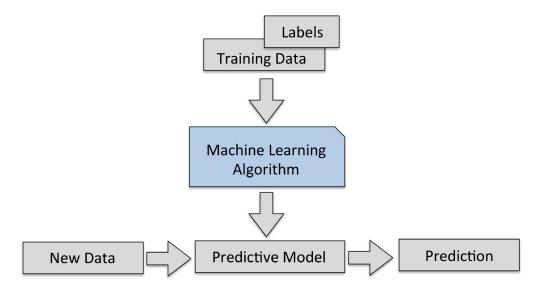|		True/Actual	
		Positive (🐶)	Negative
Predicted	Positive (🐶)	5	1
	Negative	2	2

Accuracy?
Precision?
Recall?
F1 score?

Now that we know how to evaluate a model, let us see how to build a model (supervised learning).

Supervised learning is the process of trying to infer from labeled data the underlying function that produced the labels associated with the data

- Predicting the future with supervised learning
- Classification vs. Regression

Given **labeled training data** of the form
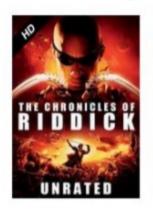
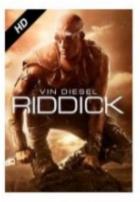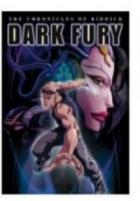$$\{(\text{data}_1, \text{label}_1), \ldots, (\text{data}_n, \text{label}_n)\}$$

Infer the function

$$f(\text{data}) \xrightarrow{?} \text{labels}$$

Suppose we want to build a movie recommender

e.g. which of these films will I rate highest?

Q: What are the **labels?**

A: **ratings** that others have given to each movie, and that I have given to other movies

Q: What is the **data?**

A: **features** about the movie and the users who evaluated it

Movie features: genre, actors, rating, length, etc.

User features: age, gender, location, etc.

Product Details

Genres	Science Fiction, Action, Horror
Director	David Twohy
Starring	Vin Diesel, Radha Mitchell
Supporting actors	Cole Hauser, Keith David, Lewis Fitz-Gerald, Claudia Black, Rhiana Gr Angela Moore, Peter Chiang, Ken Twohy
Studio	NBC Universal
MPAA rating	R (Restricted)
Captions and subtitles	English Details ▾
Rental rights	24 hour viewing period. Details ▾
Purchase rights	Stream instantly and download to 2 locations Details ▾
Format	Amazon Instant Video (streaming online video and digital download)

A. Phillips

Reviewer ranking: #17,230,554

90% helpful
votes received on reviews
(151 of 167)

ABOUT ME
Enjoy the reviews…

ACTIVITIES
Reviews (16)
Public Wish List (2)
Listmania Lists (2)
Tagged Items (1)

What is data? What is label?

No.	Model	Luxury?	Year	MPG	Horsepower	Price
1	Acura MDX	Yes	2017	20	290	$50,000
2	Honda Accord	No	2017	25	190	$25,000
3	Honda Civic	No	2012	23	160	$10,000
4	Honda Civic	No	2016	24	170	$18,000
5	Nissan Altima	No	2016	30	180	$25,000
6	Acura MDX	Yes	2015	18	280	$38,000
7	Lexus RX350	Yes	2015	21	270	$40,000
8	Toyota Prius	No	2014	45	120	$28,000
9	Toyota Prius	No	2013	40	120	$24,000
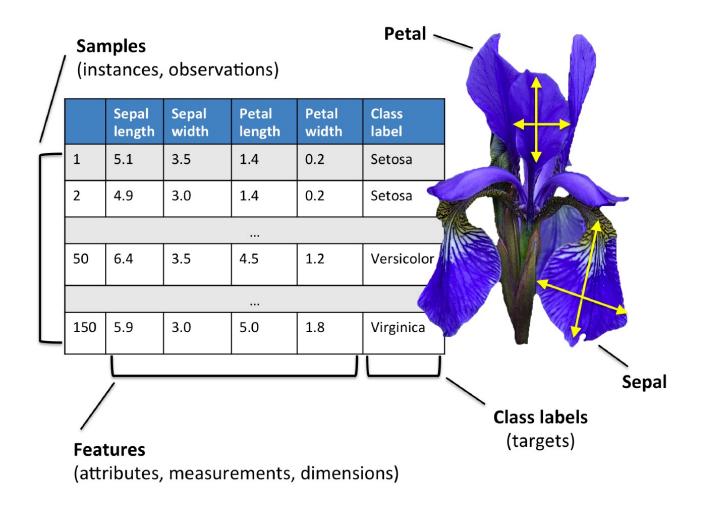
Movie recommendation:

$$f(\text{data}) \xrightarrow{?} \text{labels}$$

$$=$$

$$f(\text{user features}, \text{movie features}) \xrightarrow{?} \text{star rating}$$
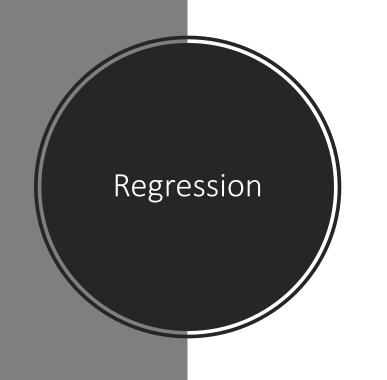
A classic example: iris dataset

Samples
(instances, observations)

	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
…					
50	6.4	3.5	4.5	1.2	Versicolor
…					
150	5.9	3.0	5.0	1.8	Virginica

Petal

Sepal

Class labels
(targets)

Features
(attributes, measurements, dimensions)

Basic Terminology
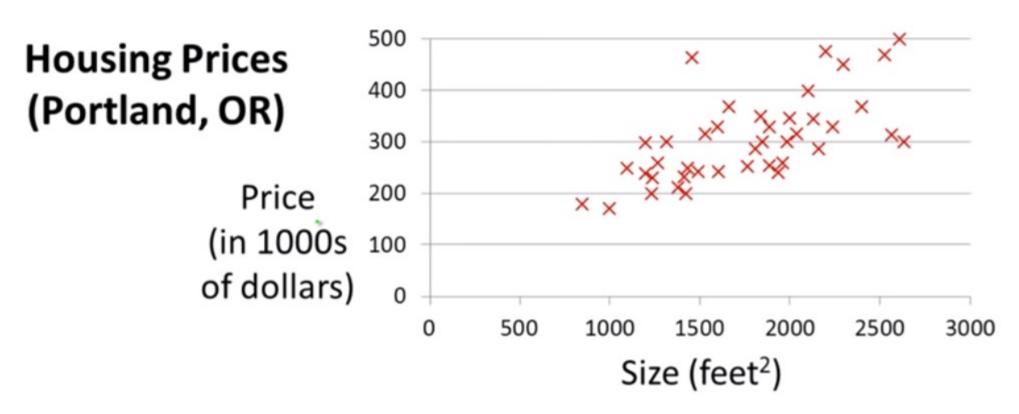
- Measurements of 150 iris flowers (150 samples / 4 features)
- From 3 different species (Setosa, Versicolor, Virginica)
- Rows are samples and columns are features
- 150×4 matrix $\mathbf{X} \in \mathbb{R}^{150 \times 4}$:

$$
\begin{bmatrix}
x_1^{(1)} & x_2^{(1)} & x_3^{(1)} & \cdots & x_4^{(1)} \\
x_1^{(2)} & x_2^{(2)} & x_3^{(2)} & \cdots & x_4^{(2)} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
x_1^{(150)} & x_2^{(150)} & x_3^{(150)} & \cdots & x_4^{(150)}
\end{bmatrix}
$$

Regression
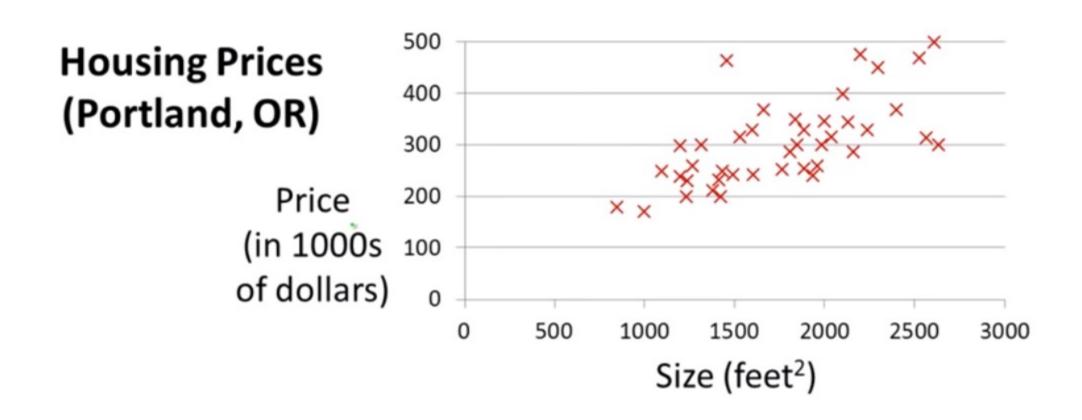
Regression is one of the simplest supervised learning approaches to learn relationships between input variables (features) and output variables (predictions)

A regression example

Housing Prices (Portland, OR)

Price (in 1000s of dollars) — Size (feet^2)
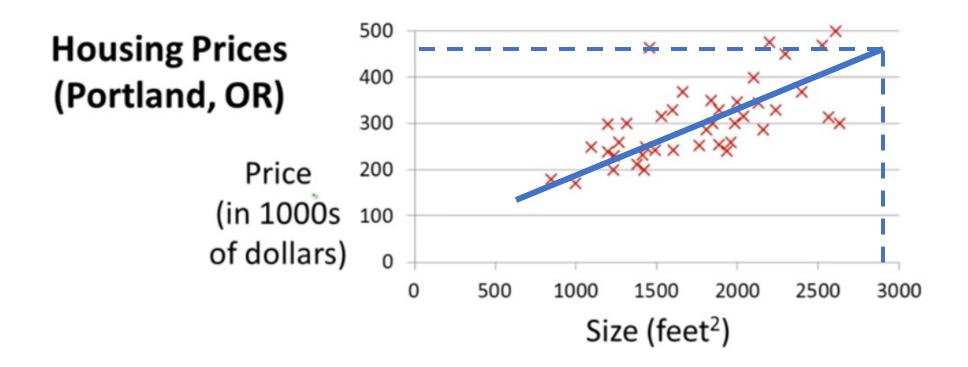
Suppose I want to build a predictive model that can predict "Price" based on "Size"
What is my input? Output?

So, how can we find such model/function?

Housing Prices (Portland, OR)

Price (in 1000s of dollars)

Size (feet^2)
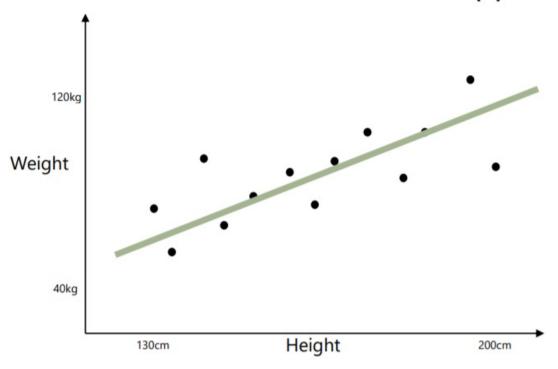
The easiest assumption

- The easiest assumption would be a linear relationship (mapping) from data to label.

Housing Prices (Portland, OR)

Price (in 1000s of dollars)

Size (feet^2)

Motivation: height vs. weight

Q: Can we find a line that (approximately) fits the data?

Motivation: height vs. weight

Q: Can we find a line that (approximately) fits the data?

- If we can find such a line, we can use it to make **predictions** (i.e., estimate a person's weight given their height)
  - How do we **formulate** the problem of finding a line?
- If no line will fit the data exactly, how to **approximate?**
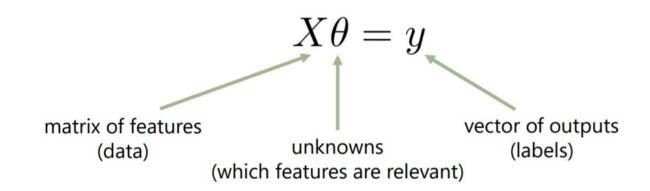 - What is the "best" line?

Goal of regression

- Given a set of training data, find the mapping function *f that can*
  - *Fit the training data*
  - *Map from X -> y*
    - *Input data → label*

To answer these questions, we need to look into the general mathematical form of linear relationship.
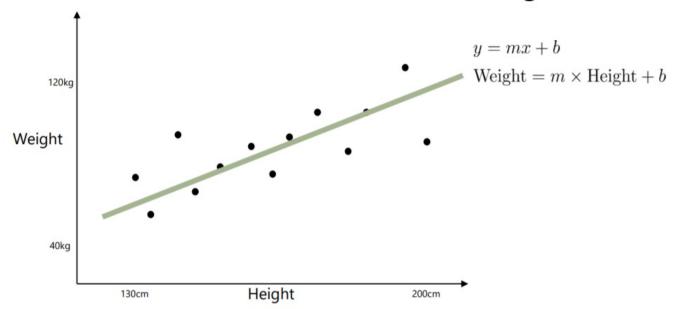
Linear regression assumes a predictor of the form

$$X\theta = y$$

matrix of features (data)

unknowns (which features are relevant)

vector of outputs (labels)

(or $Ax = b$ if you prefer)

Linear Regression

Recap: equation for a line

What about in more dimensions?

$$\text{Weight} = m_1 \times \text{Height} + m_2 \times \text{age} + b$$

$$y = m_1 x_1 + m_2 x_2 + b$$

Weight

120kg

40kg

100 years

Age

1 year

130cm   Height   200cm

Recap: equation for a line as an inner product

What about in more dimensions?

$$\text{Weight} = (\text{Height}, \text{age}, 1) \cdot (m_1, m_2, b)$$

$$y = (x_1, x_2, 1) \cdot (m_1, m_2, b)$$

Linear Regression

Linear regression assumes a predictor of the form

$$X\theta = y$$

Q: Solve for theta