Foundations of Data Science

DS 3001

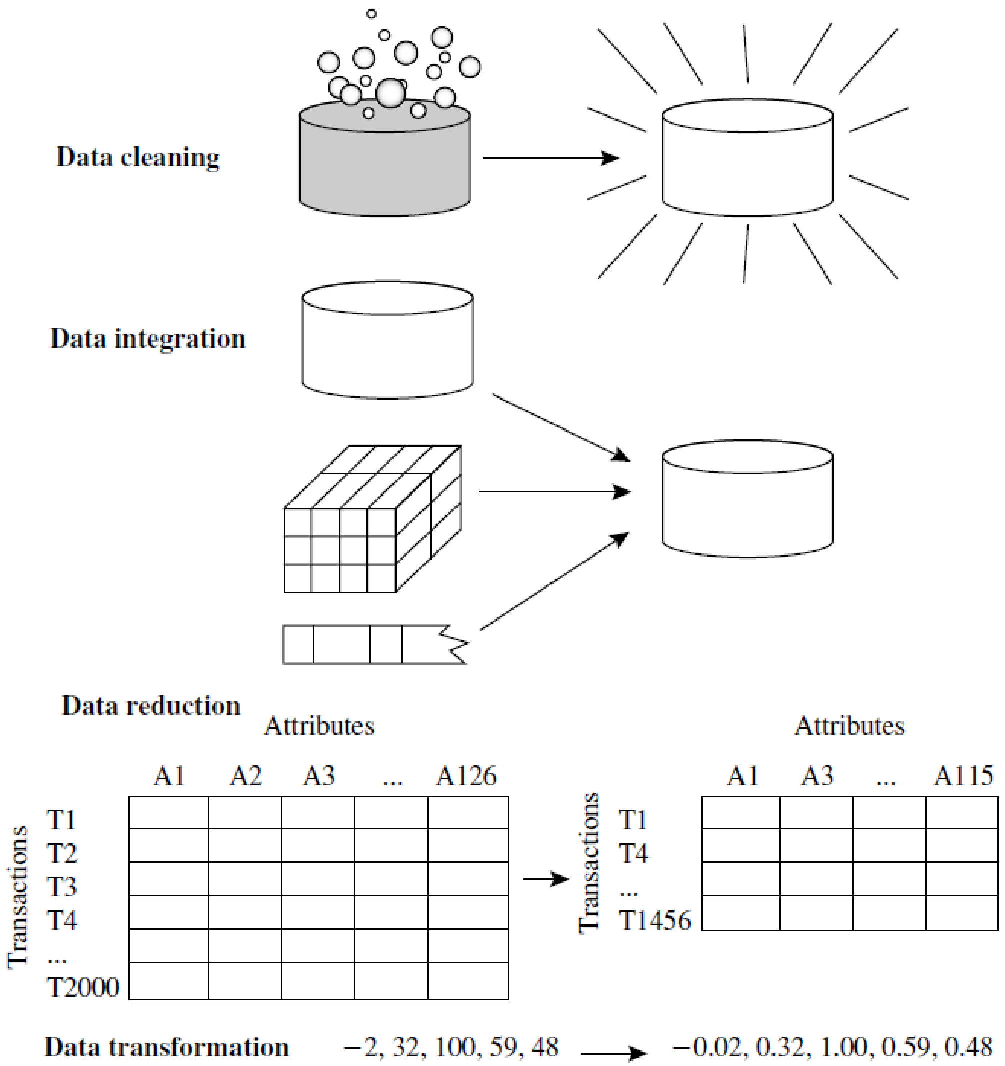
Data Science Program
Department of Computer Science
Worcester Polytechnic Institute
Instructor: Prof. Kyumin Lee

Project Teams

1. Clay Oshiro-Leavitt, Hunter Caouette, Nick Alescio
2. Danielle Angelini, Elijah Ellis, Ryan Candy, Rob Wondolowski
3. Eva (Yingbing) Lu, Manasi Danke, Erica Lee, Jonathan Dang
4. Arianna Kan, Yihan Lin, Margaret Goodwin, Ken Snoddy
5. Yang Gao, Jose Li, Sarah Burns, Daniel McDonough
6. Noah Puchovsky, Katherine Handy, Alex Tavares, Angelica Puchovsky
7. Armando Zubillaga, Gabriel Rodrigues, Humberto Leon, Joao Omena de Lucena
8. Edward Carlson, Samuel Goldman, Nick Krichevsky, Christopher Myers
9. Jessie White, Lindsay MacInnis, Bao Huynh, Ziqian Zeng
10. Suverino Frith, Nicholas Odell, Fay Whittall, Johvanni Perez
11. Alp Piskin, Robert Scalfani, Jake Barefoot, Mark Bernardo
12. Amanda Chan, Nugzar Chkhaidze, Luke Gebler
13. Daniel Pelaez, Nathan Savard, Kate Sincaglia

- So far, 49 students expressed their preferences

Forms of Data Preprocessing



Data Integration

Data Integration

- **Data integration**:
  - Combines data from multiple sources into a coherent store
- Handling Redundancy in Data Integration
  - Redundant data occur often when integration of multiple databases
    - *Object identification*: The same attribute or object may have different names in different databases
    - *Derivable data:* One attribute may be a “derived” attribute in another table, e.g., annual revenue
  - Redundant attributes may be able to be detected by correlation analysis and covariance analysis
- How to find redundant attributes or almost duplicate attributes?

Correlation Analysis (Nominal Data)

- X^2 (chi-square) test

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

- The larger the X^2 value, the more likely the variables are related

- The cells that contribute the most to the X^2 value are those whose actual count is very different from the expected count

- Expected frequency of (A_i, B_j), which can be calculated as

$$e_{ij} = \frac{count(A = a_i) \times count(B = b_j)}{n},$$

- Correlation does not imply causality
  - # of hospitals and # of car-theft in a city are correlated
  - Both are causally linked to the third variable: population

Chi-Square Calculation: An Example

	Play chess	Not play chess	Sum (row)
Like science fiction	250(90)	200(360)	450
Not like science fiction	50(210)	1000(840)	1050
Sum(col.)	300	1200	1500

- X^2 (chi-square) calculation (numbers in parenthesis are expected counts calculated based on the data distribution in the two categories)

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

- It shows that two attributes are correlated in the group

Correlation Analysis (Numeric Data)

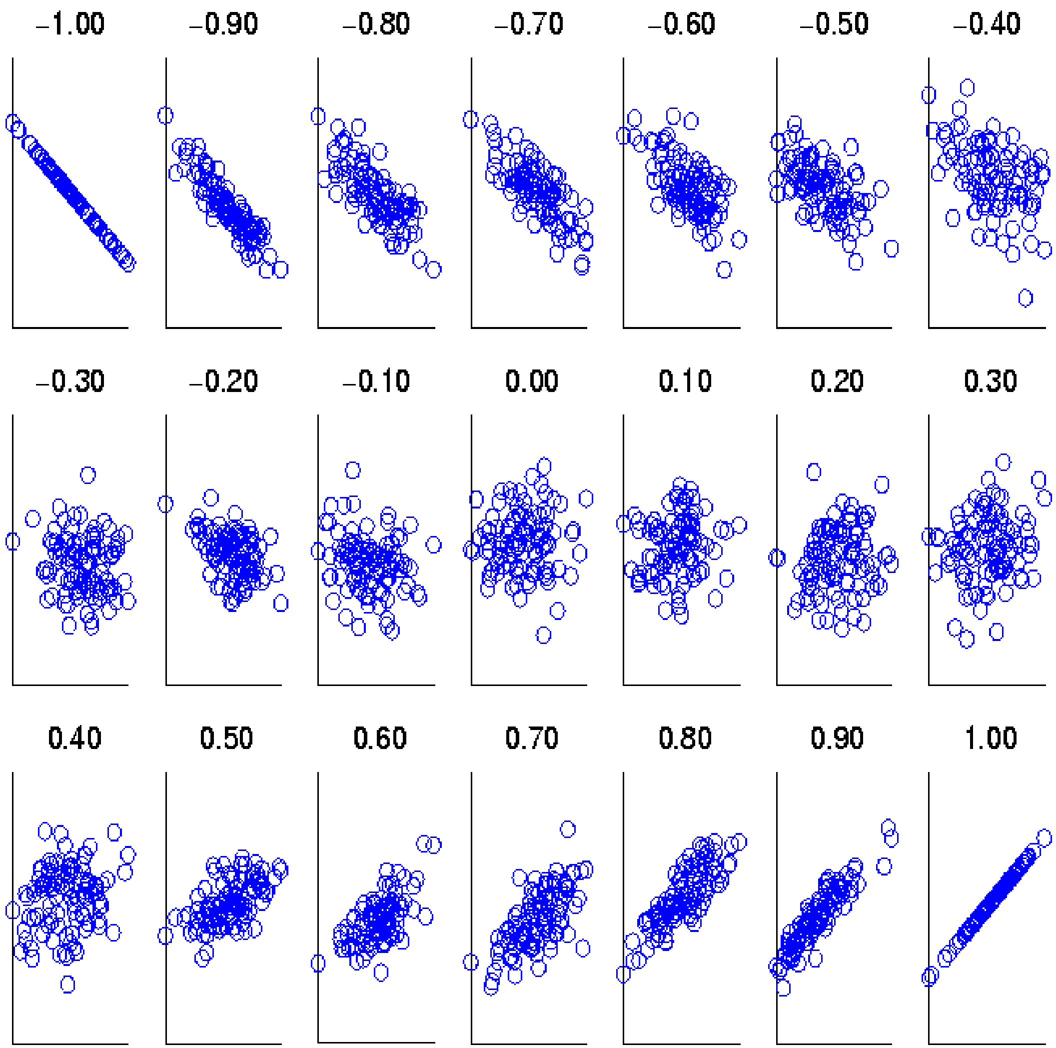
- Correlation coefficient (also called Pearson's product moment coefficient)

$$r_{A,B} = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{(n-1)\sigma_A \sigma_B} = \frac{\sum_{i=1}^{n}(a_i b_i) - n\bar{A}\bar{B}}{(n-1)\sigma_A \sigma_B}$$

where n is the number of tuples, $\bar{A}$ and $\bar{B}$ are the respective means of A and B, σ_A and σ_B are the respective standard deviation of A and B, and $\Sigma(a_i b_i)$ is the sum of the AB cross-product.

- If $r_{A,B} > 0$, A and B are positively correlated (A's values increase as B's). The higher the value, the stronger the correlation.
- $r_{A,B} = 0$: independent; $r_{AB} < 0$: negatively correlated

Visually Evaluating Correlation



Scatter plots showing the similarity from –1 to 1.

Data Reduction

Data Reduction

- **Data reduction**: Obtain a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results

- Why data reduction? — A database/data warehouse may store petabytes of data. Complex data analysis may take a very long time to run on the complete data set.

Dimensionality Reduction

- **Curse of dimensionality**
  - When dimensionality increases, data becomes increasingly sparse
  - Density and distance between points, which is critical to clustering, outlier analysis, becomes less meaningful
  - The possible combinations of subspaces will grow exponentially

- **Dimensionality reduction**
  - Avoid the curse of dimensionality
  - Help eliminate irrelevant features and reduce noise
  - Reduce time and space required in data mining
  - Allow easier visualization

Data Reduction

- Dimensionality reduction, e.g., remove unimportant attributes
  - Principal Components Analysis (PCA)
  - Feature selection (i.e., Attribute subset selection), attribute creation
- Numerosity reduction
  - data is replaced or estimated by alternative, smaller data representations
  - Parametric
    - Regression and Log-Linear Models
  - Non-parametric
    - Histograms, clustering, sampling

Data Transformation and Data Discretization

Data Transformation

- A function that maps the entire set of values of a given attribute to a new set of replacement values so that each old value can be identified with one of the new values
- Why conduct data transformation?
  - The resulting mining process may be more efficient, the patterns found may be easier to understand
- Data Transformation Methods
  - Smoothing: Remove noise from data
  - Attribute/feature construction
    - New attributes constructed from the given ones
  - Aggregation: Summarization, data cube construction
  - Normalization: Scaled to fall within a smaller, specified range
    - min-max normalization
    - z-score normalization
    - normalization by decimal scaling
  - Discretization: raw values of numeric attributes (e.g., age) replaced by interval labels (e.g., 0-10, 11-20, etc.) or conceptual labels (e.g., youth, adult, senior)

Normalization

- **Min-max normalization**: to [new_min$_A$, new_max$_A$]

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

  - Ex. Let income range \$12,000 to \$98,000 normalized to [0.0, 1.0]. Then \$73,600 is mapped to $\frac{73,600 - 12,000}{98,000 - 12,000}(1.0 - 0) + 0 = 0.716$

- **Z-score normalization** (μ: mean, σ: standard deviation):

$$v' = \frac{v - \mu_A}{\sigma_A}$$

  - Ex. Let μ = 54,000, σ = 16,000. Then $\frac{73,600 - 54,000}{16,000} = 1.225$

- **Normalization by decimal scaling**

$$v' = \frac{v}{10^j}$$ Where j is the smallest integer such that Max(|ν'|) < 1

Summary

- **Data quality**: accuracy, completeness, consistency, timeliness, believability, interpretability
- **Data cleaning**: e.g. missing/noisy values, outliers
- **Data integration** from multiple sources:
  - Entity identification problem
  - Remove redundancies
  - Detect inconsistencies
- **Data reduction**
  - Dimensionality reduction
  - Numerosity reduction
  - Data compression
- **Data transformation and data discretization**
  - Normalization
  - Concept hierarchy generation
- Read section 3 in Data Mining Concepts and Techniques

Data Science: The Context

Ask question: What data needs to be recorded? or collected?

Why? What research question am I going to answer?

What do I want it to look like?

Real World

Humans behaving
Biology
Finance
Internet
Medicine
Sociology
Olympics

→ **Raw Data is Collected / Recorded**

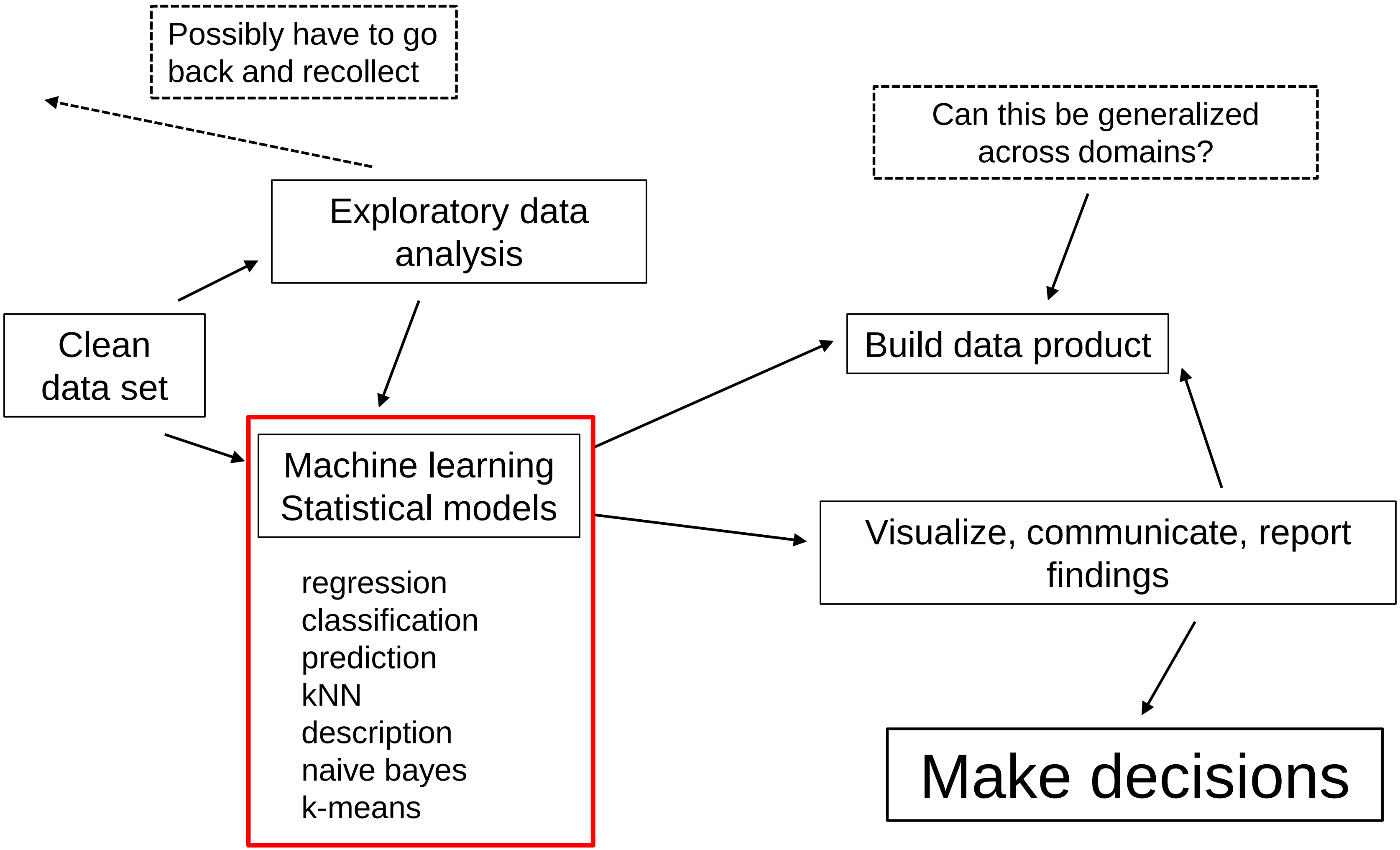
email
logs
medical records
surveys
blood drawn
(microarray)
olympic records
NYT web pages

→ **Data is Processed**

pipelines
web scraping
cleaning
munging
joining
wrangling

→ **Data Set**

“clean” table

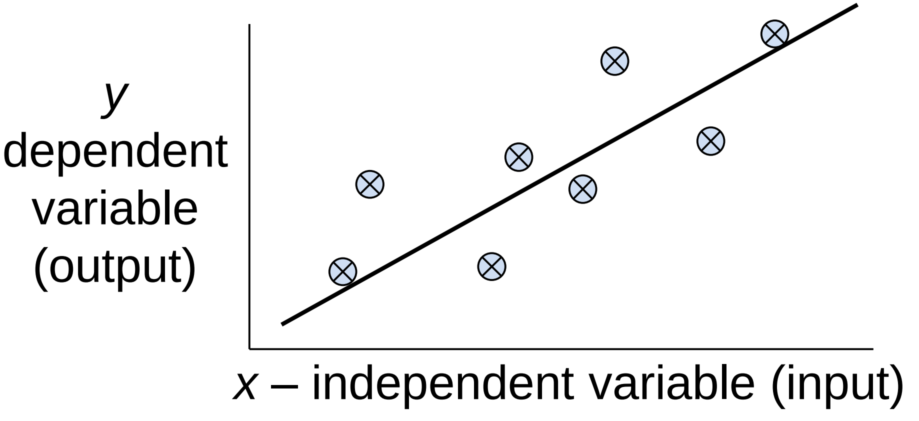

Possibly have to go back and recollect
Can this be generalized across domains?
Exploratory data analysis
Clean data set
Build data product
Machine learning Statistical models
regression
classification
prediction
kNN
description
naive bayes
k-means
Visualize, communicate, report findings
Make decisions

Mining and Analytics: Linear Regression

Some slides are adopted from Andrews Ng

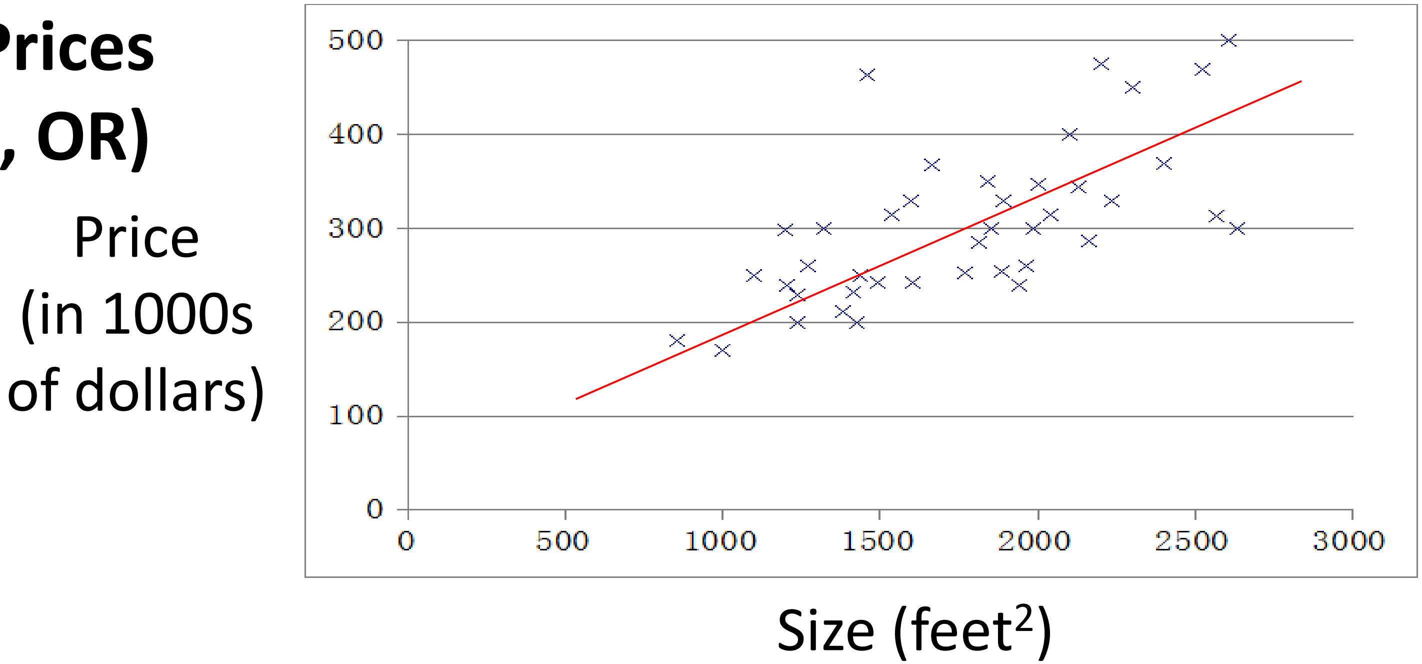
Regression

- In regression the output is continuous
  - Function Approximation
  - Also a supervised learning
    - Given the “right answer” for each example in the data.
- Many models could be used – Simplest is linear regression
  - Fit data with the best hyper-plane which "goes through" the points



Linear Regression with one Variable

Housing Prices (Portland, OR)



<u>Regression Problem</u>

Predict real-valued output

Any applications?

- advertising and sales
- consumption and income
- etc

Training set of housing prices

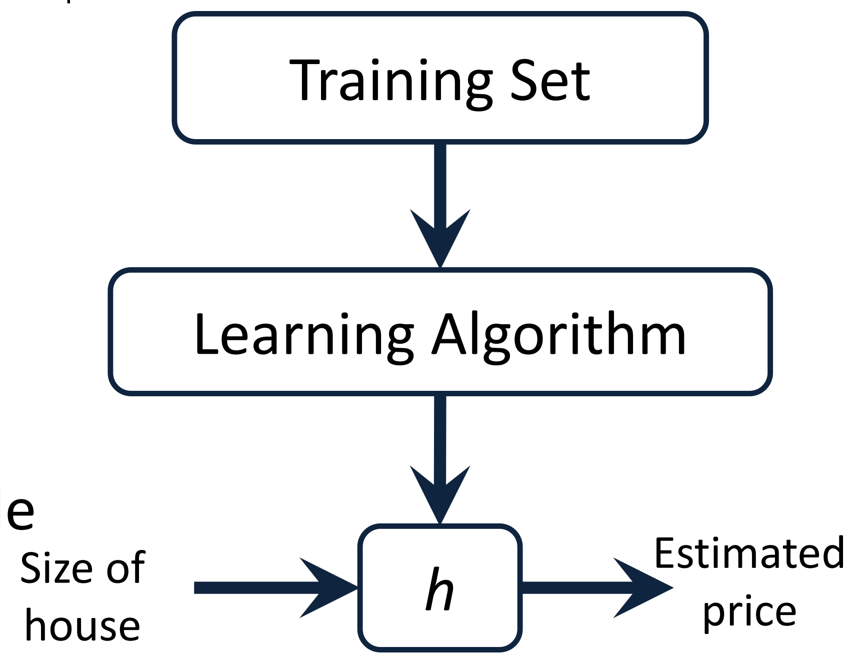
Size in feet^2 (x)	Price ($) in 1000's (y)
2104	460
1416	232
1534	315
852	178
…	…

Notation:

n = Number of training examples

x's = "input" variables / features

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



Question : How to describe ***h***?

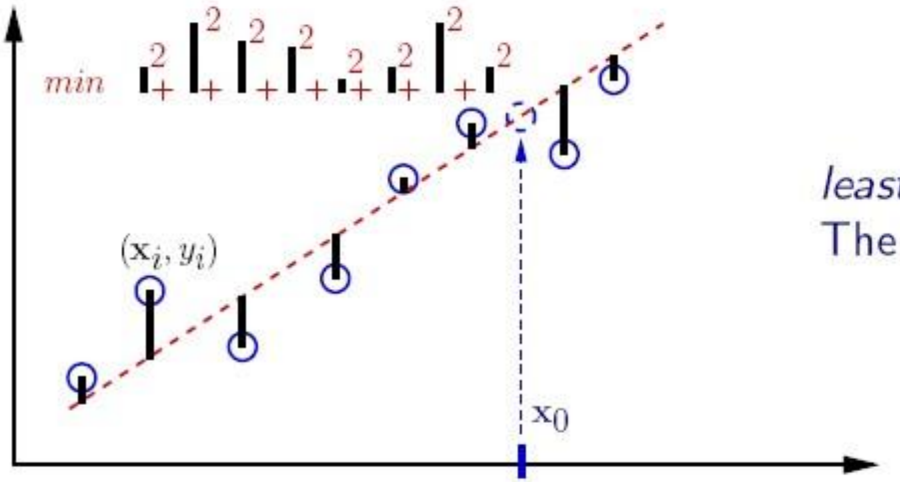
Linear Regression

- Hypothesis:

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ \ldots + \theta_d x_d = \sum_{j=0}^{d} \theta_j x_j$$

Assume $x_0 = 1$

- Fit model by minimizing sum of squared errors



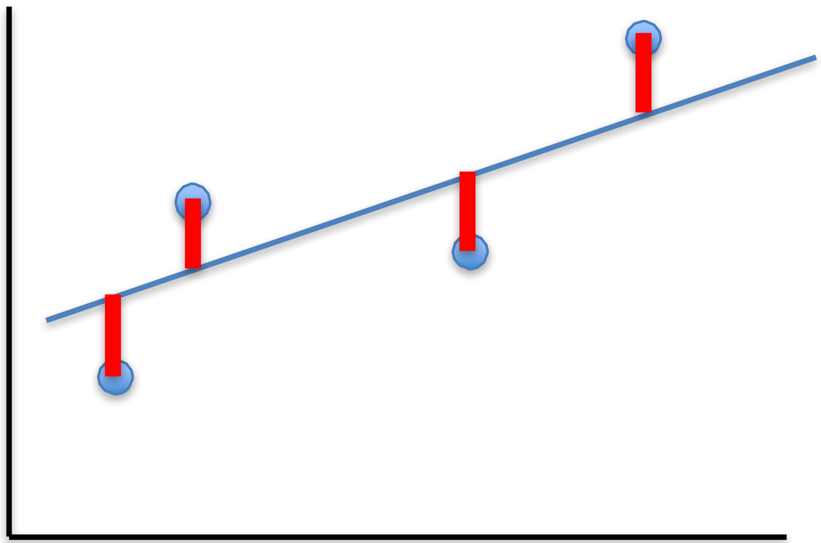
least squares (LSQ)
The fitted line is used as a predictor

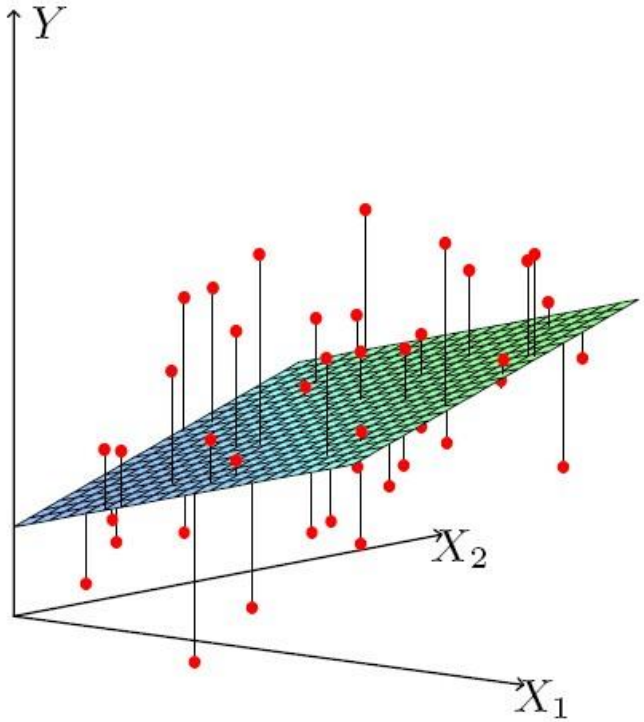
Least Squares Linear Regression

- Cost Function

$$J(\theta) = \frac{1}{2n}\sum_{i=1}^{n}(h_\theta(x^{(i)}) - y^{(i)})^2$$

- Fit by solving $\min_{\theta} J(\theta)$





Intuition Behind Cost Function

$$J(\theta) = \frac{1}{2n}\sum_{i=1}^{n}(h_\theta\left(x^{(i)}\right) - y^{(i)})^2$$

For insight on J(), let's assume $x \in \mathbb{R}$ so $\theta = [\theta_0, \theta_1]$

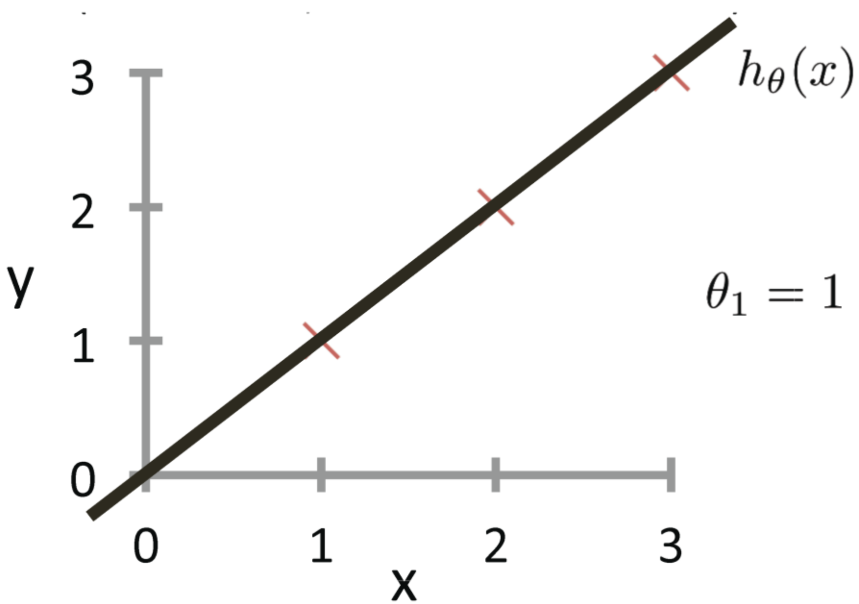
Intuition Behind Cost Function

$$J(\theta) = \frac{1}{2n}\sum_{i=1}^{n}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$$

For insight on J(), let's assume $x \in \mathbb{R}$ so $\theta = [\theta_0, \theta_1]$ ➔ $\theta_0 = 0$

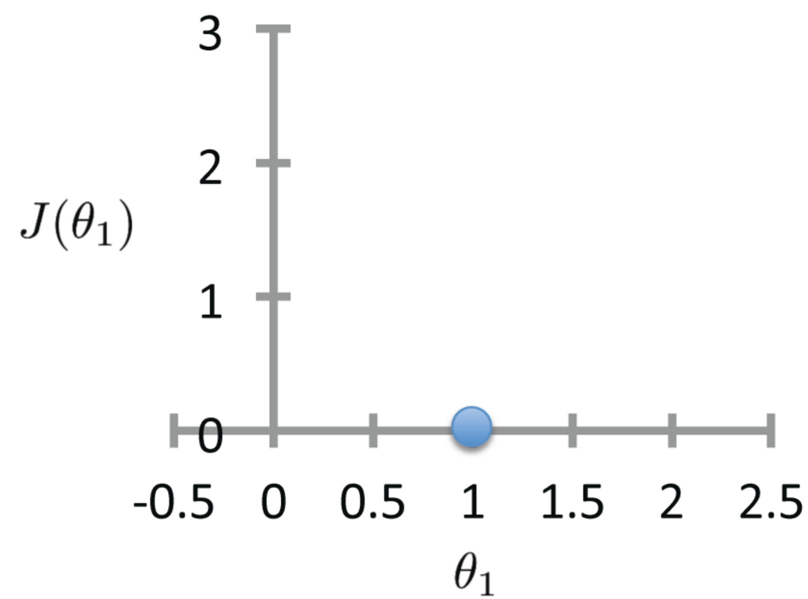
$h_\theta(x)$

(for fixed θ_1, this is a function of x)



$J(\theta)$

(function of the parameter θ_1)



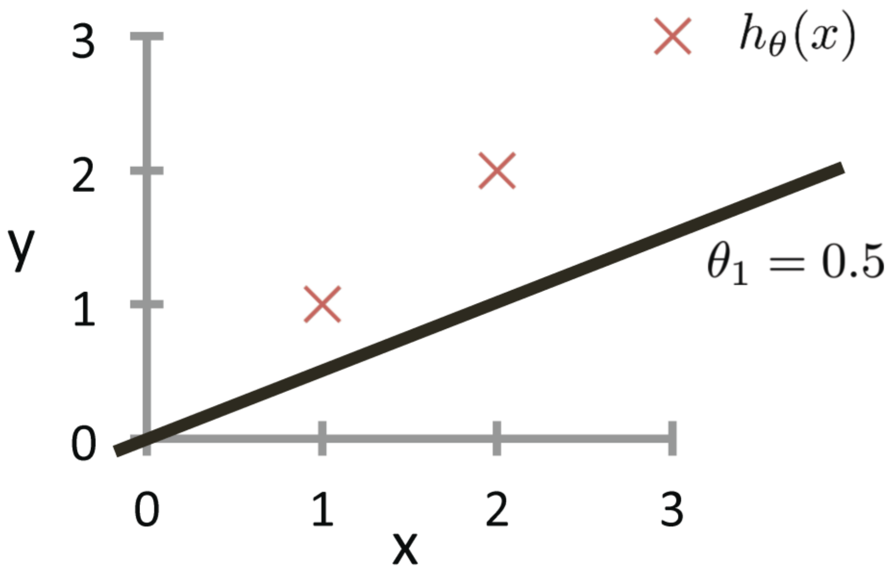
Intuition Behind Cost Function

$$J(\theta) = \frac{1}{2n}\sum_{i=1}^{n}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$$

For insight on J(), let's assume $x \in \mathbb{R}$ so $\theta = [\theta_0, \theta_1]$

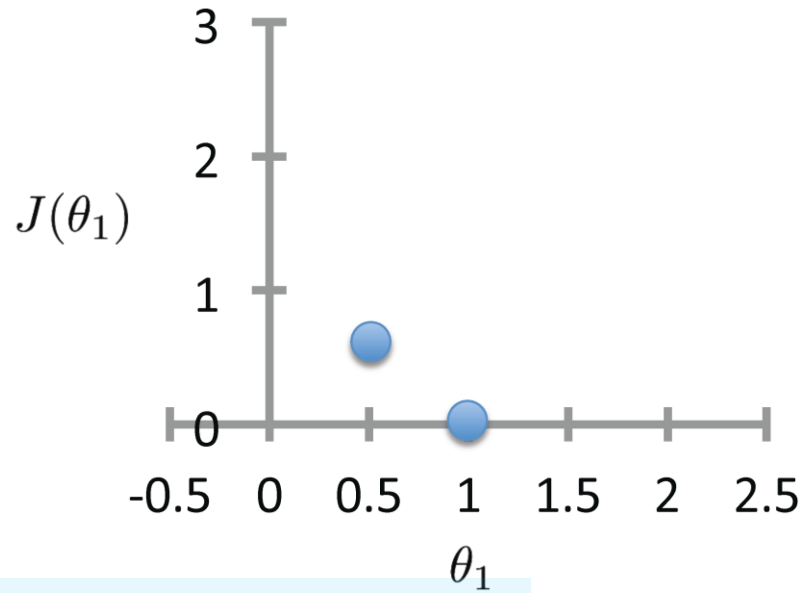
$h_\theta(x)$

(for fixed θ_1, this is a function of x)



$J(\theta)$

(function of the parameter θ_1)



$$J([0,0.5]) = \frac{1}{2 \times 3}[(0.5-1)^2 + (1-2)^2 + (1.5-3)^2 \approx 0.58$$

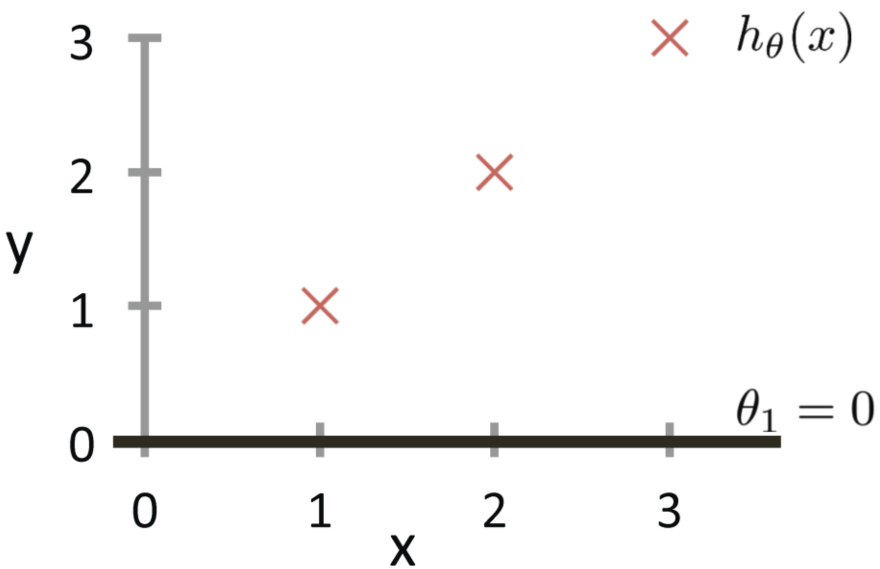
Intuition Behind Cost Function

$$J(\theta) = \frac{1}{2n}\sum_{i=1}^{n}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$$

For insight on J(), let's assume $x \in \mathbb{R}$ so $\theta = [\theta_0, \theta_1]$

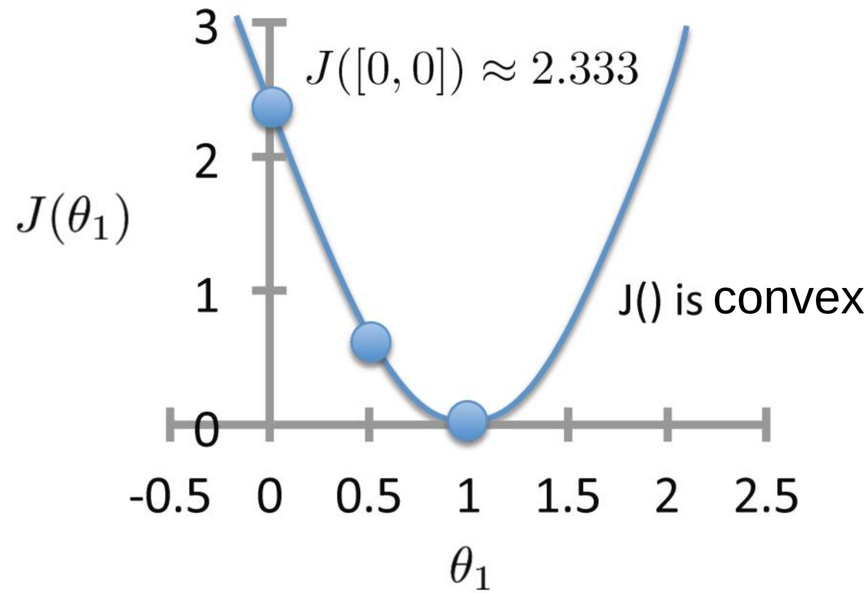
$h_\theta(x)$

(for fixed θ_1, this is a function of x)



$J(\theta)$

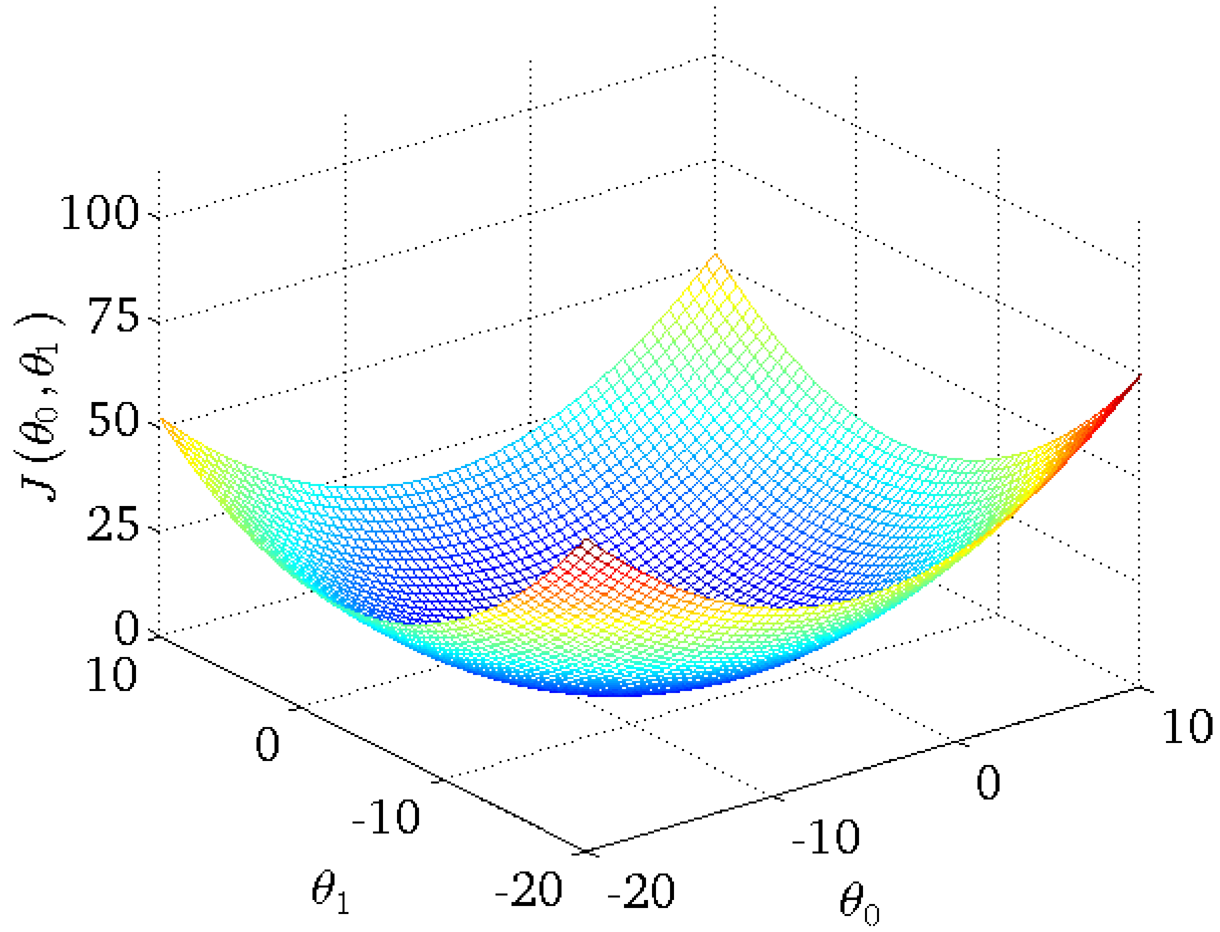
(function of the parameter θ_1)



http://mathworld.wolfram.com/ConvexFunction.html

https://www.desmos.com/calculator/kreo2ssqj8

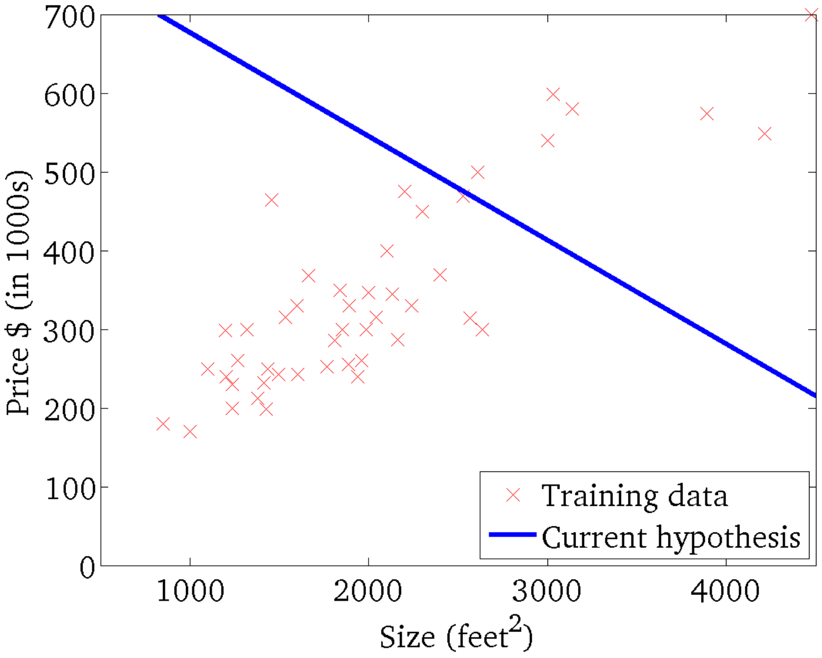
Intuition Behind Cost Function (3-D surface plot)



Intuition Behind Cost Function

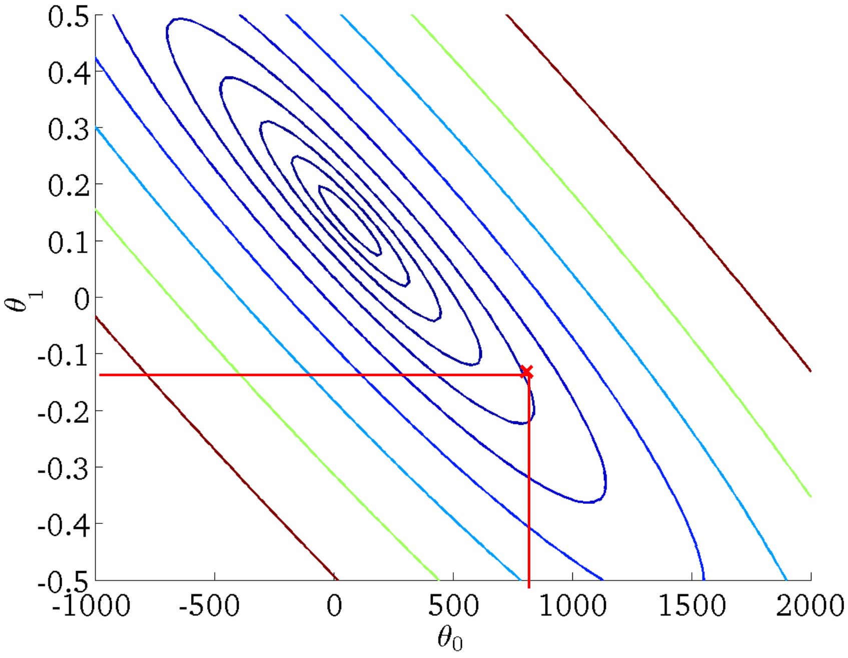
$h_\theta(x)$

(for fixed θ_0, θ_1, this is a function of x)



$J(\theta_0, \theta_1)$

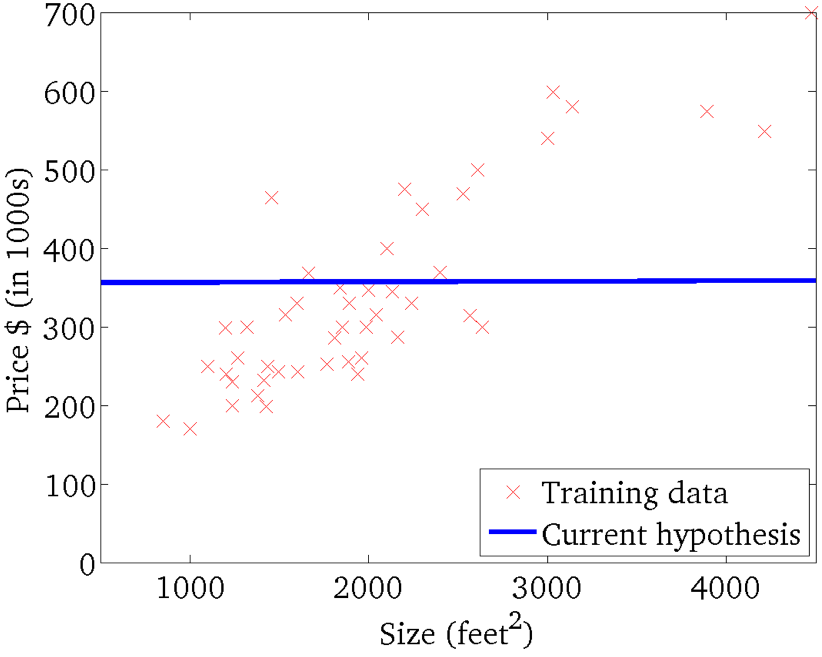
(function of the parameter θ_0, θ_1)



Intuition Behind Cost Function

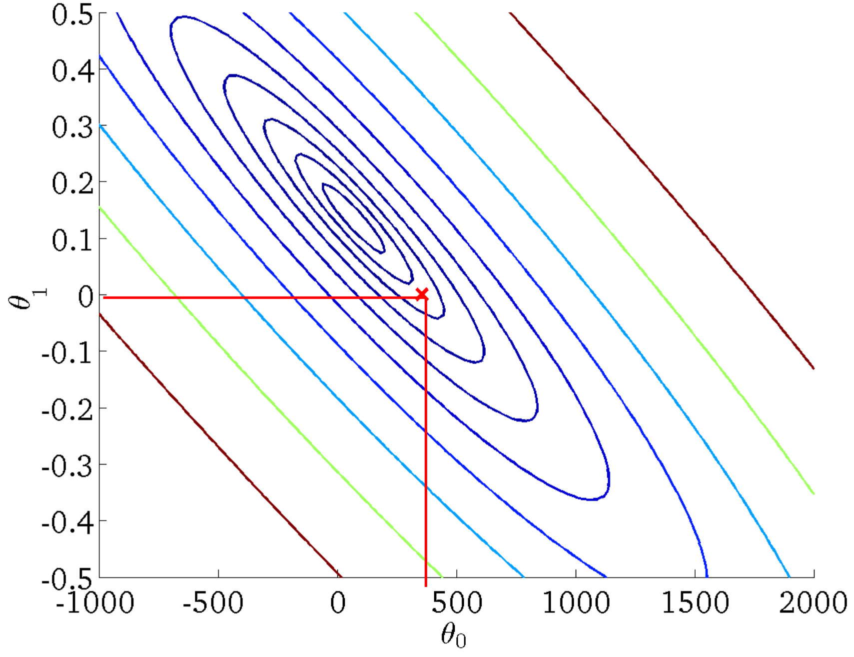
$h_\theta(x)$

(for fixedθ_0, θ_1, this is a function of x)



$J(\theta_0, \theta_1)$

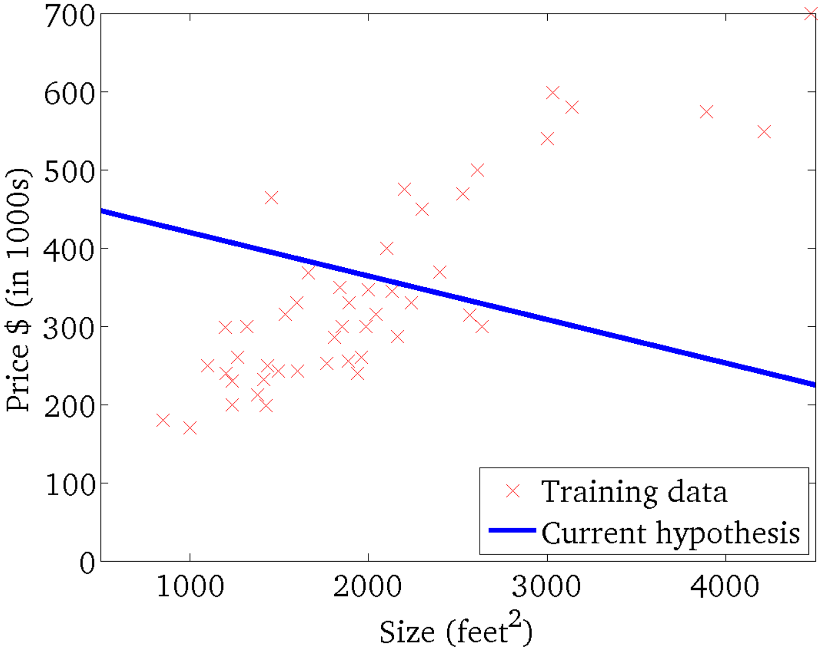
(function of the parameter θ_0, θ_1)



Intuition Behind Cost Function

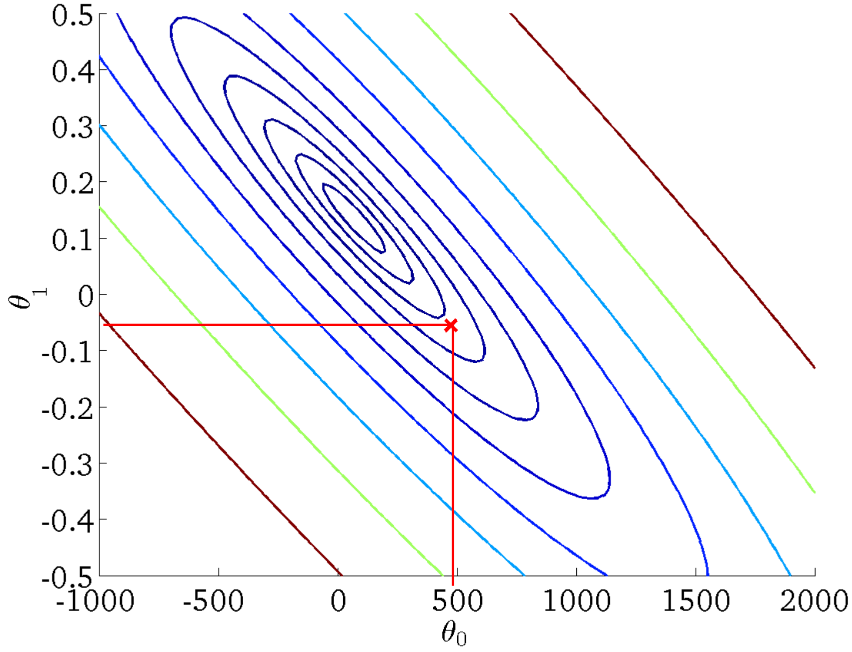
$$h_\theta(x)$$

(for fixed θ_0, θ_1, this is a function of x)



$$J(\theta_0, \theta_1)$$

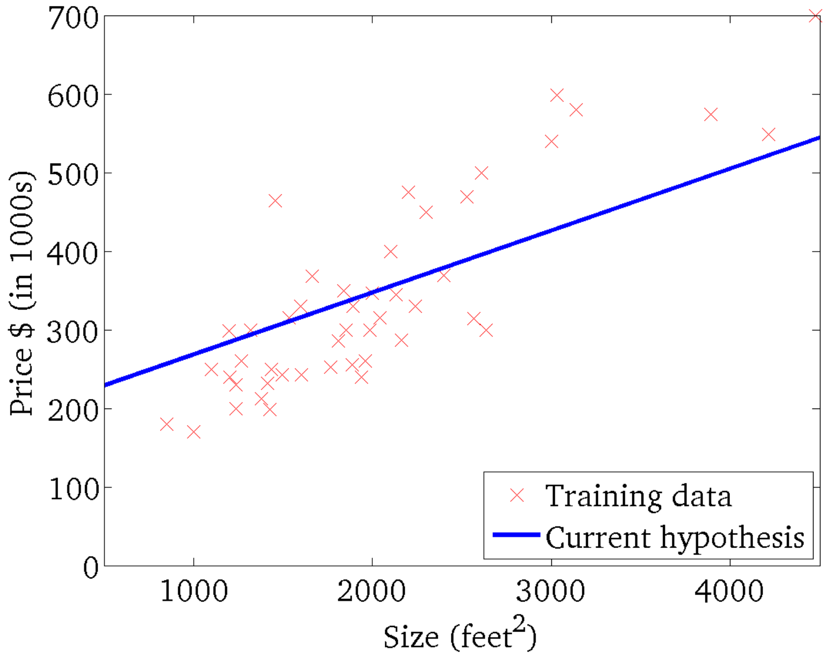
(function of the parameter θ_0, θ_1)



Intuition Behind Cost Function

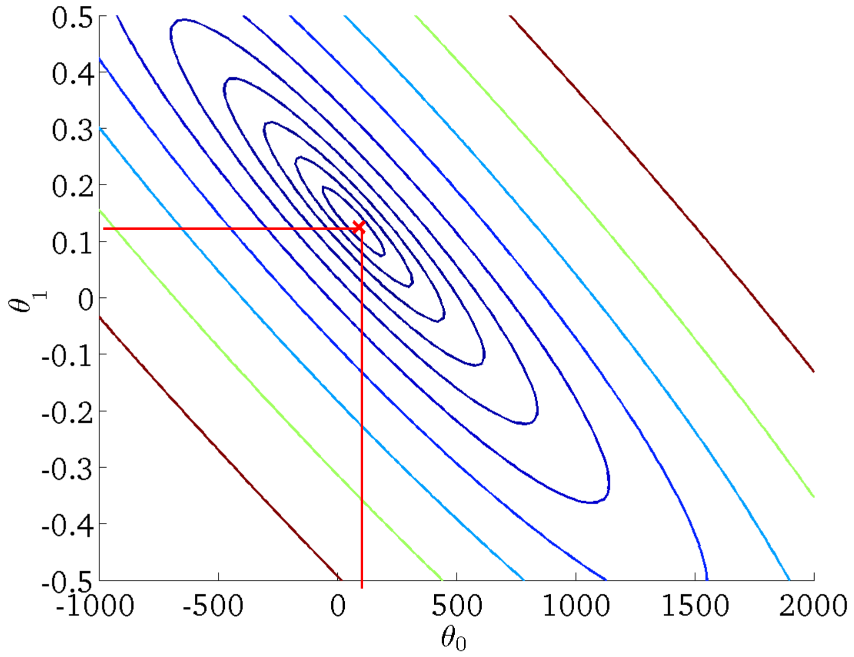
$h_\theta(x)$

(for fixed θ_0, θ_1, this is a function of x)



$J(\theta_0, \theta_1)$

(function of the parameter θ_0, θ_1)



Basic Search Procedure

- Choose initial value for θ
- Until we reach a minimum:
  - Choose a new value for θ to reduce $J(\theta)$