Programming for Data Analytics
CISA4313

Linear Regression

Dr. Mohammad Abdel-Rahman

Data mining problems

Prediction (Supervised learning)

- Classification
- Regression

Pattern discovery (Unsupervised learning)

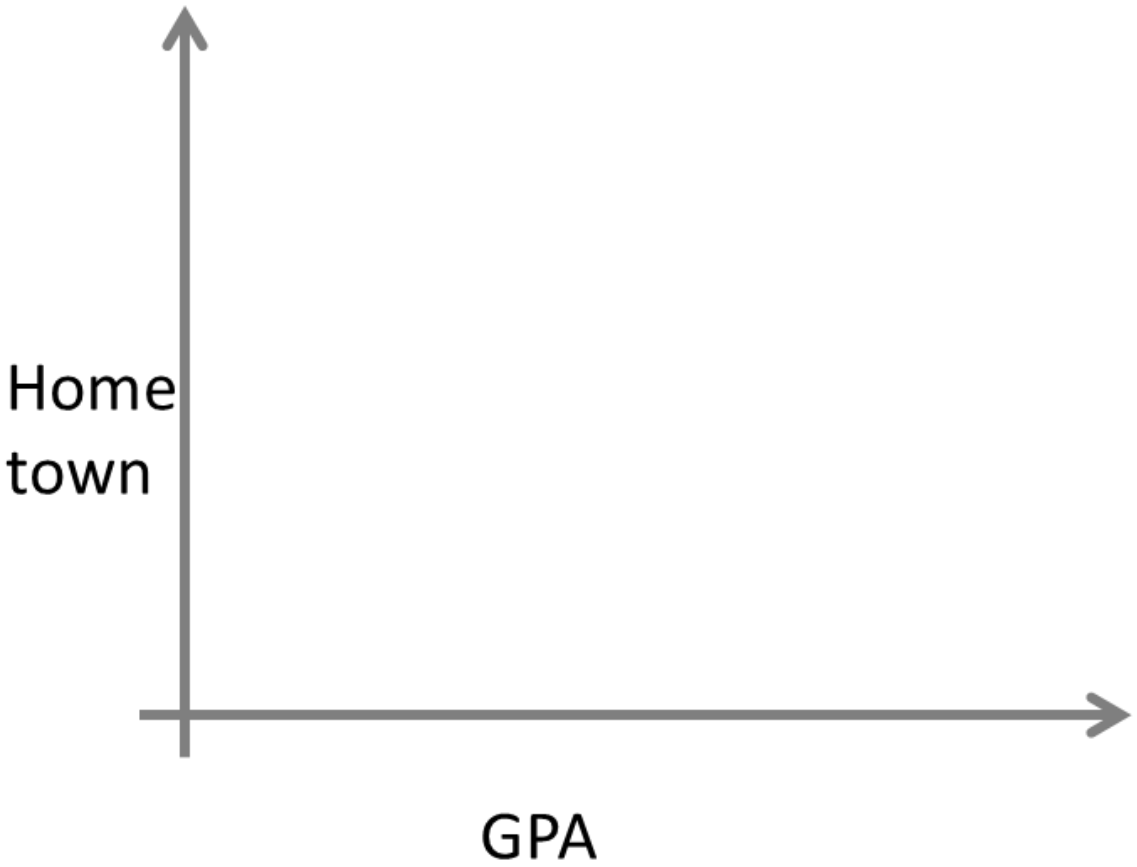
- Clustering

Classification

- Classification

- Supervised learning
- The variable of interest (i.e., output or dependent variable) is categorical.

- Case: Drop out(Y, N)

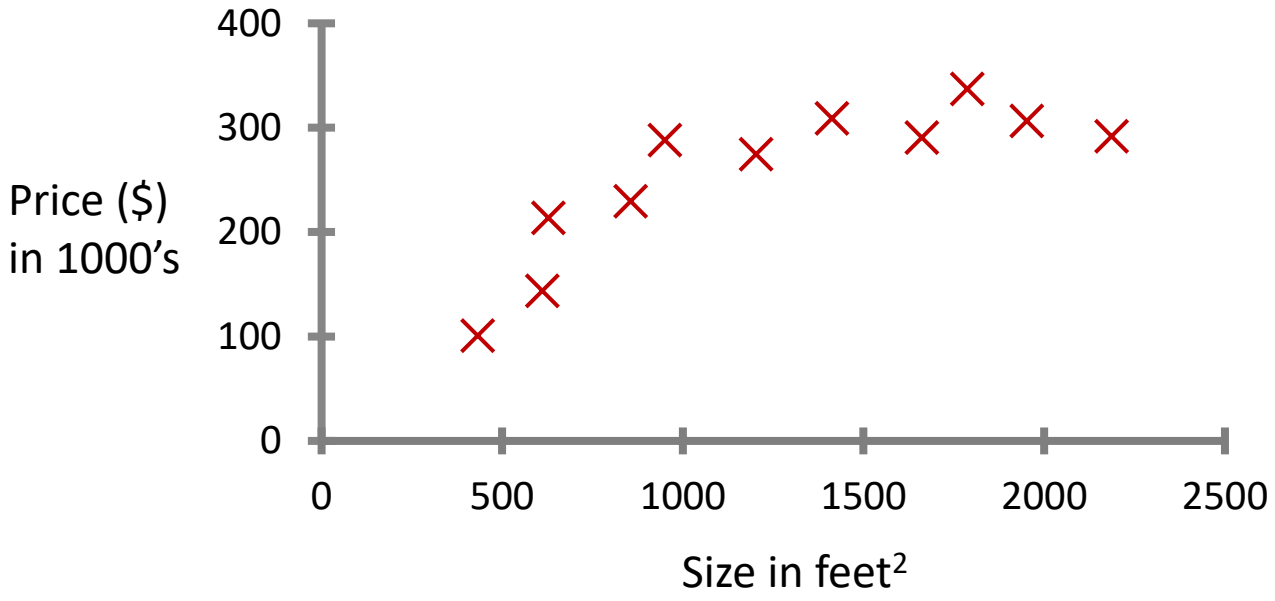


Regression

- Regression

  - Supervised learning (“right answers” given)
  - The variable of interest (output or dependent variable) is continuous.

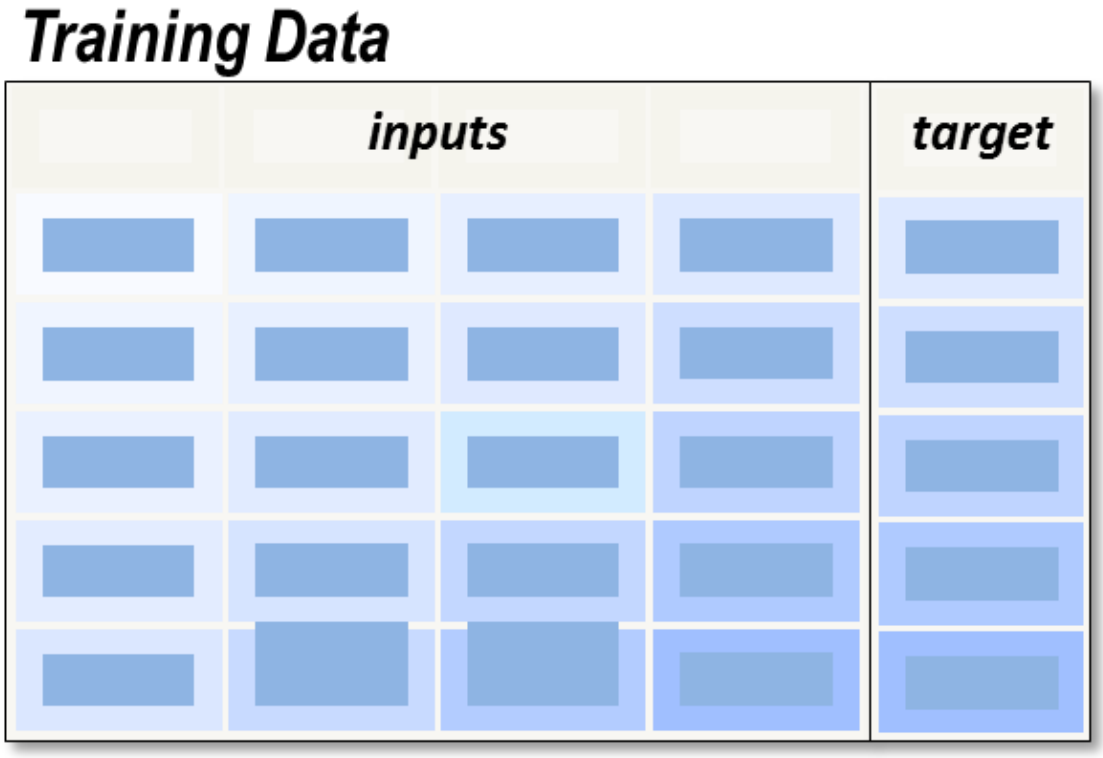
- Case: House price prediction.



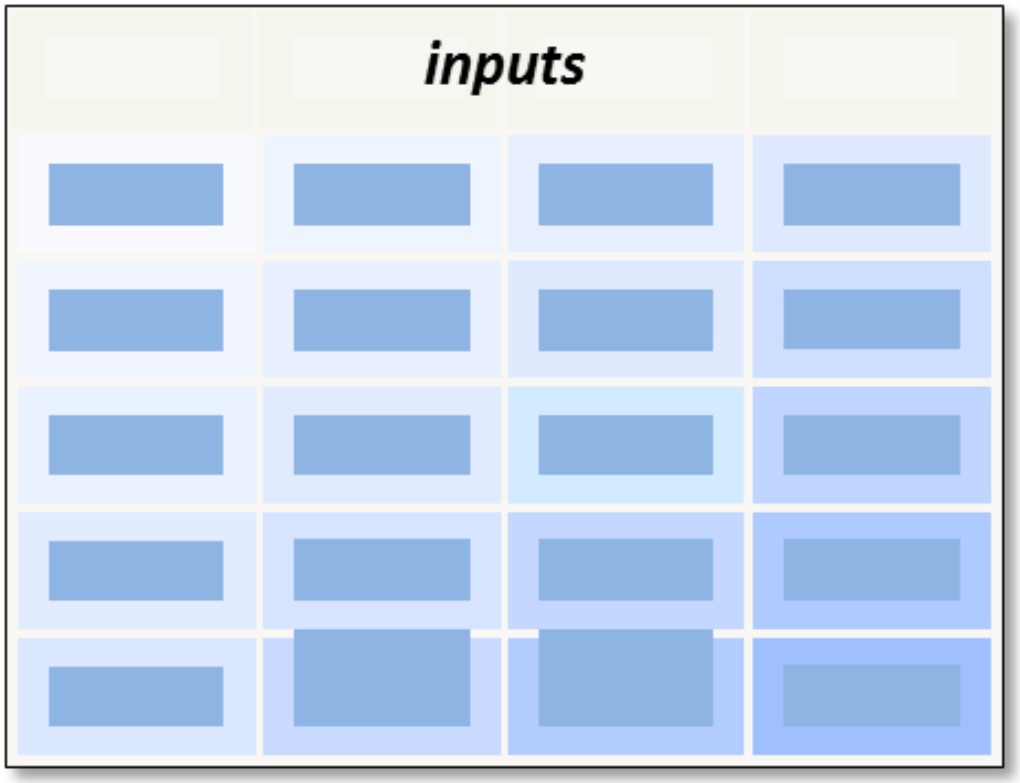
Summary of Supervised Learning

- Goal: To infer a function from *supervised* (labeled) training data.
- The training data consist of a set of *training examples*.
- In supervised learning, each example includes some input features values, and a desired output value
- A supervised learning algorithm analyzes the training data and produces an inferred function, which is called a *classifier* (if the output is discrete) or a *regression function* (if the output is continuous). The inferred function should predict the correct output value for any valid input object.

Supervised learning



Unsupervised learning



Exercise

You're running a company, and you want to develop learning algorithms to address each of two problems.

Problem 1: You have a large inventory of identical items. You want to predict how many of these items will sell over the next 3 months.
Problem 2: You'd like software to examine individual customer accounts, and for each account decide if it has been hacked/compromised.

Should you treat these as classification or as regression problems?

Treat both as classification problems.

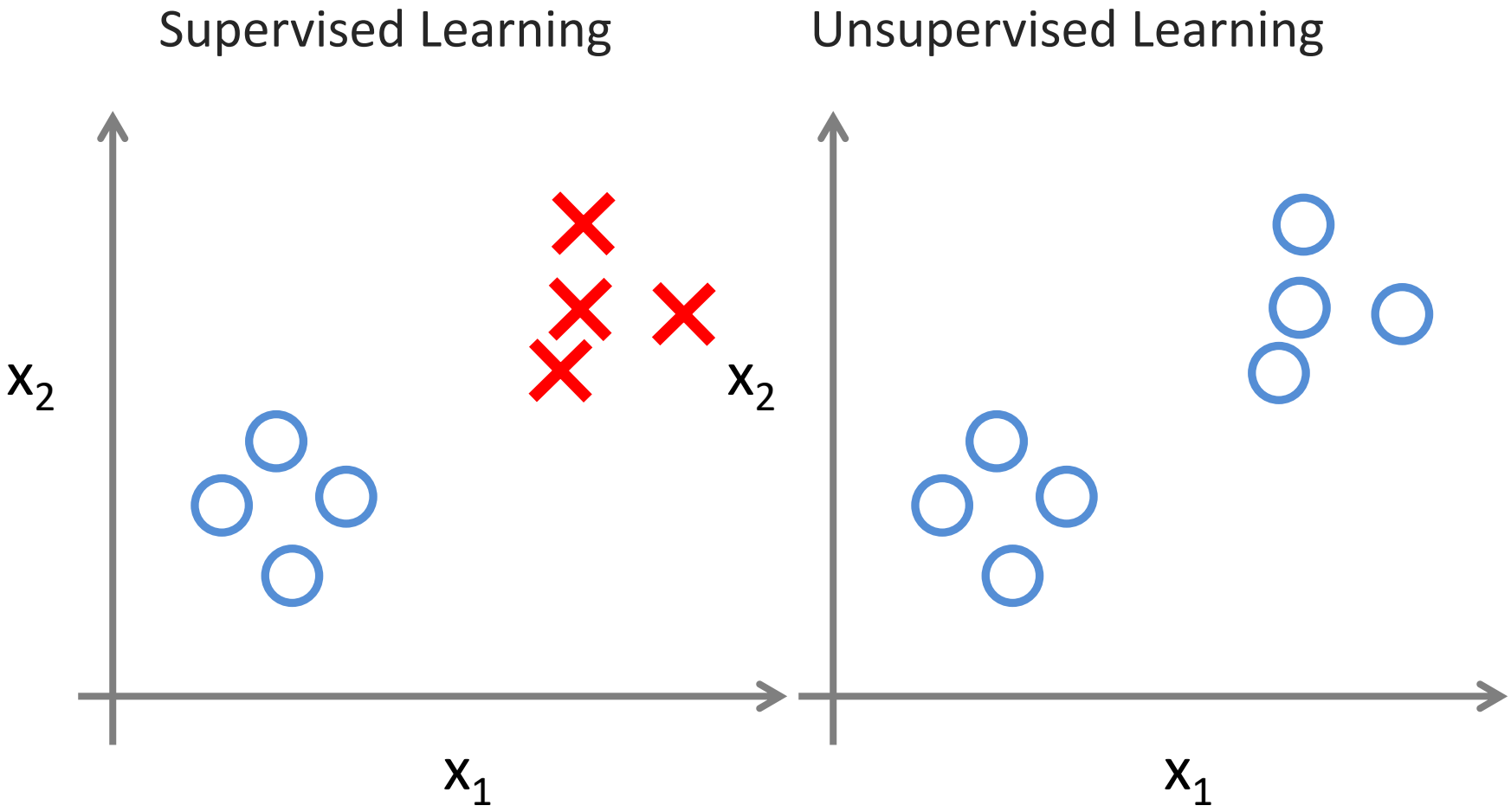
Treat problem 1 as a classification problem, problem 2 as a regression problem.

Treat problem 1 as a regression problem, problem 2 as a classification problem.

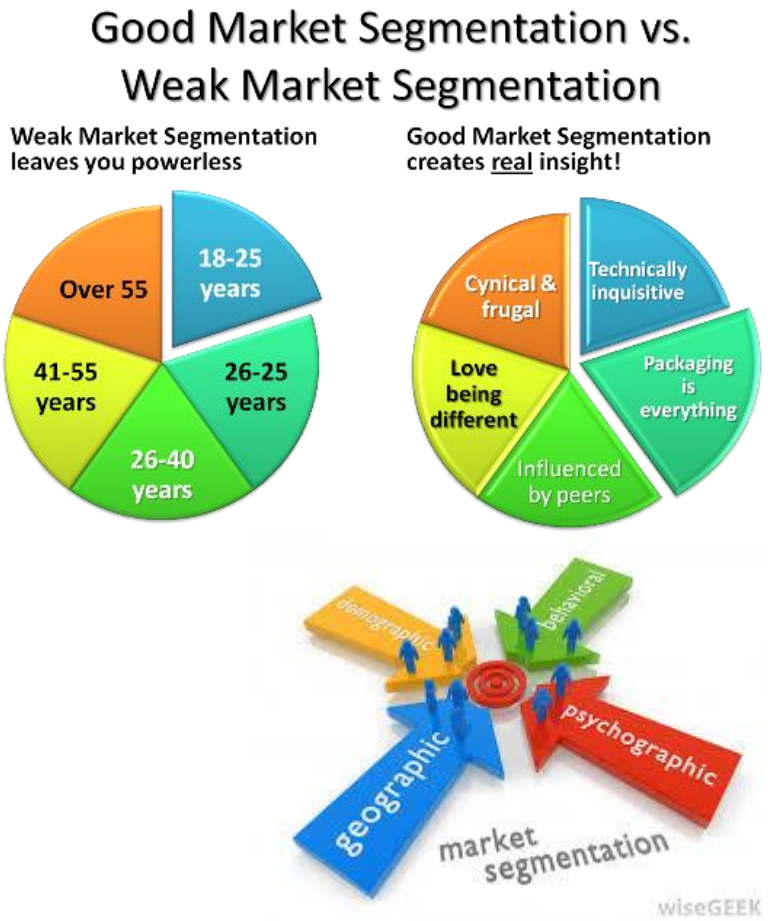
Treat both as regression problems.

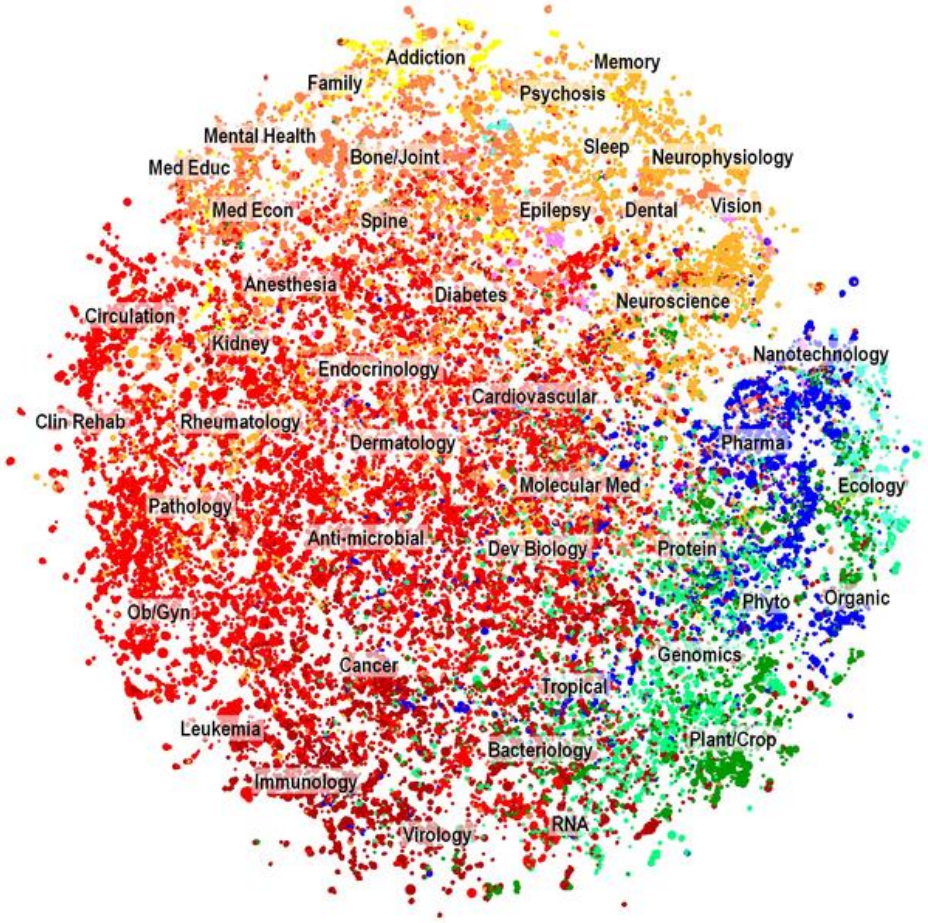
Clustering

- Unsupervised learning (e.g., the examples are unlabeled)
- E.g., Clustering of customers, symptoms, product brands , etc.



Examples of clustering

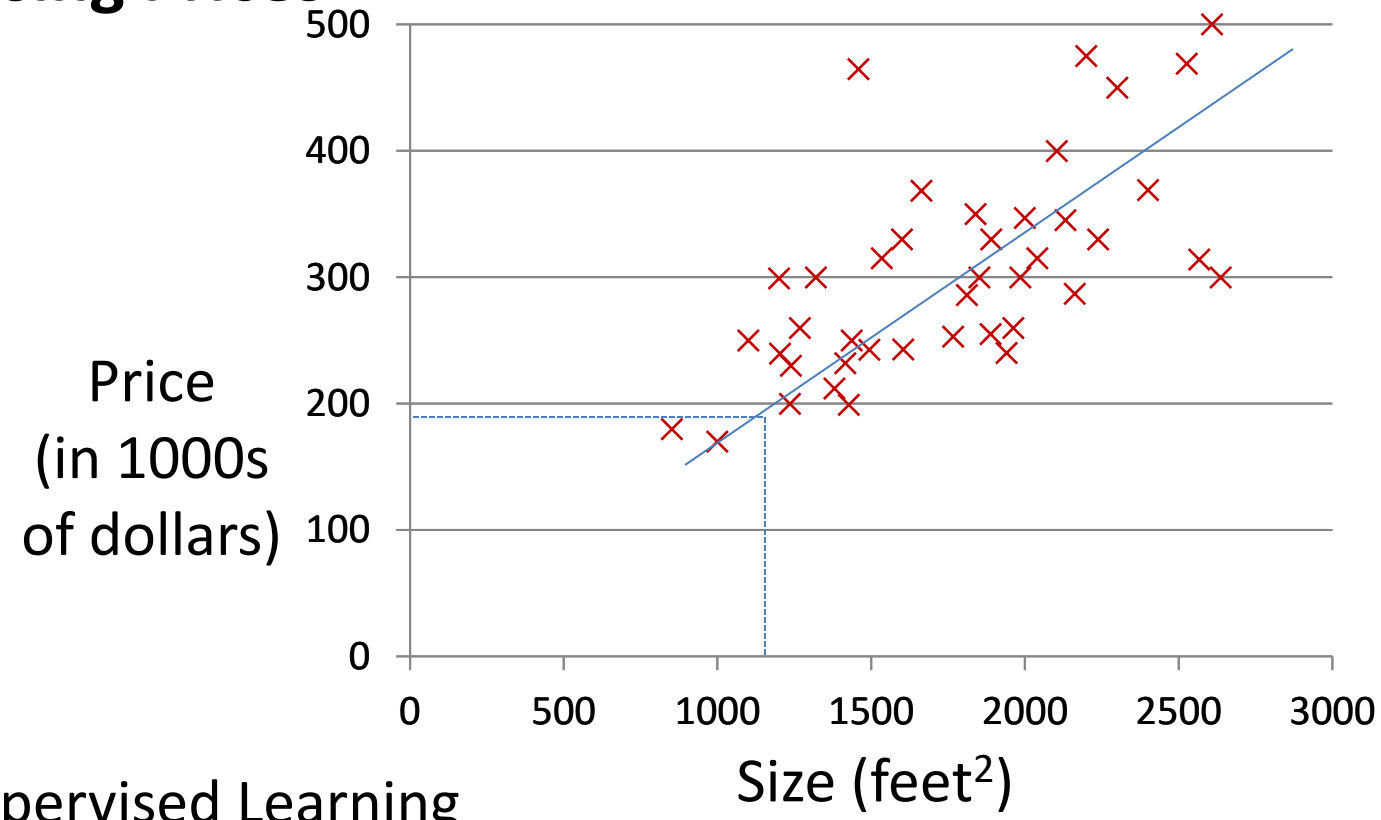




Linear regression

Linear regression

Housing Prices



Supervised Learning

Given the "right answer" for each example in the data.

Regression Problem

Predict real-valued output

Notation

Training set of housing prices

Size in feet² (x)	Price ($) in 1000's (y)
2104	460
1416	232
1534	315
852	178
…	…

Notation:

m = Number of training examples

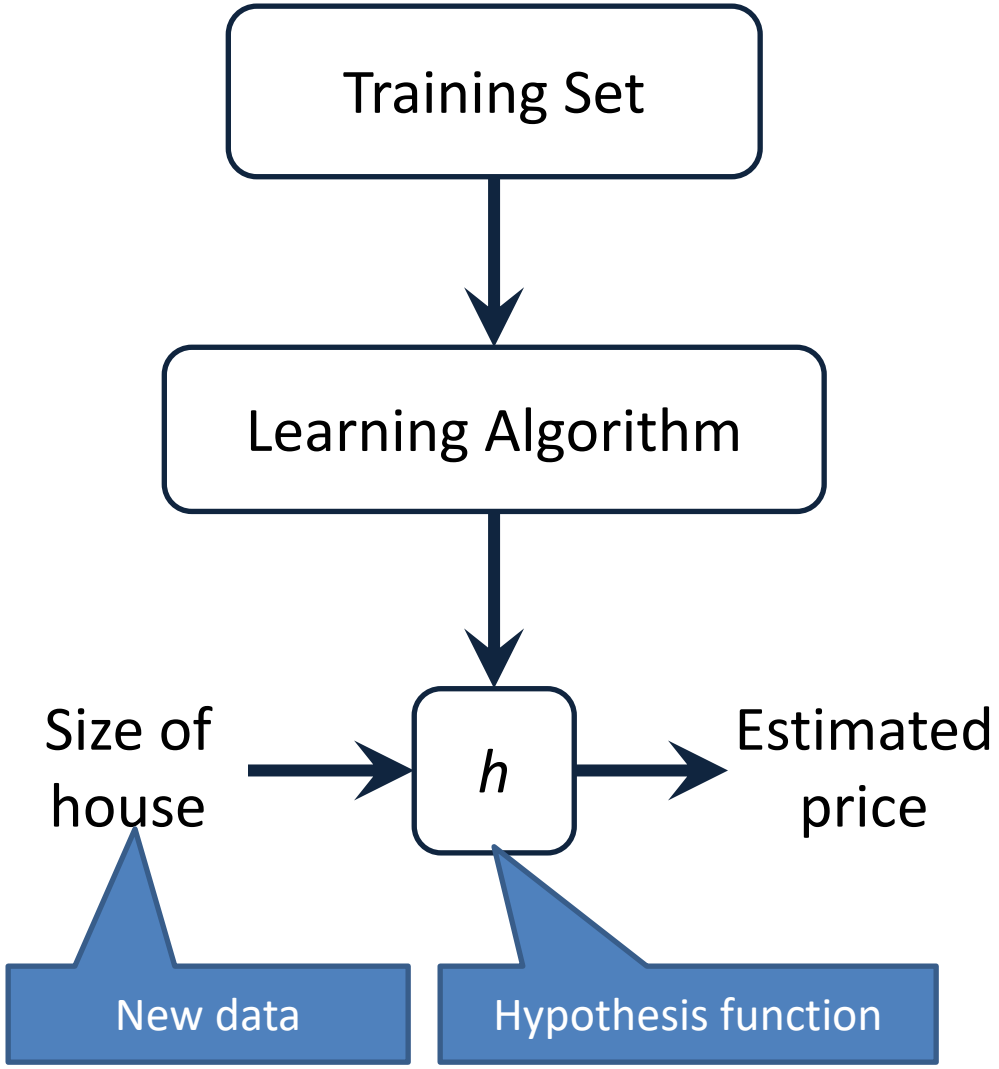
x's = "input" variable / features. x can be written as x_1 .. x_n

y's = "output" variable / "target" variable

(x, y) represents a training example

($x^{(i)}$, $y^{(i)}$) represent *i*th training example E.g., $x^{(i)} = 2104$, $y^{(i)} = 460$

How a supervised learning algorithm works



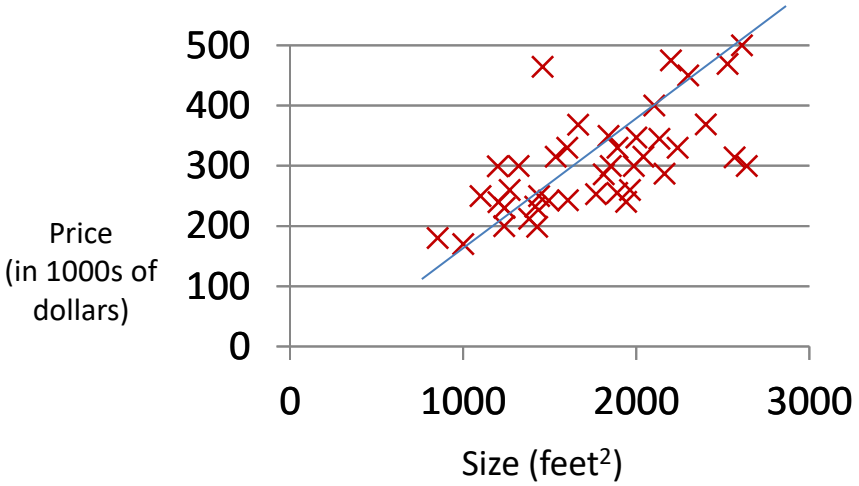
How do we represent *h* ?

Training Set

m = 47

Size in feet² (x)	Price ($) in 1000's (y)
2104	460
1416	232
1534	315
852	178
…	…

Linear regression with one variable.
Univariate linear regression.



$$h_\theta(x) = \theta_0 + \theta_1 x$$

E.g. $h_\Theta(x) = -42.8 + 0.23\ x$

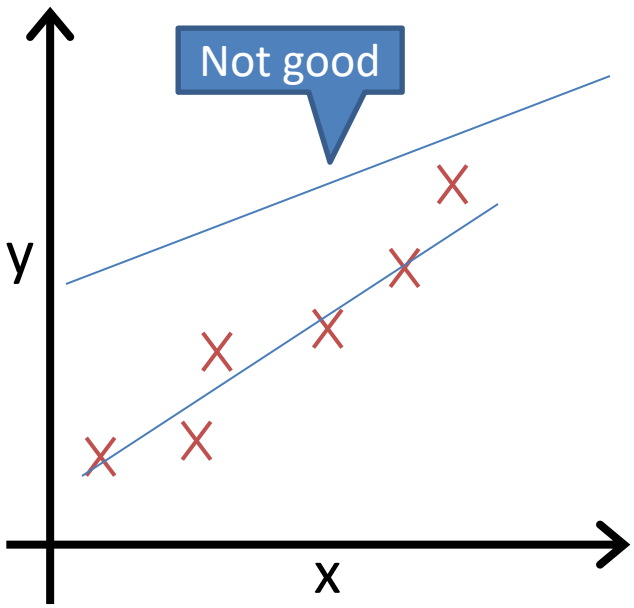
We can have multiple features.

$$h_\Theta(x) = \theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n$$

Θ: parameters

How can we find parameter Θ that corresponds to a good fit to the training data?

How to find the best Θ?



Our goal:

$$\underset{\theta_0,\theta_1}{Minimize}\ \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right)-y^{(i)}\right)^2$$

m: number of training examples

Sum of squared error

Squared error

Idea: Choose Θ so that $h_\Theta(x)$ is close to y for our training examples

Evaluation measure for linear regression

- The RMSE and R-2 metrics, two metrics commonly used to evaluate linear regression
- RMSE (Root Mean Squared Error)
  - the lower that value is, the better the fit
- R-squared: (The closer towards 1, the better the fit)

$$R^2 = \frac{SSR}{SST} = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}$$