

CSE 4705 - Artificial Intelligence

Introduction to Machine Learning - Univariate Linear Regression

Joe Johnson M.S., Ph.D., A.S.A.

Department of Computer Science and Engineering
University of Connecticut



Table of Contents

- 1 Overview of Machine Learning
 - Definition of ML
 - Types of ML
 - ML Architecture
- 2 Example - House Price Prediction
 - House Price Prediction - Training Set and Notation
- 3 Model Development - Part 1 - Model, Cost Function Selection
 - Methodology
 - Model Selection
 - Cost Function Selection
- 4 Model Development - Part 2 - Model Training
 - Review of The Chain Rule
 - Gradient Descent - Simplified Model
 - Gradient Descent - Full Model
 - Training the House Price Prediction Model

Table of Contents

1 Overview of Machine Learning

- Definition of ML
- Types of ML
- ML Architecture

2 Example - House Price Prediction

- House Price Prediction - Training Set and Notation

3 Model Development - Part 1 - Model, Cost Function Selection

- Methodology
- Model Selection
- Cost Function Selection

4 Model Development - Part 2 - Model Training

- Review of The Chain Rule
- Gradient Descent - Simplified Model
- Gradient Descent - Full Model
- Training the House Price Prediction Model

Table of Contents

1 Overview of Machine Learning

- Definition of ML
- Types of ML
- ML Architecture

2 Example - House Price Prediction

- House Price Prediction - Training Set and Notation

3 Model Development - Part 1 - Model, Cost Function Selection

- Methodology
- Model Selection
- Cost Function Selection

4 Model Development - Part 2 - Model Training

- Review of The Chain Rule
- Gradient Descent - Simplified Model
- Gradient Descent - Full Model
- Training the House Price Prediction Model

Definition of Machine Learning

Machine Learning

Definition: Machine Learning

Definition of Machine Learning

Machine Learning

Definition: **Machine Learning** is the area of computer science focused on building software in which rules are not explicitly coded but are instead implemented via the analysis of data.

Definition of Machine Learning

Machine Learning

Definition: **Machine Learning** is the area of computer science focused on building software in which rules are not explicitly coded but are instead implemented via the analysis of data.

- Contexts:

Definition of Machine Learning

Machine Learning

Definition: **Machine Learning** is the area of computer science focused on building software in which rules are not explicitly coded but are instead implemented via the analysis of data.

- Contexts:

- Rules of the problem domain are not well known,

Definition of Machine Learning

Machine Learning

Definition: **Machine Learning** is the area of computer science focused on building software in which rules are not explicitly coded but are instead implemented via the analysis of data.

- Contexts:

- Rules of the problem domain are not well known, but sufficient data is available such that the rules can be *inferred* using algorithms.

Definition of Machine Learning

Machine Learning

Definition: **Machine Learning** is the area of computer science focused on building software in which rules are not explicitly coded but are instead implemented via the analysis of data.

- Contexts:

- Rules of the problem domain are not well known, but sufficient data is available such that the rules can be *inferred* using algorithms.
- Rules are complicated

Definition of Machine Learning

Machine Learning

Definition: **Machine Learning** is the area of computer science focused on building software in which rules are not explicitly coded but are instead implemented via the analysis of data.

- Contexts:

- Rules of the problem domain are not well known, but sufficient data is available such that the rules can be *inferred* using algorithms.
- Rules are complicated and thus, cannot be easily captured in the form of explicit application requirements.

Table of Contents

1 Overview of Machine Learning

- Definition of ML
- Types of ML
- ML Architecture

2 Example - House Price Prediction

- House Price Prediction - Training Set and Notation

3 Model Development - Part 1 - Model, Cost Function Selection

- Methodology
- Model Selection
- Cost Function Selection

4 Model Development - Part 2 - Model Training

- Review of The Chain Rule
- Gradient Descent - Simplified Model
- Gradient Descent - Full Model
- Training the House Price Prediction Model

Types of Machine Learning

Types of Machine Learning:

Types of Machine Learning

Types of Machine Learning:

- Supervised Machine Learning

Types of Machine Learning

Types of Machine Learning:

- Supervised Machine Learning
- Unsupervised Machine Learning

Types of Machine Learning

Types of Machine Learning:

- Supervised Machine Learning
- Unsupervised Machine Learning
- Reinforcement Learning

Types of Machine Learning

Supervised Machine Learning

Types of Machine Learning

Supervised Machine Learning

- Data we are given is in the form of validated input-output pairs

Types of Machine Learning

Supervised Machine Learning

- Data we are given is in the form of validated input-output pairs (i.e., we know each output is the correct answer for each input)

Types of Machine Learning

Supervised Machine Learning

- Data we are given is in the form of validated input-output pairs (i.e., we know each output is the correct answer for each input) and we try to infer the model that is implied by this data.

Types of Machine Learning

Supervised Machine Learning

- Data we are given is in the form of validated input-output pairs (i.e., we know each output is the correct answer for each input) and we try to infer the model that is implied by this data.
 - The learning algorithm is given the *right answers*.

Types of Machine Learning

Supervised Machine Learning

- Data we are given is in the form of validated input-output pairs (i.e., we know each output is the correct answer for each input) and we try to infer the model that is implied by this data.
 - The learning algorithm is given the *right answers*.
 - Two Types:

Types of Machine Learning

Supervised Machine Learning

- Data we are given is in the form of validated input-output pairs (i.e., we know each output is the correct answer for each input) and we try to infer the model that is implied by this data.
 - The learning algorithm is given the *right answers*.
 - Two Types:
 - Regression

Types of Machine Learning

Supervised Machine Learning

- Data we are given is in the form of validated input-output pairs (i.e., we know each output is the correct answer for each input) and we try to infer the model that is implied by this data.
 - The learning algorithm is given the *right answers*.
 - Two Types:
 - Regression
 - Classification

Supervised Machine Learning

Examples of Supervised Machine Learning:

Input (X)	Output (Y)	Application
email	spam? (0/1)	spam filtering
audio	text transcripts	speech recognition
English	Spanish	machine translation
ad, user info	click? (0/1)	online advertising
image, radar info	position of other cars	self-driving car
image of phone	defect? (0/1)	visual inspection

Supervised Machine Learning - Regression

Supervised Machine Learning

Definition: Regression

Supervised Machine Learning - Regression

Supervised Machine Learning

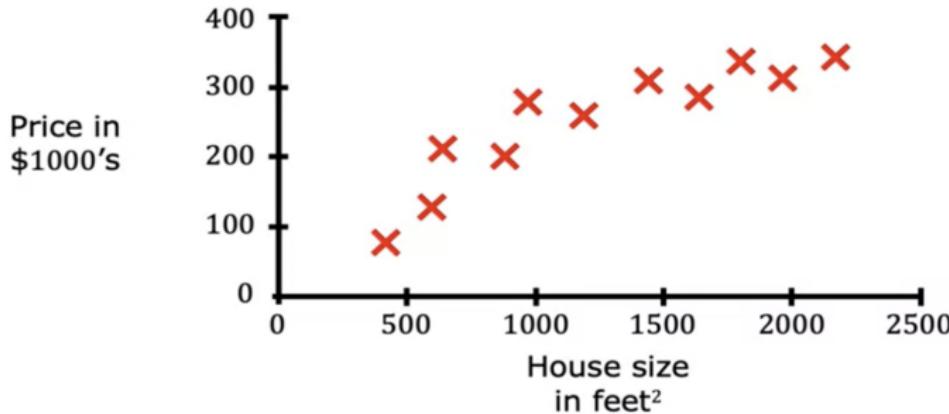
Definition: **Regression** is a form of supervised machine learning in which the target variable is *quantitative*.

Supervised Machine Learning - Regression

Supervised Machine Learning

Definition: **Regression** is a form of supervised machine learning in which the target variable is *quantitative*.

Regression: Housing price prediction



Supervised Machine Learning - Regression

Supervised Machine Learning

Definition: Classification

Supervised Machine Learning - Regression

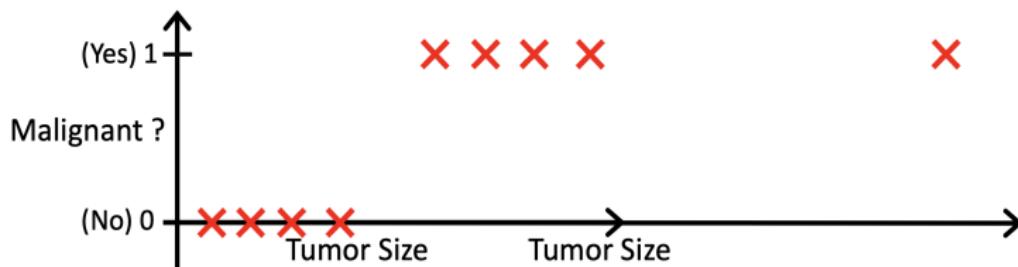
Supervised Machine Learning

Definition: **Classification** is a form of supervised machine learning in which the target variable is *categorical*.

Supervised Machine Learning - Regression

Supervised Machine Learning

Definition: **Classification** is a form of supervised machine learning in which the target variable is *categorical*.

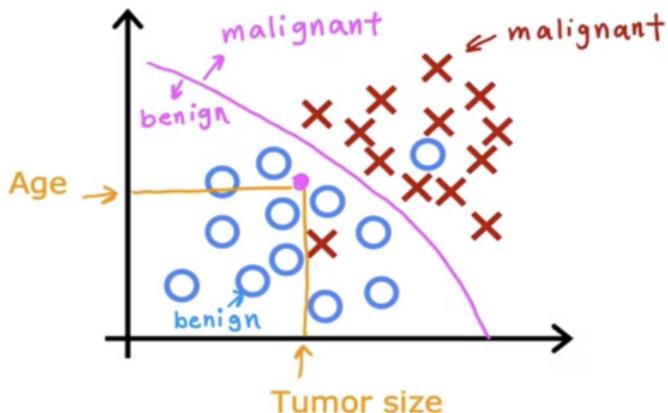


Supervised Machine Learning - Classification

Supervised Machine Learning

Definition: **Classification** is a form of supervised machine learning in which the target variable is *categorical*.

Two or more inputs



Types of Machine Learning

Unsupervised Machine Learning:

Types of Machine Learning

Unsupervised Machine Learning:

- Data we are given is *not labeled*,

Types of Machine Learning

Unsupervised Machine Learning:

- Data we are given is *not labeled*, (i.e., we are not given the *right answers*),

Types of Machine Learning

Unsupervised Machine Learning:

- Data we are given is *not labeled*, (i.e., we are not given the *right answers*), but instead is simply a collection of points from a population for which we wish to find interesting characteristics:

Types of Machine Learning

Unsupervised Machine Learning:

- Data we are given is *not labeled*, (i.e., we are not given the *right answers*), but instead is simply a collection of points from a population for which we wish to find interesting characteristics:
 - structure

Types of Machine Learning

Unsupervised Machine Learning:

- Data we are given is *not labeled*, (i.e., we are not given the *right answers*), but instead is simply a collection of points from a population for which we wish to find interesting characteristics:
 - structure
 - patterns

Types of Machine Learning

Unsupervised Machine Learning:

- Data we are given is *not labeled*, (i.e., we are not given the *right answers*), but instead is simply a collection of points from a population for which we wish to find interesting characteristics:
 - structure
 - patterns
 - groupings (clusters)

Types of Machine Learning

Unsupervised Machine Learning:

- Data we are given is *not labeled*, (i.e., we are not given the *right answers*), but instead is simply a collection of points from a population for which we wish to find interesting characteristics:
 - structure
 - patterns
 - groupings (clusters)
 - outliers (anomalies)

Unsupervised Learning

Types of Unsupervised Learning:

Unsupervised Learning

Types of Unsupervised Learning:

- Clustering

Unsupervised Learning

Types of Unsupervised Learning:

- Clustering
- Anomaly Detection

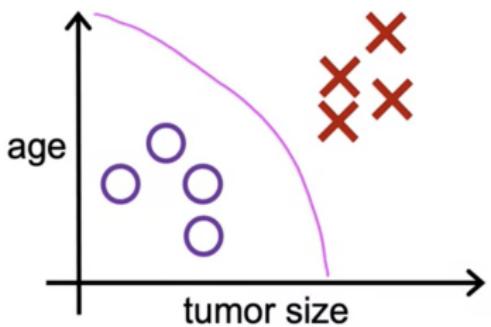
Unsupervised Learning

Types of Unsupervised Learning:

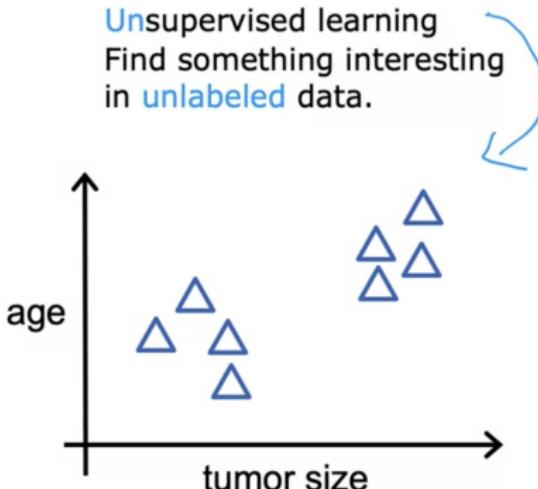
- Clustering
- Anomaly Detection
- Dimensionality Reduction

Unsupervised Machine Learning

Supervised learning
Learn from data **labeled**
with the **"right answers"**



Unsupervised learning
Find something interesting
in **unlabeled** data.



Types of Machine Learning

Reinforcement Learning:

Types of Machine Learning

Reinforcement Learning:

- ML in which input-output pairs are acquired through experience and assigned a value by a reward function relative to some task, and whose performance is optimized by maximizing the value of the reward function.

Table of Contents

1 Overview of Machine Learning

- Definition of ML
- Types of ML
- ML Architecture

2 Example - House Price Prediction

- House Price Prediction - Training Set and Notation

3 Model Development - Part 1 - Model, Cost Function Selection

- Methodology
- Model Selection
- Cost Function Selection

4 Model Development - Part 2 - Model Training

- Review of The Chain Rule
- Gradient Descent - Simplified Model
- Gradient Descent - Full Model
- Training the House Price Prediction Model

Machine Learning System Architecture

Typical Software System:

Machine Learning System Architecture

Typical Software System:

- Inputs:

Machine Learning System Architecture

Typical Software System:

- Inputs: Data: Input

Machine Learning System Architecture

Typical Software System:

- Inputs: Data: Input
- System:

Machine Learning System Architecture

Typical Software System:

- Inputs: Data: Input
- System: Algorithm

Machine Learning System Architecture

Typical Software System:

- Inputs: Data: Input
- System: Algorithm
- Outputs:

Machine Learning System Architecture

Typical Software System:

- Inputs: Data: Input
- System: Algorithm
- Outputs: Data: Output

Machine Learning System Architecture

(Supervised) Machine Learning System:

Machine Learning System Architecture

(Supervised) Machine Learning System:

- Inputs:

Machine Learning System Architecture

(Supervised) Machine Learning System:

- Inputs: Data: Input, Output

Machine Learning System Architecture

(Supervised) Machine Learning System:

- Inputs: Data: Input, Output
- System:

Machine Learning System Architecture

(Supervised) Machine Learning System:

- Inputs: Data: Input, Output
- System: Learning Algorithm

Machine Learning System Architecture

(Supervised) Machine Learning System:

- Inputs: Data: Input, Output
- System: Learning Algorithm
- Outputs:

Machine Learning System Architecture

(Supervised) Machine Learning System:

- Inputs: Data: Input, Output
- System: Learning Algorithm
- Outputs: Algorithm

Table of Contents

1 Overview of Machine Learning

- Definition of ML
- Types of ML
- ML Architecture

2 Example - House Price Prediction

- House Price Prediction - Training Set and Notation

3 Model Development - Part 1 - Model, Cost Function Selection

- Methodology
- Model Selection
- Cost Function Selection

4 Model Development - Part 2 - Model Training

- Review of The Chain Rule
- Gradient Descent - Simplified Model
- Gradient Descent - Full Model
- Training the House Price Prediction Model

Table of Contents

1 Overview of Machine Learning

- Definition of ML
- Types of ML
- ML Architecture

2 Example - House Price Prediction

- House Price Prediction - Training Set and Notation

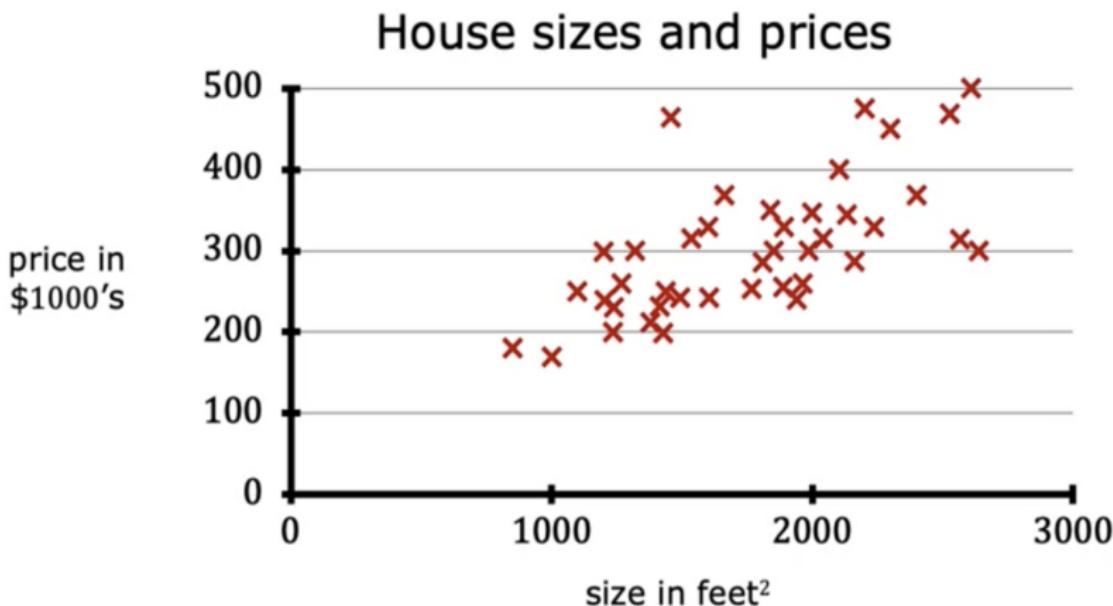
3 Model Development - Part 1 - Model, Cost Function Selection

- Methodology
- Model Selection
- Cost Function Selection

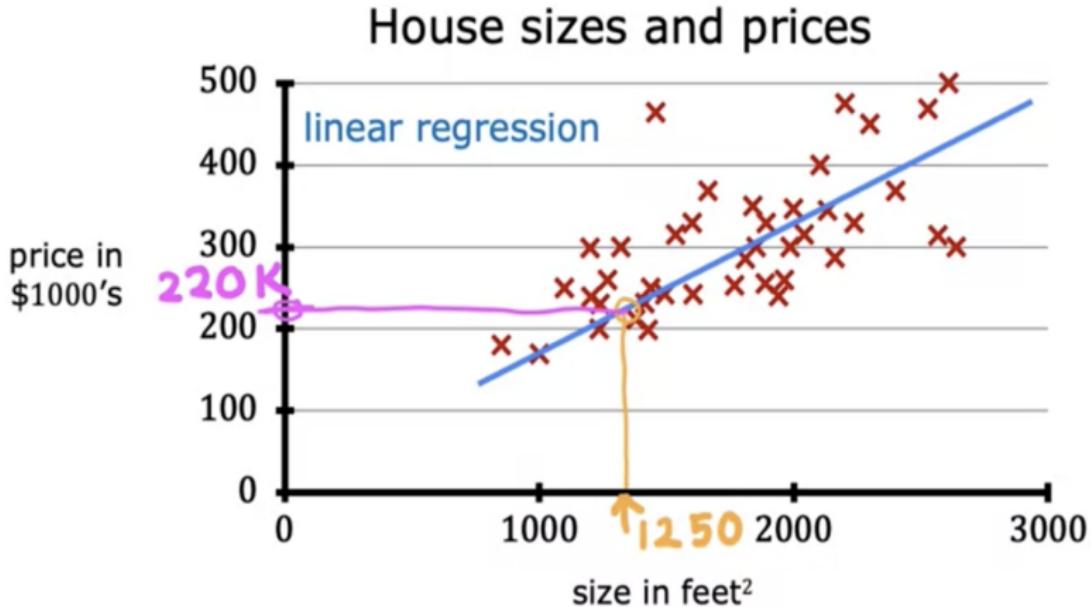
4 Model Development - Part 2 - Model Training

- Review of The Chain Rule
- Gradient Descent - Simplified Model
- Gradient Descent - Full Model
- Training the House Price Prediction Model

House Price Prediction - Training Set and Notation



House Price Prediction - Training Set and Notation



House Price Prediction - Training Set and Notation

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

House Price Prediction - Training Set and Notation

Notation:

House Price Prediction - Training Set and Notation

Notation:

- m = number of training samples

House Price Prediction - Training Set and Notation

Notation:

- m = number of training samples
- x = input variable/feature

House Price Prediction - Training Set and Notation

Notation:

- m = number of training samples
- x = input variable/feature
- y = output variable/ target

House Price Prediction - Training Set and Notation

Notation:

- m = number of training samples
- x = input variable/feature
- y = output variable/ target
- (x, y) = one training example

House Price Prediction - Training Set and Notation

Notation:

- m = number of training samples
- x = input variable/feature
- y = output variable/ target
- (x, y) = one training example
- $(x^{(i)}, y^{(i)})$ = i^{th} training example

House Price Prediction - Training Set and Notation

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

So, for our house prices example, we have:

House Price Prediction - Training Set and Notation

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

So, for our house prices example, we have:

- $x^{(1)} = 2104$

House Price Prediction - Training Set and Notation

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

So, for our house prices example, we have:

- $x^{(1)} = 2104$
- $x^{(2)} = 1416$

House Price Prediction - Training Set and Notation

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

So, for our house prices example, we have:

- $x^{(1)} = 2104$
- $x^{(2)} = 1416$
- $y^{(1)} = 460$

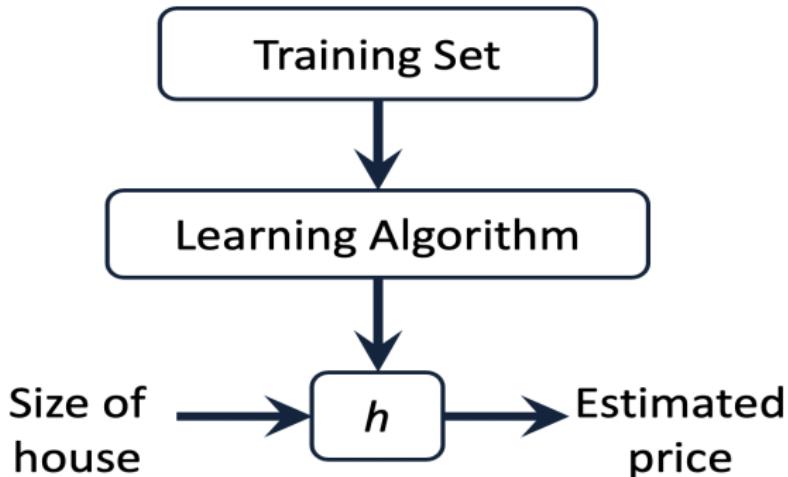
House Price Prediction - Training Set and Notation

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

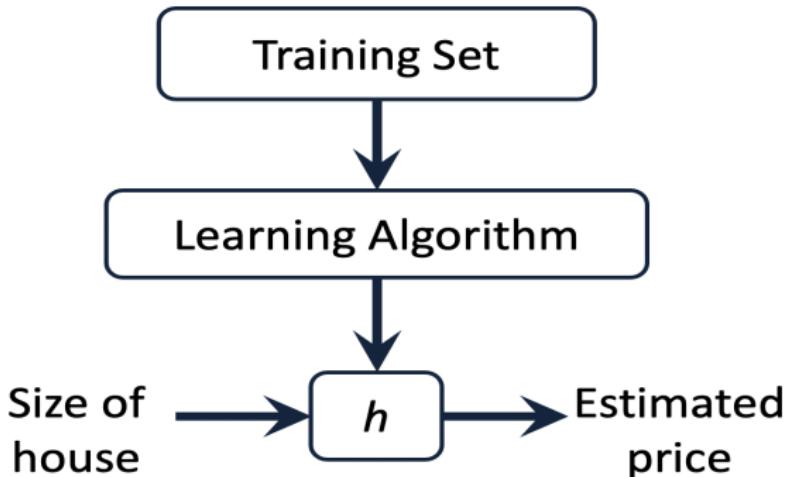
So, for our house prices example, we have:

- $x^{(1)} = 2104$
- $x^{(2)} = 1416$
- $y^{(1)} = 460$
- $y^{(2)} = 232$

House Price Prediction - Training Set and Notation

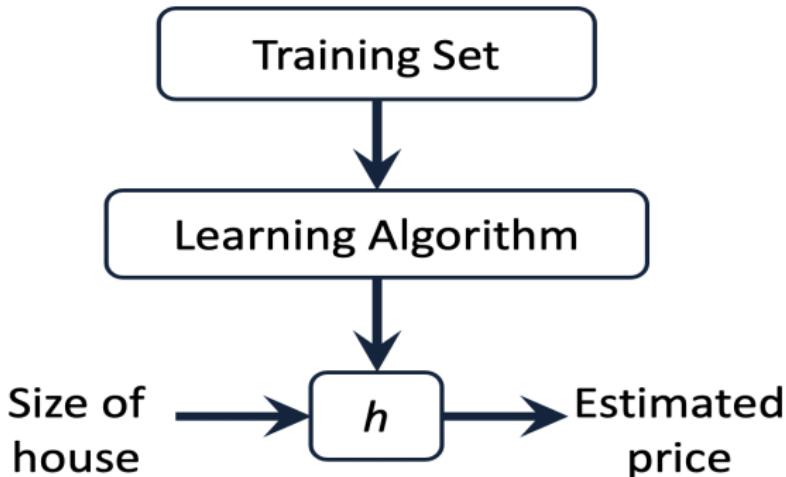


House Price Prediction - Training Set and Notation



Notation:

House Price Prediction - Training Set and Notation



Notation:

- $\hat{y}^{(i)}$ = estimated price for the i^{th} sample, based on our model, f .

Table of Contents

- 1 Overview of Machine Learning
 - Definition of ML
 - Types of ML
 - ML Architecture
- 2 Example - House Price Prediction
 - House Price Prediction - Training Set and Notation
- 3 Model Development - Part 1 - Model, Cost Function Selection
 - Methodology
 - Model Selection
 - Cost Function Selection
- 4 Model Development - Part 2 - Model Training
 - Review of The Chain Rule
 - Gradient Descent - Simplified Model
 - Gradient Descent - Full Model
 - Training the House Price Prediction Model

Table of Contents

1 Overview of Machine Learning

- Definition of ML
- Types of ML
- ML Architecture

2 Example - House Price Prediction

- House Price Prediction - Training Set and Notation

3 Model Development - Part 1 - Model, Cost Function Selection

- Methodology
- Model Selection
- Cost Function Selection

4 Model Development - Part 2 - Model Training

- Review of The Chain Rule
- Gradient Descent - Simplified Model
- Gradient Descent - Full Model
- Training the House Price Prediction Model

Methodology

Approach:

Methodology

Approach:

- ➊ Model Selection:

Methodology

Approach:

- ① Model Selection: Pick a general form for our model, f .

Methodology

Approach:

- ① Model Selection: Pick a general form for our model, f .
- ② Cost Function Selection:

Methodology

Approach:

- ① Model Selection: Pick a general form for our model, f .
- ② Cost Function Selection: Choose a cost function.

Methodology

Approach:

- ① Model Selection: Pick a general form for our model, f .
- ② Cost Function Selection: Choose a cost function. The cost function, J , will guide our search for the *best parameters* for our model.

Methodology

Approach:

- ① Model Selection: Pick a general form for our model, f .
- ② Cost Function Selection: Choose a cost function. The cost function, J , will guide our search for the *best parameters* for our model.
- ③ Model Training:

Methodology

Approach:

- ① Model Selection: Pick a general form for our model, f .
- ② Cost Function Selection: Choose a cost function. The cost function, J , will guide our search for the *best parameters* for our model.
- ③ Model Training: Train the model.

Methodology

Approach:

- ① Model Selection: Pick a general form for our model, f .
- ② Cost Function Selection: Choose a cost function. The cost function, J , will guide our search for the *best parameters* for our model.
- ③ Model Training: Train the model. That is, apply the gradient descent algorithm to find the optimal parameters,

Methodology

Approach:

- ① Model Selection: Pick a general form for our model, f .
- ② Cost Function Selection: Choose a cost function. The cost function, J , will guide our search for the *best parameters* for our model.
- ③ Model Training: Train the model. That is, apply the gradient descent algorithm to find the optimal parameters, using our general model form,

Methodology

Approach:

- ① Model Selection: Pick a general form for our model, f .
- ② Cost Function Selection: Choose a cost function. The cost function, J , will guide our search for the *best parameters* for our model.
- ③ Model Training: Train the model. That is, apply the gradient descent algorithm to find the optimal parameters, using our general model form, the cost function,

Methodology

Approach:

- ① Model Selection: Pick a general form for our model, f .
- ② Cost Function Selection: Choose a cost function. The cost function, J , will guide our search for the *best parameters* for our model.
- ③ Model Training: Train the model. That is, apply the gradient descent algorithm to find the optimal parameters, using our general model form, the cost function, and our training data.

Table of Contents

1 Overview of Machine Learning

- Definition of ML
- Types of ML
- ML Architecture

2 Example - House Price Prediction

- House Price Prediction - Training Set and Notation

3 Model Development - Part 1 - Model, Cost Function Selection

- Methodology
- Model Selection
- Cost Function Selection

4 Model Development - Part 2 - Model Training

- Review of The Chain Rule
- Gradient Descent - Simplified Model
- Gradient Descent - Full Model
- Training the House Price Prediction Model

Univariate Regression Example: House Prices Model



Univariate Regression Example: House Prices Model

- ① Pick a general form for our model, f :

Univariate Regression Example: House Prices Model

- ① Pick a general form for our model, f :
 - We choose the following form for our model f :

Univariate Regression Example: House Prices Model

- ① Pick a general form for our model, f :
 - We choose the following form for our model f :

$$f_{w,b}(x) = wx + b$$

Univariate Regression Example: House Prices Model

- ① Pick a general form for our model, f :

- We choose the following form for our model f :

$$f_{w,b}(x) = wx + b$$

- Justifications:

Univariate Regression Example: House Prices Model

- ① Pick a general form for our model, f :

- We choose the following form for our model f :

$$f_{w,b}(x) = wx + b$$

- Justifications:

- Scatter plot indicates a *roughly* linear relationship between square feet and house price.

Univariate Regression Example: House Prices Model

- ① Pick a general form for our model, f :

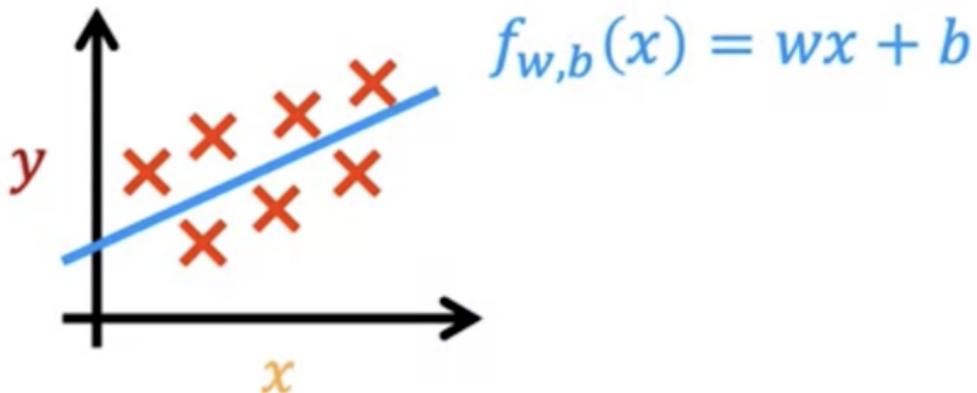
- We choose the following form for our model f :

$$f_{w,b}(x) = wx + b$$

- Justifications:

- Scatter plot indicates a *roughly* linear relationship between square feet and house price.
- We have only one input feature, square feet.

Univariate Regression Example: House Prices Model



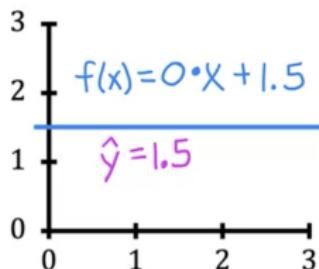
Univariate Regression Example: House Prices Model

Some examples for values of w and b :

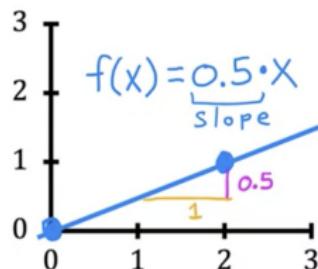
Univariate Regression Example: House Prices Model

Some examples for values of w and b :

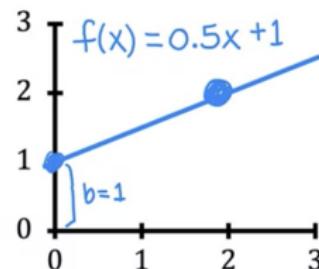
$$f_{w,b}(x) = wx + b$$



$$w = 0 \\ b = 1.5$$

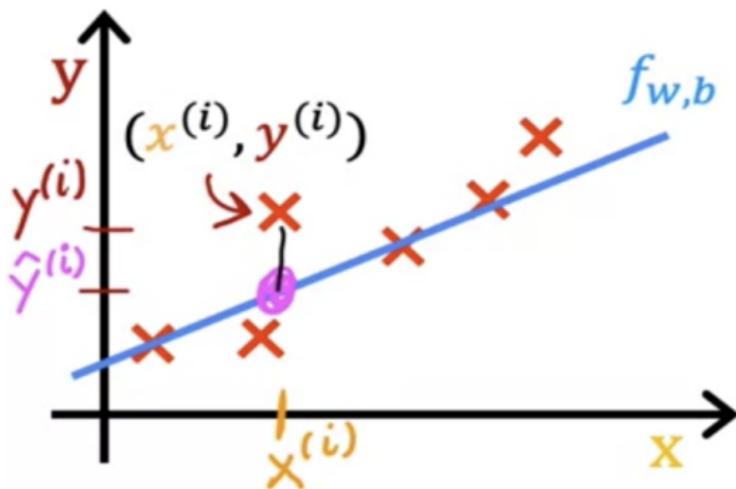


$$w = 0.5 \\ b = 0$$

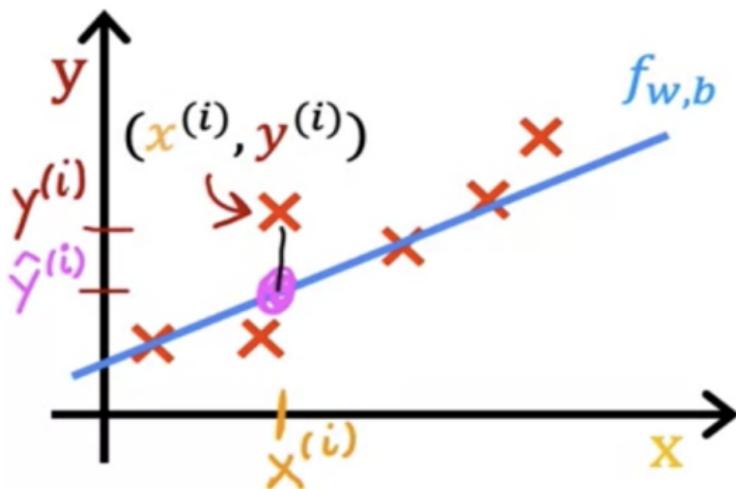


$$w = 0.5 \\ b = 1$$

Univariate Regression Example: House Prices Model

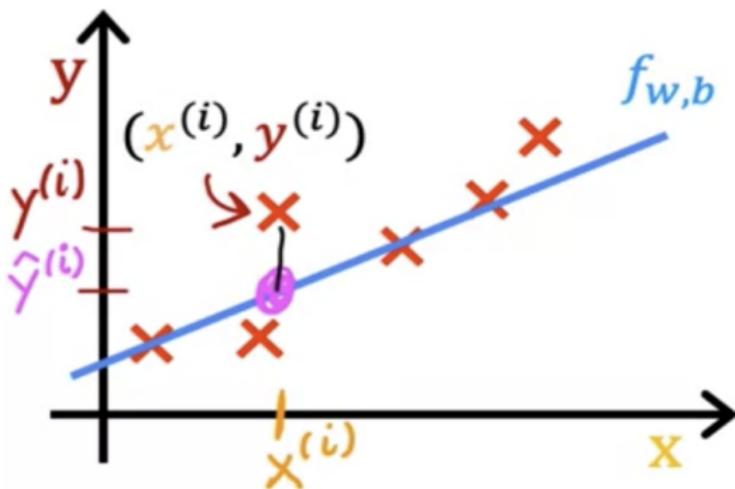


Univariate Regression Example: House Prices Model



- For each point in our training data, we have a deviation between our model's predicted value, $\hat{y}^{(i)}$) and our observed value of the target, $y^{(i)}$).

Univariate Regression Example: House Prices Model



- For each point in our training data, we have a deviation between our model's predicted value, $\hat{y}^{(i)}$ and our observed value of the target, $y^{(i)}$.
- We seek a model, $f_{w,b}(x)$ which minimizes these deviations.

Table of Contents

1 Overview of Machine Learning

- Definition of ML
- Types of ML
- ML Architecture

2 Example - House Price Prediction

- House Price Prediction - Training Set and Notation

3 Model Development - Part 1 - Model, Cost Function Selection

- Methodology
- Model Selection
- Cost Function Selection

4 Model Development - Part 2 - Model Training

- Review of The Chain Rule
- Gradient Descent - Simplified Model
- Gradient Descent - Full Model
- Training the House Price Prediction Model

Cost Function Selection

- ② Define a cost function.

Cost Function Selection

- ② Define a cost function. We choose a squared error function, (but there are other cost functions to choose from) which takes the following form:

Cost Function Selection

- ② Define a cost function. We choose a squared error function, (but there are other cost functions to choose from) which takes the following form:

$$J(w, b) = \frac{1}{2m} \sum_{i=1}^m (f_{w,b}(x^{(i)}) - y^{(i)})^2$$

Cost Function Selection

- ② Define a cost function. We choose a squared error function, (but there are other cost functions to choose from) which takes the following form:

$$\begin{aligned} J(w, b) &= \frac{1}{2m} \sum_{i=1}^m (f_{w,b}(x^{(i)}) - y^{(i)})^2 \\ &= \frac{1}{2m} \sum_{i=1}^m (wx^{(i)} + b - y^{(i)})^2 \end{aligned}$$

Cost Function Selection

- ② Define a cost function. We choose a squared error function, (but there are other cost functions to choose from) which takes the following form:

$$\begin{aligned} J(w, b) &= \frac{1}{2m} \sum_{i=1}^m (f_{w,b}(x^{(i)}) - y^{(i)})^2 \\ &= \frac{1}{2m} \sum_{i=1}^m (wx^{(i)} + b - y^{(i)})^2 \end{aligned}$$

We wish to find those values of w, b which minimize $J(w, b)$.

Cost Function Selection

- ② Define a cost function. We choose a squared error function, (but there are other cost functions to choose from) which takes the following form:

$$\begin{aligned} J(w, b) &= \frac{1}{2m} \sum_{i=1}^m (f_{w,b}(x^{(i)}) - y^{(i)})^2 \\ &= \frac{1}{2m} \sum_{i=1}^m (wx^{(i)} + b - y^{(i)})^2 \end{aligned}$$

We wish to find those values of w, b which minimize $J(w, b)$. That is, our goal is:

$$\arg \min_{w,b} J(w, b)$$

Cost Function Selection

Cost Function Intuition:

Cost Function Selection

Cost Function Intuition:

- Let's consider a simplified model,

Cost Function Selection

Cost Function Intuition:

- Let's consider a simplified model, for a moment,

Cost Function Selection

Cost Function Intuition:

- Let's consider a simplified model, for a moment, that includes only one parameter:

Cost Function Selection

Cost Function Intuition:

- Let's consider a simplified model, for a moment, that includes only one parameter:

$$f_w(x) = wx$$

Cost Function Selection

Cost Function Intuition:

- Let's consider a simplified model, for a moment, that includes only one parameter:

$$f_w(x) = wx$$

That is, we drop the parameter, b , for this model.

Cost Function Selection

Cost Function Intuition:

Cost Function Selection

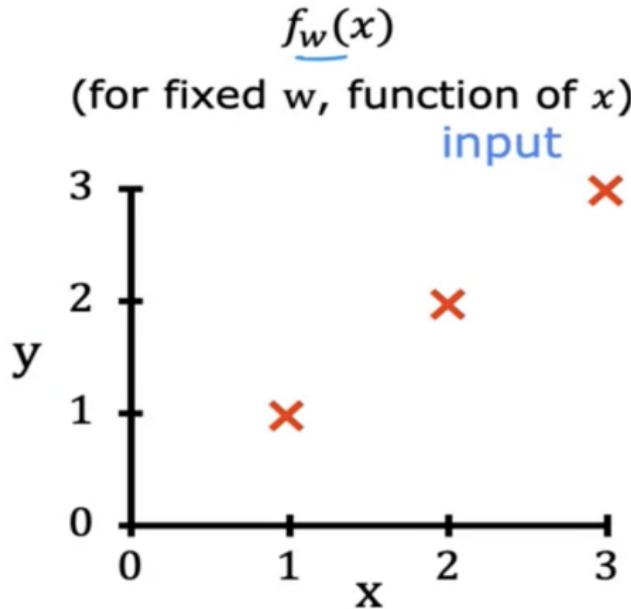
Cost Function Intuition:

- Suppose we have the following data to which we'd like to find a best fit for our model, $f_w(x)$:

Cost Function Selection

Cost Function Intuition:

- Suppose we have the following data to which we'd like to find a best fit for our model, $f_w(x)$:



Cost Function Selection

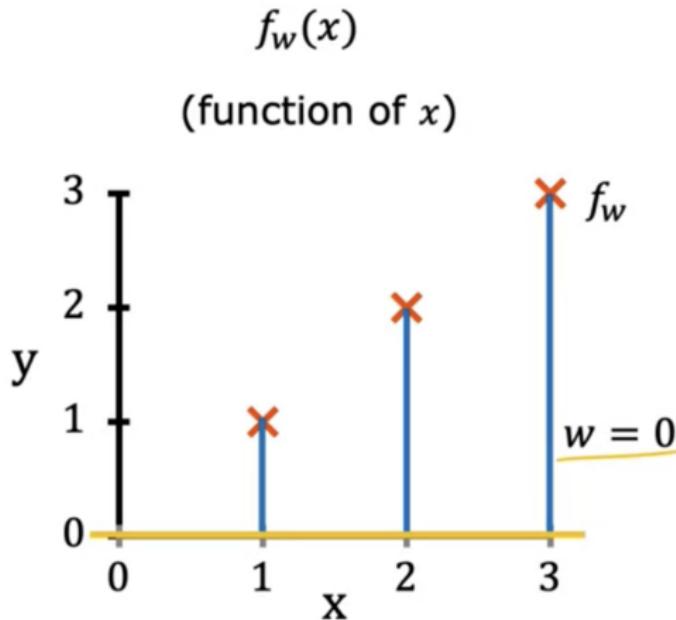
Let's look at the deviations between $f_w(x)$ and our data for different values for w .

Cost Function Selection

Let's look at the deviations between $f_w(x)$ and our data for different values for w . Consider $w = 0$:

Cost Function Selection

Let's look at the deviations between $f_w(x)$ and our data for different values for w . Consider $w = 0$:



Univariate Regression Example: House Prices Model

Calculate our cost function value when $w = 0$:

Univariate Regression Example: House Prices Model

Calculate our cost function value when $w = 0$:

$$J(0) = \frac{1}{2m} \sum_{i=1}^m (f_w(x^{(i)}) - y^{(i)})^2$$

Univariate Regression Example: House Prices Model

Calculate our cost function value when $w = 0$:

$$\begin{aligned} J(0) &= \frac{1}{2m} \sum_{i=1}^m (f_w(x^{(i)}) - y^{(i)})^2 \\ &= \frac{1}{2(3)} \sum_{i=1}^3 (wx^{(i)}) - y^{(i)})^2 \end{aligned}$$

Univariate Regression Example: House Prices Model

Calculate our cost function value when $w = 0$:

$$\begin{aligned} J(0) &= \frac{1}{2m} \sum_{i=1}^m (f_w(x^{(i)}) - y^{(i)})^2 \\ &= \frac{1}{2(3)} \sum_{i=1}^3 (wx^{(i)}) - y^{(i)})^2 \\ &= \frac{1}{6} \sum_{i=1}^3 (0 - y^{(i)})^2 \end{aligned}$$

Univariate Regression Example: House Prices Model

Calculate our cost function value when $w = 0$:

$$\begin{aligned} J(0) &= \frac{1}{2m} \sum_{i=1}^m (f_w(x^{(i)}) - y^{(i)})^2 \\ &= \frac{1}{2(3)} \sum_{i=1}^3 (wx^{(i)} - y^{(i)})^2 \\ &= \frac{1}{6} \sum_{i=1}^3 (0 - y^{(i)})^2 \\ &= \frac{1}{6} [(-1)^2 + (-2)^2 + (-3)^2] \end{aligned}$$

Univariate Regression Example: House Prices Model

Calculate our cost function value when $w = 0$:

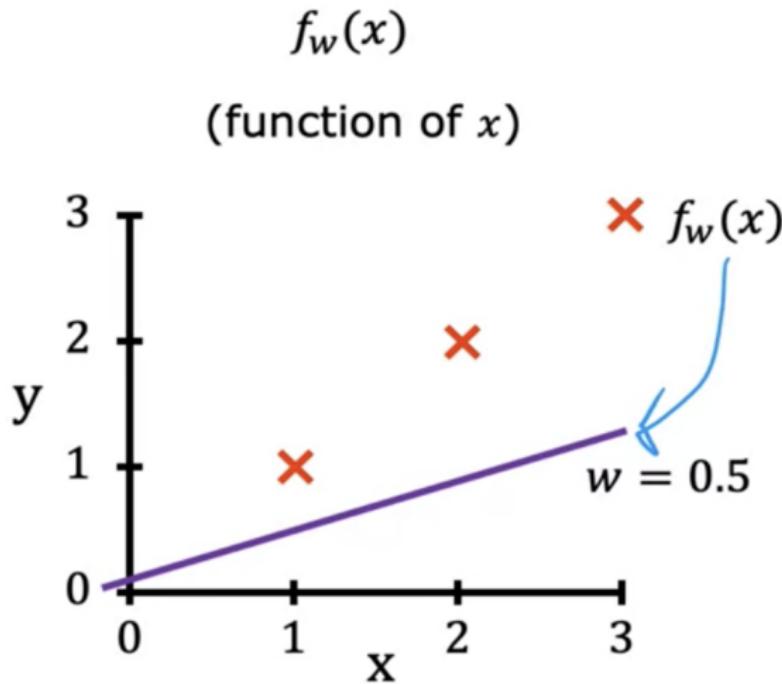
$$\begin{aligned} J(0) &= \frac{1}{2m} \sum_{i=1}^m (f_w(x^{(i)}) - y^{(i)})^2 \\ &= \frac{1}{2(3)} \sum_{i=1}^3 (wx^{(i)} - y^{(i)})^2 \\ &= \frac{1}{6} \sum_{i=1}^3 (0 - y^{(i)})^2 \\ &= \frac{1}{6} [(-1)^2 + (-2)^2 + (-3)^2] \\ &= \frac{1}{6} [1 + 4 + 9] = \frac{14}{6} \end{aligned}$$

Cost Function Selection

Now, consider $w = \frac{1}{2}$:

Cost Function Selection

Now, consider $w = \frac{1}{2}$:



Cost Function Selection

$$J\left(\frac{1}{2}\right) = \frac{1}{2m} \sum_{i=1}^m (f_w(x^{(i)}) - y^{(i)})^2$$

Cost Function Selection

$$\begin{aligned} J\left(\frac{1}{2}\right) &= \frac{1}{2m} \sum_{i=1}^m (f_w(x^{(i)}) - y^{(i)})^2 \\ &= \frac{1}{2(3)} \sum_{i=1}^3 \left(\frac{1}{2}x^{(i)}\right) - y^{(i)}\right)^2 \end{aligned}$$

Cost Function Selection

$$\begin{aligned} J\left(\frac{1}{2}\right) &= \frac{1}{2m} \sum_{i=1}^m (f_w(x^{(i)}) - y^{(i)})^2 \\ &= \frac{1}{2(3)} \sum_{i=1}^3 \left(\frac{1}{2}x^{(i)} - y^{(i)}\right)^2 \\ &= \frac{1}{6}[(0.5 - 1)^2 + (1 - 2)^2 + (1.5 - 3)^2] \end{aligned}$$

Cost Function Selection

$$\begin{aligned} J\left(\frac{1}{2}\right) &= \frac{1}{2m} \sum_{i=1}^m (f_w(x^{(i)}) - y^{(i)})^2 \\ &= \frac{1}{2(3)} \sum_{i=1}^3 \left(\frac{1}{2}x^{(i)} - y^{(i)}\right)^2 \\ &= \frac{1}{6}[(0.5 - 1)^2 + (1 - 2)^2 + (1.5 - 3)^2] \\ &= \frac{1}{6} \left[\frac{1}{4} + 1 + \frac{9}{4}\right] = \frac{1}{6} \cdot \frac{14}{4} = \frac{14}{24} \end{aligned}$$

Cost Function Selection

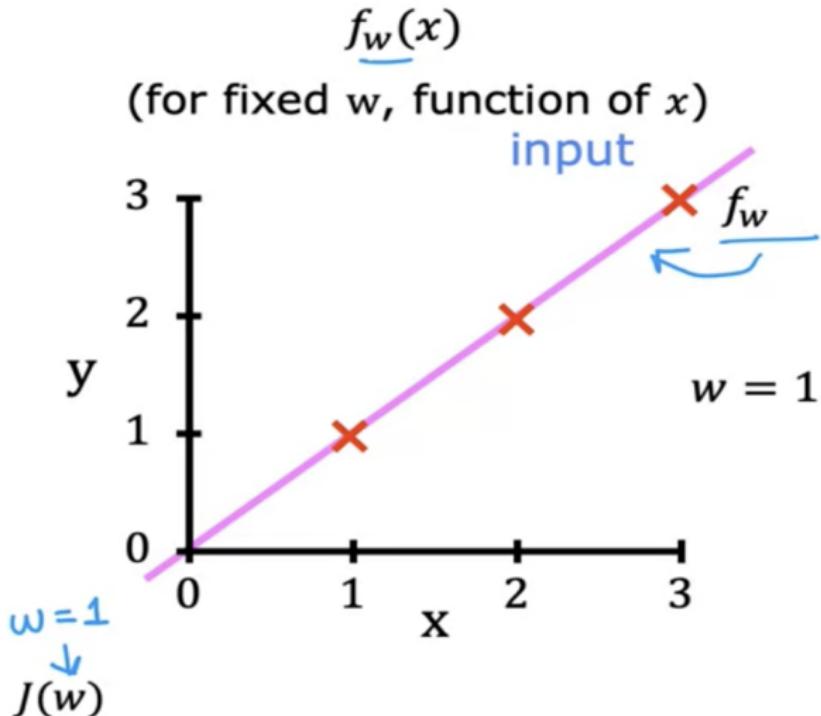
$$\begin{aligned} J\left(\frac{1}{2}\right) &= \frac{1}{2m} \sum_{i=1}^m (f_w(x^{(i)}) - y^{(i)})^2 \\ &= \frac{1}{2(3)} \sum_{i=1}^3 \left(\frac{1}{2}x^{(i)} - y^{(i)}\right)^2 \\ &= \frac{1}{6}[(0.5 - 1)^2 + (1 - 2)^2 + (1.5 - 3)^2] \\ &= \frac{1}{6} \left[\frac{1}{4} + 1 + \frac{9}{4}\right] = \frac{1}{6} \cdot \frac{14}{4} = \frac{14}{24} \\ &= \frac{7}{12} \end{aligned}$$

Cost Function Selection

Next, consider $w = 1$:

Cost Function Selection

Next, consider $w = 1$:



Cost Function Selection

$$J(1) = \frac{1}{2m} \sum_{i=1}^m (f_w(x^{(i)}) - y^{(i)})^2$$

Cost Function Selection

$$\begin{aligned} J(1) &= \frac{1}{2m} \sum_{i=1}^m (f_w(x^{(i)}) - y^{(i)})^2 \\ &= \frac{1}{2(3)} \sum_{i=1}^3 (1 \cdot x^{(i)}) - y^{(i)})^2 \end{aligned}$$

Cost Function Selection

$$\begin{aligned} J(1) &= \frac{1}{2m} \sum_{i=1}^m (f_w(x^{(i)}) - y^{(i)})^2 \\ &= \frac{1}{2(3)} \sum_{i=1}^3 (1 \cdot x^{(i)}) - y^{(i)})^2 \\ &= \frac{1}{2(3)} \sum_{i=1}^3 (x^{(i)} - y^{(i)})^2 \end{aligned}$$

Cost Function Selection

$$\begin{aligned} J(1) &= \frac{1}{2m} \sum_{i=1}^m (f_w(x^{(i)}) - y^{(i)})^2 \\ &= \frac{1}{2(3)} \sum_{i=1}^3 (1 \cdot x^{(i)}) - y^{(i)})^2 \\ &= \frac{1}{2(3)} \sum_{i=1}^3 (x^{(i)} - y^{(i)})^2 \\ &= \frac{1}{6} [(1 - 1)^2 + (2 - 2)^2 + (3 - 3)^2] \end{aligned}$$

Cost Function Selection

$$\begin{aligned} J(1) &= \frac{1}{2m} \sum_{i=1}^m (f_w(x^{(i)}) - y^{(i)})^2 \\ &= \frac{1}{2(3)} \sum_{i=1}^3 (1 \cdot x^{(i)}) - y^{(i)})^2 \\ &= \frac{1}{2(3)} \sum_{i=1}^3 (x^{(i)} - y^{(i)})^2 \\ &= \frac{1}{6} [(1 - 1)^2 + (2 - 2)^2 + (3 - 3)^2] \\ &= \frac{1}{6} [0^2 + 0^2 + 0^2] \end{aligned}$$

Cost Function Selection

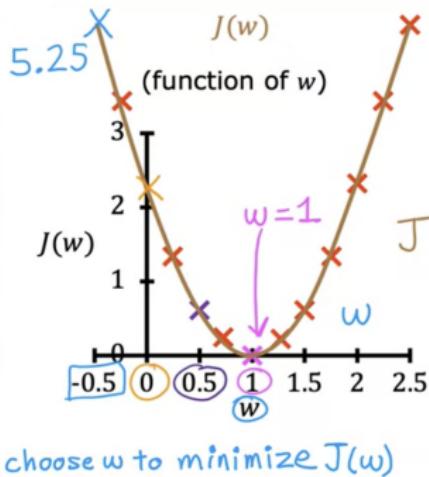
$$\begin{aligned} J(1) &= \frac{1}{2m} \sum_{i=1}^m (f_w(x^{(i)}) - y^{(i)})^2 \\ &= \frac{1}{2(3)} \sum_{i=1}^3 (1 \cdot x^{(i)}) - y^{(i)})^2 \\ &= \frac{1}{2(3)} \sum_{i=1}^3 (x^{(i)} - y^{(i)})^2 \\ &= \frac{1}{6} [(1 - 1)^2 + (2 - 2)^2 + (3 - 3)^2] \\ &= \frac{1}{6} [0^2 + 0^2 + 0^2] \\ &= 0 \end{aligned}$$

Cost Function Selection

We can plot these values of the costs function, $J(w)$ against w :

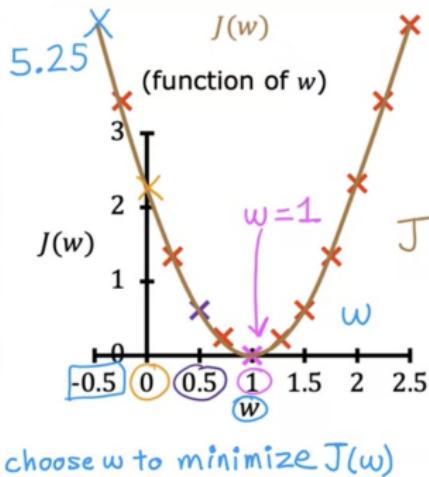
Cost Function Selection

We can plot these values of the costs function, $J(w)$ against w :



Cost Function Selection

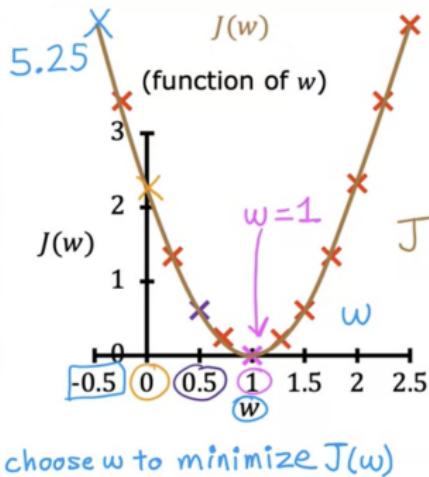
We can plot these values of the costs function, $J(w)$ against w :



- We see that $J(w)$ forms a convex parabola,

Cost Function Selection

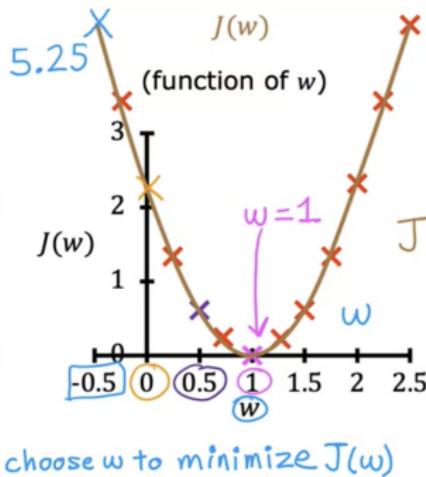
We can plot these values of the costs function, $J(w)$ against w :



- We see that $J(w)$ forms a convex parabola, with a global minimum at $w = 1$.

Cost Function Selection

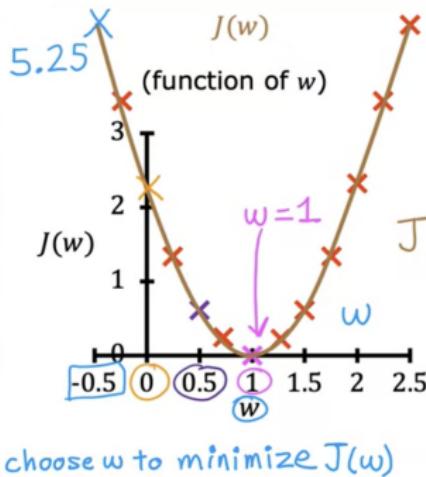
We can plot these values of the costs function, $J(w)$ against w :



- We see that $J(w)$ forms a convex parabola, with a global minimum at $w = 1$.
- It would be nice to have a computationally efficient algorithm for computing the global min...

Cost Function Selection

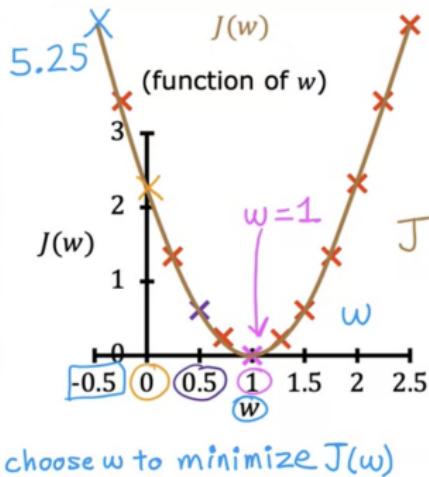
We can plot these values of the costs function, $J(w)$ against w :



- We see that $J(w)$ forms a convex parabola, with a global minimum at $w = 1$.
- It would be nice to have a computationally efficient algorithm for computing the global min... We do:

Cost Function Selection

We can plot these values of the costs function, $J(w)$ against w :



- We see that $J(w)$ forms a convex parabola, with a global minimum at $w = 1$.
- It would be nice to have a computationally efficient algorithm for computing the global min... We do: **Gradient Descent**.

Table of Contents

- 1 Overview of Machine Learning
 - Definition of ML
 - Types of ML
 - ML Architecture
- 2 Example - House Price Prediction
 - House Price Prediction - Training Set and Notation
- 3 Model Development - Part 1 - Model, Cost Function Selection
 - Methodology
 - Model Selection
 - Cost Function Selection
- 4 Model Development - Part 2 - Model Training
 - Review of The Chain Rule
 - Gradient Descent - Simplified Model
 - Gradient Descent - Full Model
 - Training the House Price Prediction Model

Table of Contents

1 Overview of Machine Learning

- Definition of ML
- Types of ML
- ML Architecture

2 Example - House Price Prediction

- House Price Prediction - Training Set and Notation

3 Model Development - Part 1 - Model, Cost Function Selection

- Methodology
- Model Selection
- Cost Function Selection

4 Model Development - Part 2 - Model Training

- Review of The Chain Rule
- Gradient Descent - Simplified Model
- Gradient Descent - Full Model
- Training the House Price Prediction Model

Review: Differential Calculus - The Chain Rule

Recall the following regarding the Chain Rule for determining the derivative of a real-valued function, $f(x)$:

- Suppose we have two real-valued functions, $f(x)$ and $g(x)$.

Review: Differential Calculus - The Chain Rule

Recall the following regarding the Chain Rule for determining the derivative of a real-valued function, $f(x)$:

- Suppose we have two real-valued functions, $f(x)$ and $g(x)$.
- g is differentiable at x and f is differentiable at the point $g(x)$.

Review: Differential Calculus - The Chain Rule

Recall the following regarding the Chain Rule for determining the derivative of a real-valued function, $f(x)$:

- Suppose we have two real-valued functions, $f(x)$ and $g(x)$.
- g is differentiable at x and f is differentiable at the point $g(x)$.
- We denote the derivative of f as f' and that of g as g' .

Review: Differential Calculus - The Chain Rule

Recall the following regarding the Chain Rule for determining the derivative of a real-valued function, $f(x)$:

- Suppose we have two real-valued functions, $f(x)$ and $g(x)$.
- g is differentiable at x and f is differentiable at the point $g(x)$.
- We denote the derivative of f as f' and that of g as g' .

Then, we have:

Review: Differential Calculus - The Chain Rule

Recall the following regarding the Chain Rule for determining the derivative of a real-valued function, $f(x)$:

- Suppose we have two real-valued functions, $f(x)$ and $g(x)$.
- g is differentiable at x and f is differentiable at the point $g(x)$.
- We denote the derivative of f as f' and that of g as g' .

Then, we have:

$$f(g(x))' = f'(g(x))g'(x)$$

Review: Differential Calculus - The Chain Rule

We can re-express the Chain Rule in terms of Leibniz notation (which is sort of easier):

Review: Differential Calculus - The Chain Rule

We can re-express the Chain Rule in terms of Leibniz notation (which is sort of easier):

- Let $y = f(u)$ and $u = g(x)$.

Review: Differential Calculus - The Chain Rule

We can re-express the Chain Rule in terms of Leibniz notation (which is sort of easier):

- Let $y = f(u)$ and $u = g(x)$.
- Then $y = f(g(x))$.

Review: Differential Calculus - The Chain Rule

We can re-express the Chain Rule in terms of Leibniz notation (which is sort of easier):

- Let $y = f(u)$ and $u = g(x)$.
- Then $y = f(g(x))$.
- And we have:

Review: Differential Calculus - The Chain Rule

We can re-express the Chain Rule in terms of Leibniz notation (which is sort of easier):

- Let $y = f(u)$ and $u = g(x)$.
- Then $y = f(g(x))$.
- And we have:

$$\frac{dy}{du} = f'(u) = f'(g(x)) \quad \text{and} \quad \frac{du}{dx} = g'(x)$$

Review: Differential Calculus - The Chain Rule

We can re-express the Chain Rule in terms of Leibniz notation (which is sort of easier):

- Let $y = f(u)$ and $u = g(x)$.
- Then $y = f(g(x))$.
- And we have:

$$\frac{dy}{du} = f'(u) = f'(g(x)) \quad \text{and} \quad \frac{du}{dx} = g'(x)$$

- This gives us:

Review: Differential Calculus - The Chain Rule

We can re-express the Chain Rule in terms of Leibniz notation (which is sort of easier):

- Let $y = f(u)$ and $u = g(x)$.
- Then $y = f(g(x))$.
- And we have:

$$\frac{dy}{du} = f'(u) = f'(g(x)) \quad \text{and} \quad \frac{du}{dx} = g'(x)$$

- This gives us:

$$\frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx}$$

Review: Differential Calculus - The Chain Rule

$$\frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx}$$

Review: Differential Calculus - The Chain Rule

$$\frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx}$$

Interpretation:

Review: Differential Calculus - The Chain Rule

$$\frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx}$$

Interpretation:

- 1 As x changes, u changes $\frac{du}{dx}$ times as fast as x .

Review: Differential Calculus - The Chain Rule

$$\frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx}$$

Interpretation:

- ① As x changes, u changes $\frac{du}{dx}$ times as fast as x .
- ② And y changes $\frac{dy}{du}$ times as fast as u .

Review: Differential Calculus - The Chain Rule

$$\frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx}$$

Interpretation:

- ① As x changes, u changes $\frac{du}{dx}$ times as fast as x .
- ② And y changes $\frac{dy}{du}$ times as fast as u .
- ③ Thus,

Review: Differential Calculus - The Chain Rule

$$\frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx}$$

Interpretation:

- ① As x changes, u changes $\frac{du}{dx}$ times as fast as x .
- ② And y changes $\frac{dy}{du}$ times as fast as u .
- ③ Thus, y changes $(\frac{dy}{du})(\frac{du}{dx})$ times as fast as x .

Review: Differential Calculus - The Chain Rule

Example 1 - The Chain Rule

Suppose $y = \sqrt{3x - 2}$.

Review: Differential Calculus - The Chain Rule

Example 1 - The Chain Rule

Suppose $y = \sqrt{3x - 2}$. Find $\frac{dy}{dx}$.

Review: Differential Calculus - The Chain Rule

Example 1 - The Chain Rule

Suppose $y = \sqrt{3x - 2}$. Find $\frac{dy}{dx}$.

Solution:

Review: Differential Calculus - The Chain Rule

Example 1 - The Chain Rule

Suppose $y = \sqrt{3x - 2}$. Find $\frac{dy}{dx}$.

Solution: Let $u = 3x - 2$.

Review: Differential Calculus - The Chain Rule

Example 1 - The Chain Rule

Suppose $y = \sqrt{3x - 2}$. Find $\frac{dy}{dx}$.

Solution: Let $u = 3x - 2$. Then $y = \sqrt{u}$.

Review: Differential Calculus - The Chain Rule

Example 1 - The Chain Rule

Suppose $y = \sqrt{3x - 2}$. Find $\frac{dy}{dx}$.

Solution: Let $u = 3x - 2$. Then $y = \sqrt{u}$.

$$\frac{dy}{du} = \frac{1}{2}u^{-1/2} = \frac{1}{2\sqrt{u}}$$

Review: Differential Calculus - The Chain Rule

Example 1 - The Chain Rule

Suppose $y = \sqrt{3x - 2}$. Find $\frac{dy}{dx}$.

Solution: Let $u = 3x - 2$. Then $y = \sqrt{u}$.

$$\frac{dy}{du} = \frac{1}{2}u^{-1/2} = \frac{1}{2\sqrt{u}}$$

$$\frac{du}{dx} = 3$$

Review: Differential Calculus - The Chain Rule

Example 1 - The Chain Rule

Suppose $y = \sqrt{3x - 2}$. Find $\frac{dy}{dx}$.

Solution: Let $u = 3x - 2$. Then $y = \sqrt{u}$.

$$\frac{dy}{du} = \frac{1}{2}u^{-1/2} = \frac{1}{2\sqrt{u}}$$

$$\frac{du}{dx} = 3$$

$$\frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx} = \frac{1}{2\sqrt{u}} \cdot 3 = \frac{3}{2\sqrt{3x - 2}}$$

Review: Differential Calculus - The Chain Rule

Example 2 - The Chain Rule

Suppose $y = (5x^2 + 8x)^2$.

Review: Differential Calculus - The Chain Rule

Example 2 - The Chain Rule

Suppose $y = (5x^2 + 8x)^2$. Find $\frac{dy}{dx}$.

Review: Differential Calculus - The Chain Rule

Example 2 - The Chain Rule

Suppose $y = (5x^2 + 8x)^2$. Find $\frac{dy}{dx}$.

Solution:

Review: Differential Calculus - The Chain Rule

Example 2 - The Chain Rule

Suppose $y = (5x^2 + 8x)^2$. Find $\frac{dy}{dx}$.

Solution: Let $u = 5x^2 + 8x$.

Review: Differential Calculus - The Chain Rule

Example 2 - The Chain Rule

Suppose $y = (5x^2 + 8x)^2$. Find $\frac{dy}{dx}$.

Solution: Let $u = 5x^2 + 8x$. Then $y = u^2$.

Review: Differential Calculus - The Chain Rule

Example 2 - The Chain Rule

Suppose $y = (5x^2 + 8x)^2$. Find $\frac{dy}{dx}$.

Solution: Let $u = 5x^2 + 8x$. Then $y = u^2$.

$$\frac{dy}{du} = 2u$$

Review: Differential Calculus - The Chain Rule

Example 2 - The Chain Rule

Suppose $y = (5x^2 + 8x)^2$. Find $\frac{dy}{dx}$.

Solution: Let $u = 5x^2 + 8x$. Then $y = u^2$.

$$\frac{dy}{du} = 2u$$

$$\frac{du}{dx} = 5(2x) + 8 = 10x + 8$$

Review: Differential Calculus - The Chain Rule

Example 2 - The Chain Rule

Suppose $y = (5x^2 + 8x)^2$. Find $\frac{dy}{dx}$.

Solution: Let $u = 5x^2 + 8x$. Then $y = u^2$.

$$\frac{dy}{du} = 2u$$

$$\frac{du}{dx} = 5(2x) + 8 = 10x + 8$$

$$\frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx} = 2u(10x + 8) = 2(5x^2 + 8x)(10x + 8)$$

Table of Contents

1 Overview of Machine Learning

- Definition of ML
- Types of ML
- ML Architecture

2 Example - House Price Prediction

- House Price Prediction - Training Set and Notation

3 Model Development - Part 1 - Model, Cost Function Selection

- Methodology
- Model Selection
- Cost Function Selection

4 Model Development - Part 2 - Model Training

- Review of The Chain Rule
- Gradient Descent - Simplified Model
- Gradient Descent - Full Model
- Training the House Price Prediction Model

Gradient Descent - Simplified Model

$$f_w(x) = wx$$

Gradient Descent - Simplified Model

$$f_w(x) = wx$$

Gradient Descent for $f_w(x)$:

Gradient Descent - Simplified Model

$$f_w(x) = wx$$

Gradient Descent for $f_w(x)$:

- ➊ Initialize w to a random value.

Gradient Descent - Simplified Model

$$f_w(x) = wx$$

Gradient Descent for $f_w(x)$:

- ➊ Initialize w to a random value.
- ➋ Repeat until convergence:

Gradient Descent - Simplified Model

$$f_w(x) = wx$$

Gradient Descent for $f_w(x)$:

- ➊ Initialize w to a random value.
- ➋ Repeat until convergence:

$$w := w - \alpha \frac{\partial J(w)}{\partial w}$$

Gradient Descent - Simplified Model

$$f_w(x) = wx$$

Gradient Descent for $f_w(x)$:

- ➊ Initialize w to a random value.
- ➋ Repeat until convergence:

$$w := w - \alpha \frac{\partial J(w)}{\partial w}$$

where $\alpha \in \mathbb{R} > 0$ is a learning rate

Gradient Descent - Simplified Model

$$f_w(x) = wx$$

Gradient Descent for $f_w(x)$:

- ➊ Initialize w to a random value.
- ➋ Repeat until convergence:

$$w := w - \alpha \frac{\partial J(w)}{\partial w}$$

where $\alpha \in \mathbb{R} > 0$ is a learning rate - we must assign this value judiciously (more on this later).

Gradient Descent - Simplified Model

Finding $\frac{\partial J(w)}{\partial w}$:

Gradient Descent - Simplified Model

Finding $\frac{\partial J(w)}{\partial w}$:

$$J(w) = \frac{1}{2m} \sum_{i=1}^m (f_w(x^{(i)}) - y^{(i)})^2$$

Gradient Descent - Simplified Model

Finding $\frac{\partial J(w)}{\partial w}$:

$$\begin{aligned}J(w) &= \frac{1}{2m} \sum_{i=1}^m (f_w(x^{(i)}) - y^{(i)})^2 \\&= \frac{1}{2m} \sum_{i=1}^m (wx^{(i)} - y^{(i)})^2\end{aligned}$$

Gradient Descent - Simplified Model

Finding $\frac{\partial J(w)}{\partial w}$:

$$\begin{aligned}J(w) &= \frac{1}{2m} \sum_{i=1}^m (f_w(x^{(i)}) - y^{(i)})^2 \\&= \frac{1}{2m} \sum_{i=1}^m (wx^{(i)} - y^{(i)})^2 \\&= \frac{1}{2m} \left((wx^{(1)} - y^{(1)})^2 + (wx^{(2)} - y^{(2)})^2 + \dots \right. \\&\quad \left. + (wx^{(m)} - y^{(m)})^2 \right)\end{aligned}$$

Gradient Descent - Simplified Model

Finding $\frac{\partial J(w)}{\partial w}$.

Gradient Descent - Simplified Model

Finding $\frac{\partial J(w)}{\partial w}$:

$$J(w) = \frac{1}{2m} \left((wx^{(1)} - y^{(1)})^2 + (wx^{(2)} - y^{(2)})^2 + \dots + (wx^{(m)} - y^{(m)})^2 \right)$$

Gradient Descent - Simplified Model

Finding $\frac{\partial J(w)}{\partial w}$:

$$J(w) = \frac{1}{2m} \left((wx^{(1)} - y^{(1)})^2 + (wx^{(2)} - y^{(2)})^2 + \dots + (wx^{(m)} - y^{(m)})^2 \right)$$

Let's take just one term of this sum:

Gradient Descent - Simplified Model

Finding $\frac{\partial J(w)}{\partial w}$:

$$J(w) = \frac{1}{2m} \left((wx^{(1)} - y^{(1)})^2 + (wx^{(2)} - y^{(2)})^2 + \dots + (wx^{(m)} - y^{(m)})^2 \right)$$

Let's take just one term of this sum:

$$J_1(w) = (wx^{(1)} - y^{(1)})^2$$

Gradient Descent - Simplified Model

Finding $\frac{\partial J(w)}{\partial w}$:

$$J(w) = \frac{1}{2m} \left((wx^{(1)} - y^{(1)})^2 + (wx^{(2)} - y^{(2)})^2 + \dots + (wx^{(m)} - y^{(m)})^2 \right)$$

Let's take just one term of this sum:

$$J_1(w) = (wx^{(1)} - y^{(1)})^2$$

Let's find $\frac{\partial J_1(w)}{\partial w}$.

Gradient Descent - Simplified Model

Let $J_1(w) = (wx^{(1)} - y^{(1)})^2$.

Gradient Descent - Simplified Model

Let $J_1(w) = (wx^{(1)} - y^{(1)})^2$. Find $\frac{\partial J_1(w)}{\partial w}$:

Gradient Descent - Simplified Model

Let $J_1(w) = (wx^{(1)} - y^{(1)})^2$. Find $\frac{\partial J_1(w)}{\partial w}$:

Use the Chain Rule.

Gradient Descent - Simplified Model

Let $J_1(w) = (wx^{(1)} - y^{(1)})^2$. Find $\frac{\partial J_1(w)}{\partial w}$:

Use the Chain Rule. Let $u = wx^{(1)} - y^{(1)}$.

Gradient Descent - Simplified Model

Let $J_1(w) = (wx^{(1)} - y^{(1)})^2$. Find $\frac{\partial J_1(w)}{\partial w}$:

Use the Chain Rule. Let $u = wx^{(1)} - y^{(1)}$. Then, $y = u^2 = J_1(w)$.

Gradient Descent - Simplified Model

Let $J_1(w) = (wx^{(1)} - y^{(1)})^2$. Find $\frac{\partial J_1(w)}{\partial w}$:

Use the Chain Rule. Let $u = wx^{(1)} - y^{(1)}$. Then, $y = u^2 = J_1(w)$.

$$\frac{dy}{du} = 2u$$

$$\frac{du}{dw} = x^{(1)}$$

$$\frac{dy}{dw} = \frac{dy}{du} \cdot \frac{du}{dw} = 2u \cdot x^{(1)} = 2(wx^{(1)} - y^{(1)})x^{(1)}$$

Gradient Descent - Simplified Model

Let $J_1(w) = (wx^{(1)} - y^{(1)})^2$. Find $\frac{\partial J_1(w)}{\partial w}$:

Use the Chain Rule. Let $u = wx^{(1)} - y^{(1)}$. Then, $y = u^2 = J_1(w)$.

$$\frac{dy}{du} = 2u$$

$$\frac{du}{dw} = x^{(1)}$$

$$\frac{dy}{dw} = \frac{dy}{du} \cdot \frac{du}{dw} = 2u \cdot x^{(1)} = 2(wx^{(1)} - y^{(1)})x^{(1)}$$

$$\therefore \frac{\partial J_1(w)}{\partial w} = \frac{dy}{dw} = 2(wx^{(1)} - y^{(1)})x^{(1)}$$

Gradient Descent - Simplified Model

In general, let's define:

Gradient Descent - Simplified Model

In general, let's define:

$$J_i(w) \equiv (wx^{(i)} - y^{(i)})^2 \quad i = 1 \dots m$$

Gradient Descent - Simplified Model

In general, let's define:

$$J_i(w) \equiv (wx^{(i)} - y^{(i)})^2 \quad i = 1 \dots m$$

Then, we have:

Gradient Descent - Simplified Model

In general, let's define:

$$J_i(w) \equiv (wx^{(i)} - y^{(i)})^2 \quad i = 1 \dots m$$

Then, we have:

$$\frac{\partial J_i(w)}{\partial w} = 2(wx^{(i)} - y^{(i)})x^{(i)} \quad i = 1 \dots m$$

Gradient Descent - Simplified Model

Now, we return to finding $\frac{\partial J(w)}{\partial w}$:

Gradient Descent - Simplified Model

Now, we return to finding $\frac{\partial J(w)}{\partial w}$:

$$J(w) = \frac{1}{2m} \sum_{i=1}^m (f_w(x^{(i)}) - y^{(i)})^2$$

Gradient Descent - Simplified Model

Now, we return to finding $\frac{\partial J(w)}{\partial w}$:

$$\begin{aligned}J(w) &= \frac{1}{2m} \sum_{i=1}^m (f_w(x^{(i)}) - y^{(i)})^2 \\&= \frac{1}{2m} \left((wx^{(1)} - y^{(1)})^2 + (wx^{(2)} - y^{(2)})^2 + \dots \right. \\&\quad \left. + (wx^{(m)} - y^{(m)})^2 \right)\end{aligned}$$

Gradient Descent - Simplified Model

Now, we return to finding $\frac{\partial J(w)}{\partial w}$:

$$\begin{aligned} J(w) &= \frac{1}{2m} \sum_{i=1}^m (f_w(x^{(i)}) - y^{(i)})^2 \\ &= \frac{1}{2m} \left((wx^{(1)} - y^{(1)})^2 + (wx^{(2)} - y^{(2)})^2 + \dots \right. \\ &\quad \left. + (wx^{(m)} - y^{(m)})^2 \right) \\ &= \frac{1}{2m} \left(J_1(w) + J_2(w) + \dots + J_m(w) \right) \end{aligned}$$

Gradient Descent - Simplified Model

Now, we return to finding $\frac{\partial J(w)}{\partial w}$:

$$\begin{aligned} J(w) &= \frac{1}{2m} \sum_{i=1}^m (f_w(x^{(i)}) - y^{(i)})^2 \\ &= \frac{1}{2m} \left((wx^{(1)} - y^{(1)})^2 + (wx^{(2)} - y^{(2)})^2 + \dots \right. \\ &\quad \left. + (wx^{(m)} - y^{(m)})^2 \right) \\ &= \frac{1}{2m} \left(J_1(w) + J_2(w) + \dots + J_m(w) \right) \\ \therefore \frac{\partial J(w)}{\partial w} &= \frac{\partial}{\partial w} \left[\frac{1}{2m} \left(J_1(w) + J_2(w) + \dots + J_m(w) \right) \right] \end{aligned}$$

Gradient Descent - Simplified Model

Now, we return to finding $\frac{\partial J(w)}{\partial w}$:

$$\begin{aligned}J(w) &= \frac{1}{2m} \sum_{i=1}^m (f_w(x^{(i)}) - y^{(i)})^2 \\&= \frac{1}{2m} \left((wx^{(1)} - y^{(1)})^2 + (wx^{(2)} - y^{(2)})^2 + \dots \right. \\&\quad \left. + (wx^{(m)} - y^{(m)})^2 \right) \\&= \frac{1}{2m} \left(J_1(w) + J_2(w) + \dots + J_m(w) \right)\end{aligned}$$

$$\begin{aligned}\therefore \frac{\partial J(w)}{\partial w} &= \frac{\partial}{\partial w} \left[\frac{1}{2m} \left(J_1(w) + J_2(w) + \dots + J_m(w) \right) \right] \\&= \frac{1}{2m} \left(\frac{\partial J_1(w)}{\partial w} + \frac{\partial J_2(w)}{\partial w} + \dots + \frac{\partial J_m(w)}{\partial w} \right)\end{aligned}$$

Gradient Descent - Simplified Model

From the prior slide, we have:

Gradient Descent - Simplified Model

From the prior slide, we have:

$$\frac{\partial J(w)}{\partial w} = \frac{1}{2m} \left(\frac{\partial J_1(w)}{\partial w} + \frac{\partial J_2(w)}{\partial w} + \dots + \frac{\partial J_m(w)}{\partial w} \right)$$

Gradient Descent - Simplified Model

From the prior slide, we have:

$$= \frac{1}{2m} \left(2(wx^{(1)} - y^{(1)})x^{(1)} + 2(wx^{(2)} - y^{(1)})x^{(2)} + \dots + 2(wx^{(m)} - y^{(m)})x^{(m)} \right)$$

Gradient Descent - Simplified Model

From the prior slide, we have:

$$\begin{aligned} &= \frac{1}{2m} \left(2(wx^{(1)} - y^{(1)})x^{(1)} + 2(wx^{(2)} - y^{(1)})x^{(2)} + \dots \right. \\ &\quad \left. + 2(wx^{(m)} - y^{(m)})x^{(m)} \right) \\ &= \frac{2}{2m} \left((wx^{(1)} - y^{(1)})x^{(1)} + (wx^{(2)} - y^{(1)})x^{(2)} + \dots \right. \\ &\quad \left. + (wx^{(m)} - y^{(m)})x^{(m)} \right) \end{aligned}$$

Gradient Descent - Simplified Model

From the prior slide, we have:

$$\begin{aligned} &= \frac{1}{2m} \left(2(wx^{(1)} - y^{(1)})x^{(1)} + 2(wx^{(2)} - y^{(1)})x^{(2)} + \dots \right. \\ &\quad \left. + 2(wx^{(m)} - y^{(m)})x^{(m)} \right) \\ &= \frac{2}{2m} \left((wx^{(1)} - y^{(1)})x^{(1)} + (wx^{(2)} - y^{(1)})x^{(2)} + \dots \right. \\ &\quad \left. + (wx^{(m)} - y^{(m)})x^{(m)} \right) \\ &= \frac{1}{m} \left((wx^{(1)} - y^{(1)})x^{(1)} + (wx^{(2)} - y^{(1)})x^{(2)} + \dots \right. \\ &\quad \left. + (wx^{(m)} - y^{(m)})x^{(m)} \right) \end{aligned}$$

Gradient Descent - Simplified Model

From the prior slide, we have:

Gradient Descent - Simplified Model

From the prior slide, we have:

$$\begin{aligned}\frac{\partial J(w)}{\partial w} = \frac{1}{m} & \left((wx^{(1)} - y^{(1)})x^{(1)} + (wx^{(2)} - y^{(2)})x^{(2)} + \dots \right. \\ & \left. + (wx^{(m)} - y^{(m)})x^{(m)} \right)\end{aligned}$$

Gradient Descent - Simplified Model

From the prior slide, we have:

$$\frac{\partial J(w)}{\partial w} = \frac{1}{m} \left((wx^{(1)} - y^{(1)})x^{(1)} + (wx^{(2)} - y^{(2)})x^{(2)} + \dots + (wx^{(m)} - y^{(m)})x^{(m)} \right)$$

$$\therefore \frac{\partial J(w)}{\partial w} = \frac{1}{m} \sum_{i=1}^m (wx^{(i)} - y^{(i)})x^{(i)}$$

Gradient Descent - Simplified Model

So, our Gradient Descent algorithm for $f_w(x)$:

Gradient Descent - Simplified Model

So, our Gradient Descent algorithm for $f_w(x)$:

- ➊ Initialize w to a random value.

Gradient Descent - Simplified Model

So, our Gradient Descent algorithm for $f_w(x)$:

- ① Initialize w to a random value.
- ② Repeat until convergence:

Gradient Descent - Simplified Model

So, our Gradient Descent algorithm for $f_w(x)$:

- ➊ Initialize w to a random value.
- ➋ Repeat until convergence:

$$w := w - \alpha \frac{1}{m} \sum_{i=1}^m (wx^{(i)} - y^{(i)})x^{(i)}$$

Gradient Descent - Simplified Model

Gradient Descent:

Gradient Descent - Simplified Model

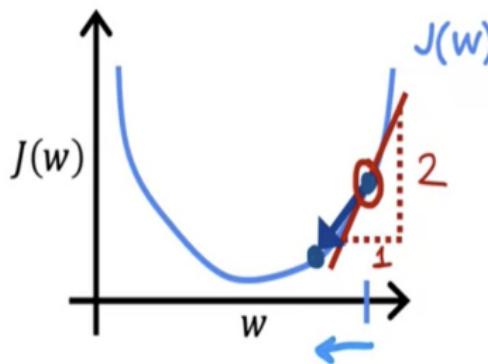
Gradient Descent:

- Consider when w is larger than the minimum:

Gradient Descent - Simplified Model

Gradient Descent:

- Consider when w is larger than the minimum:



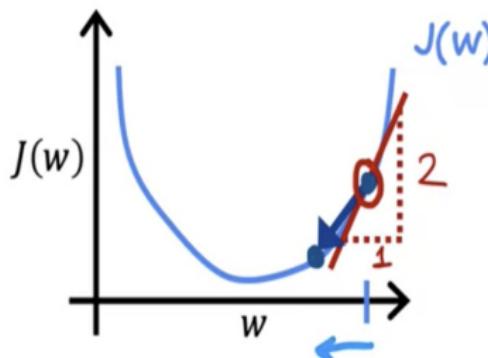
$$w = w - \alpha \frac{\frac{d}{dw} J(w)}{> 0}$$

$w = w - \underline{\alpha} \cdot (\text{positive number})$

Gradient Descent - Simplified Model

Gradient Descent:

- Consider when w is larger than the minimum:



$$w = w - \alpha \frac{\frac{d}{dw} J(w)}{> 0}$$

$$w = w - \underline{\alpha} \cdot (\text{positive number})$$

- The parameter w is adjusted downward toward the value that minimizes $J(w)$.

Gradient Descent - Simplified Model

Gradient Descent:

Gradient Descent - Simplified Model

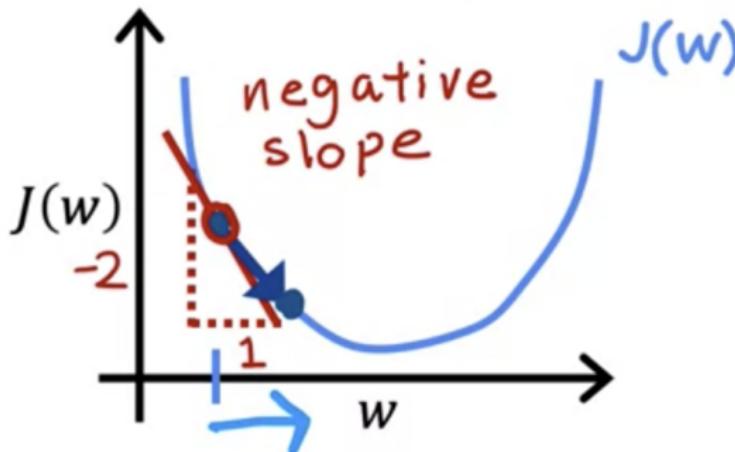
Gradient Descent:

- Consider when w is smaller than the minimum:

Gradient Descent - Simplified Model

Gradient Descent:

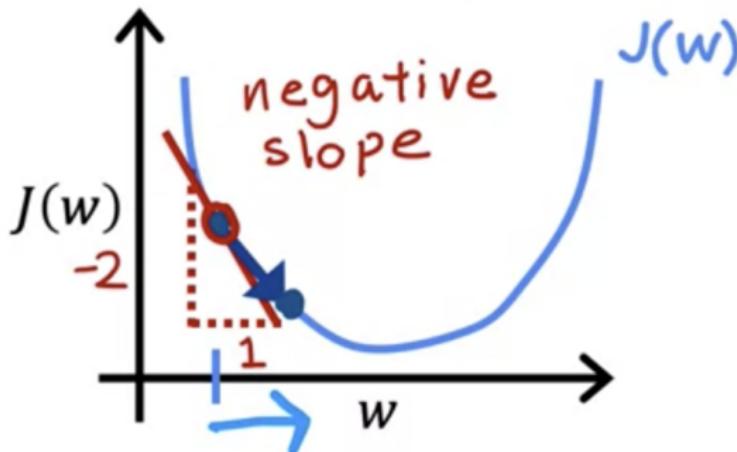
- Consider when w is smaller than the minimum:



Gradient Descent - Simplified Model

Gradient Descent:

- Consider when w is smaller than the minimum:



- The parameter w is adjusted upward toward the value that minimizes $J(w)$.

Gradient Descent - Learning Rate

Learning Rate, α :

Gradient Descent - Learning Rate

Learning Rate, α :

- The adjustment formula for parameter w is: $w := w - \alpha \frac{\partial J(w)}{\partial w}$.

Gradient Descent - Learning Rate

Learning Rate, α :

- The adjustment formula for parameter w is: $w := w - \alpha \frac{\partial J(w)}{\partial w}$.
- If α is too small,

Gradient Descent - Learning Rate

Learning Rate, α :

- The adjustment formula for parameter w is: $w := w - \alpha \frac{\partial J(w)}{\partial w}$.
- If α is too small, the algorithm converges very slowly, causing very long runtimes.

Gradient Descent - Learning Rate

Learning Rate, α :

- The adjustment formula for parameter w is: $w := w - \alpha \frac{\partial J(w)}{\partial w}$.
- If α is too small, the algorithm converges very slowly, causing very long runtimes.
- If α is too large,

Gradient Descent - Learning Rate

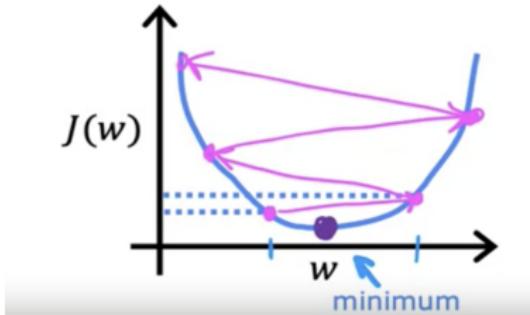
Learning Rate, α :

- The adjustment formula for parameter w is: $w := w - \alpha \frac{\partial J(w)}{\partial w}$.
- If α is too small, the algorithm converges very slowly, causing very long runtimes.
- If α is too large, the algorithm may not converge at all,

Gradient Descent - Learning Rate

Learning Rate, α :

- The adjustment formula for parameter w is: $w := w - \alpha \frac{\partial J(w)}{\partial w}$.
- If α is too small, the algorithm converges very slowly, causing very long runtimes.
- If α is too large, the algorithm may not converge at all, i.e., we may have divergence.



Gradient Descent - Learning Rate

More on the Learning Rate, α :

Gradient Descent - Learning Rate

More on the Learning Rate, α :

- Question:

Gradient Descent - Learning Rate

More on the Learning Rate, α :

- Question: Do we need to shrink α as we get close to the minimum

Gradient Descent - Learning Rate

More on the Learning Rate, α :

- Question: Do we need to shrink α as we get close to the minimum to make sure we don't have divergence

Gradient Descent - Learning Rate

More on the Learning Rate, α :

- Question: Do we need to shrink α as we get close to the minimum to make sure we don't have divergence (even when we have a judicious choice for α)?

Gradient Descent - Learning Rate

More on the Learning Rate, α :

- Question: Do we need to shrink α as we get close to the minimum to make sure we don't have divergence (even when we have a judicious choice for α)?
- Answer:

Gradient Descent - Learning Rate

More on the Learning Rate, α :

- Question: Do we need to shrink α as we get close to the minimum to make sure we don't have divergence (even when we have a judicious choice for α)?
- Answer: No.

Gradient Descent - Learning Rate

More on the Learning Rate, α :

- Question: Do we need to shrink α as we get close to the minimum to make sure we don't have divergence (even when we have a judicious choice for α)?
- Answer: No. As we approach a minimum,

Gradient Descent - Learning Rate

More on the Learning Rate, α :

- Question: Do we need to shrink α as we get close to the minimum to make sure we don't have divergence (even when we have a judicious choice for α)?
- Answer: No. As we approach a minimum, the partial derivative of $J(w)$ will get smaller and smaller,

Gradient Descent - Learning Rate

More on the Learning Rate, α :

- Question: Do we need to shrink α as we get close to the minimum to make sure we don't have divergence (even when we have a judicious choice for α)?
- Answer: No. As we approach a minimum, the partial derivative of $J(w)$ will get smaller and smaller, approaching 0,

Gradient Descent - Learning Rate

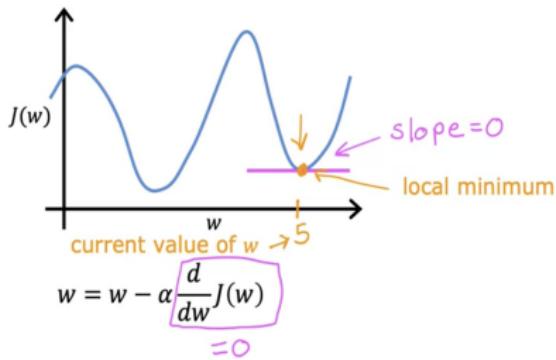
More on the Learning Rate, α :

- Question: Do we need to shrink α as we get close to the minimum to make sure we don't have divergence (even when we have a judicious choice for α)?
- Answer: No. As we approach a minimum, the partial derivative of $J(w)$ will get smaller and smaller, approaching 0, and thus, make our adjustment at each step, $\alpha \frac{\partial J(w)}{\partial w}$ very small.

Gradient Descent - Learning Rate

More on the Learning Rate, α :

- Question: Do we need to shrink α as we get close to the minimum to make sure we don't have divergence (even when we have a judicious choice for α)?
- Answer: No. As we approach a minimum, the partial derivative of $J(w)$ will get smaller and smaller, approaching 0, and thus, make our adjustment at each step, $\alpha \frac{\partial J(w)}{\partial w}$ very small.



Gradient Descent - Simplified Model

More on the Learning Rate, α : (contd.)

Gradient Descent - Simplified Model

More on the Learning Rate, α : (contd.)

- Thus, if α is initialized properly,

Gradient Descent - Simplified Model

More on the Learning Rate, α : (contd.)

- Thus, if α is initialized properly, it can and *should* remain fixed

Gradient Descent - Simplified Model

More on the Learning Rate, α : (contd.)

- Thus, if α is initialized properly, it can and *should* remain fixed throughout the execution of the gradient descent algorithm.

Gradient Descent - Simplified Model

More on the Learning Rate, α : (contd.)

- Thus, if α is initialized properly, it can and *should* remain fixed throughout the execution of the gradient descent algorithm.

Can reach local minimum with fixed learning rate α

$$w = w - \alpha \frac{d}{dw} J(w)$$

smaller
not as large
large

Near a local minimum,
- Derivative becomes smaller
- Update steps become smaller

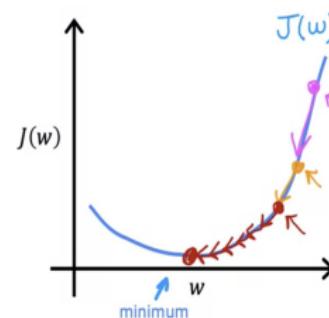


Table of Contents

1 Overview of Machine Learning

- Definition of ML
- Types of ML
- ML Architecture

2 Example - House Price Prediction

- House Price Prediction - Training Set and Notation

3 Model Development - Part 1 - Model, Cost Function Selection

- Methodology
- Model Selection
- Cost Function Selection

4 Model Development - Part 2 - Model Training

- Review of The Chain Rule
- Gradient Descent - Simplified Model
- Gradient Descent - Full Model
- Training the House Price Prediction Model

Gradient Descent - Full Model

$$f_{w,b}(x) = wx + b$$

Gradient Descent - Full Model

$$f_{w,b}(x) = wx + b$$

Gradient Descent for $f_{w,b}(x)$:

Gradient Descent - Full Model

$$f_{w,b}(x) = wx + b$$

Gradient Descent for $f_{w,b}(x)$:

- 1 Initialize w, b to random values.

Gradient Descent - Full Model

$$f_{w,b}(x) = wx + b$$

Gradient Descent for $f_{w,b}(x)$:

- ➊ Initialize w, b to random values.
- ➋ Repeat until convergence:

Gradient Descent - Full Model

$$f_{w,b}(x) = wx + b$$

Gradient Descent for $f_{w,b}(x)$:

- ➊ Initialize w, b to random values.
- ➋ Repeat until convergence:

$$w := w - \alpha \frac{\partial J(w, b)}{\partial w}$$
$$b := b - \alpha \frac{\partial J(w, b)}{\partial b}$$

Gradient Descent - Full Model

$$f_{w,b}(x) = wx + b$$

Gradient Descent for $f_{w,b}(x)$:

- ➊ Initialize w, b to random values.
- ➋ Repeat until convergence:

$$w := w - \alpha \frac{\partial J(w, b)}{\partial w}$$
$$b := b - \alpha \frac{\partial J(w, b)}{\partial b}$$

Note: For each step, updates to w and b should be made at the same time. That is, do not update any partial derivatives of J until both w and b have been adjusted for a given step.

Gradient Descent - Full Model

We need to find both $\frac{\partial J(w,b)}{\partial w}$ and $\frac{\partial J(w,b)}{\partial b}$.

Gradient Descent - Full Model

We need to find both $\frac{\partial J(w,b)}{\partial w}$ and $\frac{\partial J(w,b)}{\partial b}$. First, we find the partial with respect to w :

Gradient Descent - Full Model

We need to find both $\frac{\partial J(w,b)}{\partial w}$ and $\frac{\partial J(w,b)}{\partial b}$. First, we find the partial with respect to w :

$$J(w, b) = \frac{1}{2m} \sum_{i=1}^m (f_{w,b}(x^{(i)}) - y^{(i)})^2$$

Gradient Descent - Full Model

We need to find both $\frac{\partial J(w,b)}{\partial w}$ and $\frac{\partial J(w,b)}{\partial b}$. First, we find the partial with respect to w :

$$J(w, b) = \frac{1}{2m} \sum_{i=1}^m (f_{w,b}(x^{(i)}) - y^{(i)})^2$$

$$J(w, b) = \frac{1}{2m} \sum_{i=1}^m (wx^{(i)} + b - y^{(i)})^2$$

Gradient Descent - Introduction

Finding $\frac{\partial J(w, b)}{\partial w}$:

Gradient Descent - Introduction

Finding $\frac{\partial J(w, b)}{\partial w}$:

$$J(w, b) = \frac{1}{2m} \sum_{i=1}^m (wx^{(i)} + b - y^{(i)})^2$$

Gradient Descent - Introduction

Finding $\frac{\partial J(w, b)}{\partial w}$:

$$J(w, b) = \frac{1}{2m} \sum_{i=1}^m (wx^{(i)} + b - y^{(i)})^2$$

$$\begin{aligned} J(w, b) = \frac{1}{2m} & \left((wx^{(1)} + b - y^{(1)})^2 + (wx^{(2)} + b - y^{(2)})^2 + \dots \right. \\ & \left. + (wx^{(m)} + b - y^{(m)})^2 \right) \end{aligned}$$

Gradient Descent - Introduction

Finding $\frac{\partial J(w, b)}{\partial w}$:

$$J(w, b) = \frac{1}{2m} \sum_{i=1}^m (wx^{(i)} + b - y^{(i)})^2$$

$$\begin{aligned} J(w, b) = \frac{1}{2m} & \left((wx^{(1)} + b - y^{(1)})^2 + (wx^{(2)} + b - y^{(2)})^2 + \dots \right. \\ & \left. + (wx^{(m)} + b - y^{(m)})^2 \right) \end{aligned}$$

Let's take just one term of this sum:

Gradient Descent - Introduction

Finding $\frac{\partial J(w, b)}{\partial w}$:

$$J(w, b) = \frac{1}{2m} \sum_{i=1}^m (wx^{(i)} + b - y^{(i)})^2$$

$$\begin{aligned} J(w, b) = \frac{1}{2m} & \left((wx^{(1)} + b - y^{(1)})^2 + (wx^{(2)} + b - y^{(2)})^2 + \dots \right. \\ & \left. + (wx^{(m)} + b - y^{(m)})^2 \right) \end{aligned}$$

Let's take just one term of this sum:

$$J_1(w, b) = (wx^{(1)} + b - y^{(1)})^2$$

Gradient Descent - Introduction

Finding $\frac{\partial J(w, b)}{\partial w}$:

$$J(w, b) = \frac{1}{2m} \sum_{i=1}^m (wx^{(i)} + b - y^{(i)})^2$$

$$\begin{aligned} J(w, b) = \frac{1}{2m} & \left((wx^{(1)} + b - y^{(1)})^2 + (wx^{(2)} + b - y^{(2)})^2 + \dots \right. \\ & \left. + (wx^{(m)} + b - y^{(m)})^2 \right) \end{aligned}$$

Let's take just one term of this sum:

$$J_1(w, b) = (wx^{(1)} + b - y^{(1)})^2$$

Let's find $\frac{\partial J_1(w, b)}{\partial w}$.

Gradient Descent - Introduction

Let $J_1(w, b) \equiv (wx^{(1)} + b - y^{(1)})^2$.

Gradient Descent - Introduction

Let $J_1(w, b) \equiv (wx^{(1)} + b - y^{(1)})^2$. Find $\frac{\partial J_1(w, b)}{\partial w}$:

Gradient Descent - Introduction

Let $J_1(w, b) \equiv (wx^{(1)} + b - y^{(1)})^2$. Find $\frac{\partial J_1(w, b)}{\partial w}$:

Use the Chain Rule.

Gradient Descent - Introduction

Let $J_1(w, b) \equiv (wx^{(1)} + b - y^{(1)})^2$. Find $\frac{\partial J_1(w, b)}{\partial w}$:

Use the Chain Rule. Let $u = wx^{(1)} + b - y^{(1)}$.

Gradient Descent - Introduction

Let $J_1(w, b) \equiv (wx^{(1)} + b - y^{(1)})^2$. Find $\frac{\partial J_1(w, b)}{\partial w}$:

Use the Chain Rule. Let $u = wx^{(1)} + b - y^{(1)}$. Then, $y = u^2$.

Gradient Descent - Introduction

Let $J_1(w, b) \equiv (wx^{(1)} + b - y^{(1)})^2$. Find $\frac{\partial J_1(w, b)}{\partial w}$:

Use the Chain Rule. Let $u = wx^{(1)} + b - y^{(1)}$. Then, $y = u^2$.

$$\frac{dy}{du} = 2u$$

Gradient Descent - Introduction

Let $J_1(w, b) \equiv (wx^{(1)} + b - y^{(1)})^2$. Find $\frac{\partial J_1(w, b)}{\partial w}$:

Use the Chain Rule. Let $u = wx^{(1)} + b - y^{(1)}$. Then, $y = u^2$.

$$\frac{dy}{du} = 2u$$

$$\frac{du}{dw} = x^{(1)}$$

Gradient Descent - Introduction

Let $J_1(w, b) \equiv (wx^{(1)} + b - y^{(1)})^2$. Find $\frac{\partial J_1(w, b)}{\partial w}$:

Use the Chain Rule. Let $u = wx^{(1)} + b - y^{(1)}$. Then, $y = u^2$.

$$\frac{dy}{du} = 2u$$

$$\frac{du}{dw} = x^{(1)}$$

$$\frac{dy}{dw} = \frac{dy}{du} \cdot \frac{du}{dw} = 2u \cdot x^{(1)} = 2(wx^{(1)} + b - y^{(1)})x^{(1)}$$

Gradient Descent - Introduction

Let $J_1(w, b) \equiv (wx^{(1)} + b - y^{(1)})^2$. Find $\frac{\partial J_1(w, b)}{\partial w}$:

Use the Chain Rule. Let $u = wx^{(1)} + b - y^{(1)}$. Then, $y = u^2$.

$$\frac{dy}{du} = 2u$$

$$\frac{du}{dw} = x^{(1)}$$

$$\frac{dy}{dw} = \frac{dy}{du} \cdot \frac{du}{dw} = 2u \cdot x^{(1)} = 2(wx^{(1)} + b - y^{(1)})x^{(1)}$$

$$\therefore \frac{\partial J_1(w)}{\partial w} = 2(wx^{(1)} + b - y^{(1)})x^{(1)}$$

Gradient Descent - Introduction

In general, let's define:

Gradient Descent - Introduction

In general, let's define:

$$J_i(w, b) \equiv (wx^{(i)} + b - y^{(i)})^2 \quad i = 1 \dots m$$

Gradient Descent - Introduction

In general, let's define:

$$J_i(w, b) \equiv (wx^{(i)} + b - y^{(i)})^2 \quad i = 1 \dots m$$

Then, we have:

Gradient Descent - Introduction

In general, let's define:

$$J_i(w, b) \equiv (wx^{(i)} + b - y^{(i)})^2 \quad i = 1 \dots m$$

Then, we have:

$$\frac{\partial J_i(w, b)}{\partial w} = 2(wx^{(i)} + b - y^{(i)})x^{(i)} \quad i = 1 \dots m$$

Gradient Descent - Introduction

Now, we return to finding $\frac{\partial J(w, b)}{\partial w}$:

Gradient Descent - Introduction

Now, we return to finding $\frac{\partial J(w, b)}{\partial w}$:

$$J(w, b) = \frac{1}{2m} \sum_{i=1}^m (f_w(x^{(i)}) - y^{(i)})^2$$

Gradient Descent - Introduction

Now, we return to finding $\frac{\partial J(w, b)}{\partial w}$:

$$\begin{aligned} J(w, b) &= \frac{1}{2m} \sum_{i=1}^m (f_w(x^{(i)}) - y^{(i)})^2 \\ &= \frac{1}{2m} \left((wx^{(1)} + b - y^{(1)})^2 + (wx^{(2)} + b - y^{(2)})^2 + \dots \right. \\ &\quad \left. + (wx^{(m)} + b - y^{(m)})^2 \right) \end{aligned}$$

Gradient Descent - Introduction

Now, we return to finding $\frac{\partial J(w, b)}{\partial w}$:

$$\begin{aligned} J(w, b) &= \frac{1}{2m} \sum_{i=1}^m (f_w(x^{(i)}) - y^{(i)})^2 \\ &= \frac{1}{2m} \left((wx^{(1)} + b - y^{(1)})^2 + (wx^{(2)} + b - y^{(2)})^2 + \dots \right. \\ &\quad \left. + (wx^{(m)} + b - y^{(m)})^2 \right) \\ &= \frac{1}{2m} \left(J_1(w, b) + J_2(w, b) + \dots + J_m(w, b) \right) \end{aligned}$$

Gradient Descent - Introduction

Now, we return to finding $\frac{\partial J(w, b)}{\partial w}$:

$$\begin{aligned} J(w, b) &= \frac{1}{2m} \sum_{i=1}^m (f_w(x^{(i)}) - y^{(i)})^2 \\ &= \frac{1}{2m} \left((wx^{(1)} + b - y^{(1)})^2 + (wx^{(2)} + b - y^{(2)})^2 + \dots \right. \\ &\quad \left. + (wx^{(m)} + b - y^{(m)})^2 \right) \\ &= \frac{1}{2m} \left(J_1(w, b) + J_2(w, b) + \dots + J_m(w, b) \right) \\ \therefore \frac{\partial J(w, b)}{\partial w} &= \frac{\partial}{\partial w} \left[\frac{1}{2m} \left(J_1(w, b) + J_2(w, b) + \dots + J_m(w, b) \right) \right] \end{aligned}$$

Gradient Descent - Introduction

Now, we return to finding $\frac{\partial J(w, b)}{\partial w}$:

$$\begin{aligned} J(w, b) &= \frac{1}{2m} \sum_{i=1}^m (f_w(x^{(i)}) - y^{(i)})^2 \\ &= \frac{1}{2m} \left((wx^{(1)} + b - y^{(1)})^2 + (wx^{(2)} + b - y^{(2)})^2 + \dots \right. \\ &\quad \left. + (wx^{(m)} + b - y^{(m)})^2 \right) \\ &= \frac{1}{2m} \left(J_1(w, b) + J_2(w, b) + \dots + J_m(w, b) \right) \\ \therefore \frac{\partial J(w, b)}{\partial w} &= \frac{\partial}{\partial w} \left[\frac{1}{2m} \left(J_1(w, b) + J_2(w, b) + \dots + J_m(w, b) \right) \right] \\ &= \frac{1}{2m} \left(\frac{\partial J_1(w, b)}{\partial w} + \frac{\partial J_2(w, b)}{\partial w} + \dots + \frac{\partial J_m(w, b)}{\partial w} \right) \end{aligned}$$

Gradient Descent - Introduction

From the prior slide, we have:

Gradient Descent - Introduction

From the prior slide, we have:

$$\frac{\partial J(w, b)}{\partial w} = \frac{1}{2m} \left(\frac{\partial J_1(w, b)}{\partial w} + \frac{\partial J_2(w, b)}{\partial w} + \dots + \frac{\partial J_m(w, b)}{\partial w} \right)$$

Gradient Descent - Introduction

From the prior slide, we have:

$$\begin{aligned}\frac{\partial J(w, b)}{\partial w} &= \frac{1}{2m} \left(\frac{\partial J_1(w, b)}{\partial w} + \frac{\partial J_2(w, b)}{\partial w} + \dots + \frac{\partial J_m(w, b)}{\partial w} \right) \\ &= \frac{1}{2m} \left(2(wx^{(1)} + b - y^{(1)})x^{(1)} + 2(wx^{(2)} + b - y^{(2)})x^{(2)} + \dots \right. \\ &\quad \left. + 2(wx^{(m)} + b - y^{(m)})x^{(m)} \right)\end{aligned}$$

Gradient Descent - Introduction

From the prior slide, we have:

$$\begin{aligned}\frac{\partial J(w, b)}{\partial w} &= \frac{1}{2m} \left(\frac{\partial J_1(w, b)}{\partial w} + \frac{\partial J_2(w, b)}{\partial w} + \dots + \frac{\partial J_m(w, b)}{\partial w} \right) \\ &= \frac{1}{2m} \left(2(wx^{(1)} + b - y^{(1)})x^{(1)} + 2(wx^{(2)} + b - y^{(1)})x^{(2)} + \dots \right. \\ &\quad \left. + 2(wx^{(m)} + b - y^{(m)})x^{(m)} \right) \\ &= \frac{2}{2m} \left((wx^{(1)} + b - y^{(1)})x^{(1)} + (wx^{(2)} + b - y^{(1)})x^{(2)} + \dots \right. \\ &\quad \left. + (wx^{(m)} + b - y^{(m)})x^{(m)} \right)\end{aligned}$$

Gradient Descent - Introduction

From the prior slide, we have:

$$\begin{aligned}\frac{\partial J(w, b)}{\partial w} &= \frac{1}{2m} \left(\frac{\partial J_1(w, b)}{\partial w} + \frac{\partial J_2(w, b)}{\partial w} + \dots + \frac{\partial J_m(w, b)}{\partial w} \right) \\ &= \frac{1}{2m} \left(2(wx^{(1)} + b - y^{(1)})x^{(1)} + 2(wx^{(2)} + b - y^{(1)})x^{(2)} + \dots \right. \\ &\quad \left. + 2(wx^{(m)} + b - y^{(m)})x^{(m)} \right) \\ &= \frac{2}{2m} \left((wx^{(1)} + b - y^{(1)})x^{(1)} + (wx^{(2)} + b - y^{(1)})x^{(2)} + \dots \right. \\ &\quad \left. + (wx^{(m)} + b - y^{(m)})x^{(m)} \right) \\ &= \frac{1}{m} \left((wx^{(1)} + b - y^{(1)})x^{(1)} + (wx^{(2)} + b - y^{(1)})x^{(2)} + \dots \right. \\ &\quad \left. + (wx^{(m)} + b - y^{(m)})x^{(m)} \right)\end{aligned}$$

Gradient Descent - Introduction

From the prior slide, we have:

Gradient Descent - Introduction

From the prior slide, we have:

$$\frac{\partial J(w, b)}{\partial w} = \frac{1}{m} \left((wx^{(1)} + b - y^{(1)})x^{(1)} + (wx^{(2)} + b - y^{(1)})x^{(2)} + \dots + (wx^{(m)} + b - y^{(m)})x^{(m)} \right)$$

Gradient Descent - Introduction

From the prior slide, we have:

$$\frac{\partial J(w, b)}{\partial w} = \frac{1}{m} \left((wx^{(1)} + b - y^{(1)})x^{(1)} + (wx^{(2)} + b - y^{(2)})x^{(2)} + \dots + (wx^{(m)} + b - y^{(m)})x^{(m)} \right)$$

$$\therefore \frac{\partial J(w)}{\partial w} = \frac{1}{m} \sum_{i=1}^m (wx^{(i)} + b - y^{(i)})x^{(i)}$$

Gradient Descent - Introduction

Next, we must find $\frac{\partial J(w, b)}{\partial b}$.

Gradient Descent - Introduction

Next, we must find $\frac{\partial J(w, b)}{\partial b}$.

$$J(w, b) = \frac{1}{2m} \sum_{i=1}^m (wx^{(i)} + b - y^{(i)})^2$$

Gradient Descent - Introduction

Next, we must find $\frac{\partial J(w, b)}{\partial b}$.

$$J(w, b) = \frac{1}{2m} \sum_{i=1}^m (wx^{(i)} + b - y^{(i)})^2$$

$$\begin{aligned} J(w, b) = \frac{1}{2m} & \left((wx^{(1)} + b - y^{(1)})^2 + (wx^{(2)} + b - y^{(2)})^2 + \dots \right. \\ & \left. + (wx^{(m)} + b - y^{(m)})^2 \right) \end{aligned}$$

Gradient Descent - Introduction

Next, we must find $\frac{\partial J(w, b)}{\partial b}$.

$$J(w, b) = \frac{1}{2m} \sum_{i=1}^m (wx^{(i)} + b - y^{(i)})^2$$

$$\begin{aligned} J(w, b) = \frac{1}{2m} & \left((wx^{(1)} + b - y^{(1)})^2 + (wx^{(2)} + b - y^{(2)})^2 + \dots \right. \\ & \left. + (wx^{(m)} + b - y^{(m)})^2 \right) \end{aligned}$$

Let's take just one term of this sum:

Gradient Descent - Introduction

Next, we must find $\frac{\partial J(w, b)}{\partial b}$.

$$J(w, b) = \frac{1}{2m} \sum_{i=1}^m (wx^{(i)} + b - y^{(i)})^2$$

$$\begin{aligned} J(w, b) = \frac{1}{2m} & \left((wx^{(1)} + b - y^{(1)})^2 + (wx^{(2)} + b - y^{(2)})^2 + \dots \right. \\ & \left. + (wx^{(m)} + b - y^{(m)})^2 \right) \end{aligned}$$

Let's take just one term of this sum:

Let's define $J_1(w, b) \equiv (wx^{(1)} + b - y^{(1)})^2$

Gradient Descent - Introduction

Next, we must find $\frac{\partial J(w, b)}{\partial b}$.

$$J(w, b) = \frac{1}{2m} \sum_{i=1}^m (wx^{(i)} + b - y^{(i)})^2$$

$$\begin{aligned} J(w, b) = \frac{1}{2m} & \left((wx^{(1)} + b - y^{(1)})^2 + (wx^{(2)} + b - y^{(2)})^2 + \dots \right. \\ & \left. + (wx^{(m)} + b - y^{(m)})^2 \right) \end{aligned}$$

Let's take just one term of this sum:

Let's define $J_1(w, b) \equiv (wx^{(1)} + b - y^{(1)})^2$

Let's find $\frac{\partial J_1(w, b)}{\partial b}$.

Gradient Descent - Introduction

Let $J_1(w, b) \equiv (wx^{(1)} + b - y^{(1)})^2$.

Gradient Descent - Introduction

Let $J_1(w, b) \equiv (wx^{(1)} + b - y^{(1)})^2$. Find $\frac{\partial J_1(w, b)}{\partial b}$:

Use the Chain Rule.

Gradient Descent - Introduction

Let $J_1(w, b) \equiv (wx^{(1)} + b - y^{(1)})^2$. Find $\frac{\partial J_1(w, b)}{\partial b}$:

Use the Chain Rule. Let $u = wx^{(1)} + b - y^{(1)}$. Then, $y = u^2$.

Gradient Descent - Introduction

Let $J_1(w, b) \equiv (wx^{(1)} + b - y^{(1)})^2$. Find $\frac{\partial J_1(w, b)}{\partial b}$:

Use the Chain Rule. Let $u = wx^{(1)} + b - y^{(1)}$. Then, $y = u^2$.

$$\frac{dy}{du} = 2u$$

Gradient Descent - Introduction

Let $J_1(w, b) \equiv (wx^{(1)} + b - y^{(1)})^2$. Find $\frac{\partial J_1(w, b)}{\partial b}$:

Use the Chain Rule. Let $u = wx^{(1)} + b - y^{(1)}$. Then, $y = u^2$.

$$\frac{dy}{du} = 2u$$

$$\frac{du}{db} = 1$$

Gradient Descent - Introduction

Let $J_1(w, b) \equiv (wx^{(1)} + b - y^{(1)})^2$. Find $\frac{\partial J_1(w, b)}{\partial b}$:

Use the Chain Rule. Let $u = wx^{(1)} + b - y^{(1)}$. Then, $y = u^2$.

$$\frac{dy}{du} = 2u$$

$$\frac{du}{db} = 1$$

$$\frac{dy}{db} = \frac{dy}{du} \cdot \frac{du}{db} = 2u \cdot 1 = 2(wx^{(1)} + b - y^{(1)}) \cdot 1$$

Gradient Descent - Introduction

Let $J_1(w, b) \equiv (wx^{(1)} + b - y^{(1)})^2$. Find $\frac{\partial J_1(w, b)}{\partial b}$:

Use the Chain Rule. Let $u = wx^{(1)} + b - y^{(1)}$. Then, $y = u^2$.

$$\frac{dy}{du} = 2u$$

$$\frac{du}{db} = 1$$

$$\frac{dy}{db} = \frac{dy}{du} \cdot \frac{du}{db} = 2u \cdot 1 = 2(wx^{(1)} + b - y^{(1)}) \cdot 1$$

$$\therefore \frac{\partial J_1(w)}{\partial b} = 2(wx^{(1)} + b - y^{(1)}) \cdot 1 = 2(wx^{(1)} + b - y^{(1)})$$

Gradient Descent - Introduction

In general, let's define:

Gradient Descent - Introduction

In general, let's define:

$$J_i(w, b) \equiv (wx^{(i)} + b - y^{(i)})^2 \quad i = 1 \dots m$$

Gradient Descent - Introduction

In general, let's define:

$$J_i(w, b) \equiv (wx^{(i)} + b - y^{(i)})^2 \quad i = 1 \dots m$$

Then, we have:

Gradient Descent - Introduction

In general, let's define:

$$J_i(w, b) \equiv (wx^{(i)} + b - y^{(i)})^2 \quad i = 1 \dots m$$

Then, we have:

$$\frac{\partial J_i(w, b)}{\partial b} = 2(wx^{(i)} + b - y^{(i)}) \quad i = 1 \dots m$$

Gradient Descent - Introduction

Now, we return to finding $\frac{\partial J(w, b)}{\partial w}$:

Gradient Descent - Introduction

Now, we return to finding $\frac{\partial J(w, b)}{\partial w}$:

$$J(w, b) = \frac{1}{2m} \sum_{i=1}^m (f_w(x^{(i)}) - y^{(i)})^2$$

Gradient Descent - Introduction

Now, we return to finding $\frac{\partial J(w, b)}{\partial w}$:

$$\begin{aligned}J(w, b) &= \frac{1}{2m} \sum_{i=1}^m (f_w(x^{(i)}) - y^{(i)})^2 \\&= \frac{1}{2m} \left((wx^{(1)} + b - y^{(1)})^2 + (wx^{(2)} + b - y^{(2)})^2 + \dots \right. \\&\quad \left. + (wx^{(m)} + b - y^{(m)})^2 \right)\end{aligned}$$

Gradient Descent - Introduction

Now, we return to finding $\frac{\partial J(w, b)}{\partial w}$:

$$\begin{aligned} J(w, b) &= \frac{1}{2m} \sum_{i=1}^m (f_w(x^{(i)}) - y^{(i)})^2 \\ &= \frac{1}{2m} \left((wx^{(1)} + b - y^{(1)})^2 + (wx^{(2)} + b - y^{(2)})^2 + \dots \right. \\ &\quad \left. + (wx^{(m)} + b - y^{(m)})^2 \right) \\ &= \frac{1}{2m} \left(J_1(w, b) + J_2(w, b) + \dots + J_m(w, b) \right) \end{aligned}$$

Gradient Descent - Introduction

Now, we return to finding $\frac{\partial J(w, b)}{\partial w}$:

$$\begin{aligned} J(w, b) &= \frac{1}{2m} \sum_{i=1}^m (f_w(x^{(i)}) - y^{(i)})^2 \\ &= \frac{1}{2m} \left((wx^{(1)} + b - y^{(1)})^2 + (wx^{(2)} + b - y^{(2)})^2 + \dots \right. \\ &\quad \left. + (wx^{(m)} + b - y^{(m)})^2 \right) \\ &= \frac{1}{2m} \left(J_1(w, b) + J_2(w, b) + \dots + J_m(w, b) \right) \\ \therefore \frac{\partial J(w, b)}{\partial b} &= \frac{\partial}{\partial w} \left[\frac{1}{2m} \left(J_1(w, b) + J_2(w, b) + \dots + J_m(w, b) \right) \right] \end{aligned}$$

Gradient Descent - Introduction

Now, we return to finding $\frac{\partial J(w, b)}{\partial w}$:

$$\begin{aligned}
 J(w, b) &= \frac{1}{2m} \sum_{i=1}^m (f_w(x^{(i)}) - y^{(i)})^2 \\
 &= \frac{1}{2m} \left((wx^{(1)} + b - y^{(1)})^2 + (wx^{(2)} + b - y^{(2)})^2 + \dots \right. \\
 &\quad \left. + (wx^{(m)} + b - y^{(m)})^2 \right) \\
 &= \frac{1}{2m} \left(J_1(w, b) + J_2(w, b) + \dots + J_m(w, b) \right)
 \end{aligned}$$

$$\begin{aligned}
 \therefore \frac{\partial J(w, b)}{\partial b} &= \frac{\partial}{\partial w} \left[\frac{1}{2m} \left(J_1(w, b) + J_2(w, b) + \dots + J_m(w, b) \right) \right] \\
 &= \frac{1}{2m} \left(\frac{\partial J_1(w, b)}{\partial b} + \frac{\partial J_2(w, b)}{\partial b} + \dots + \frac{\partial J_m(w, b)}{\partial b} \right)
 \end{aligned}$$

Gradient Descent - Introduction

From the prior slide, we have:

Gradient Descent - Introduction

From the prior slide, we have:

$$\frac{\partial J(w, b)}{\partial b} = \frac{1}{2m} \left(\frac{\partial J_1(w, b)}{\partial b} + \frac{\partial J_2(w, b)}{\partial b} + \dots + \frac{\partial J_m(w, b)}{\partial b} \right)$$

Gradient Descent - Introduction

From the prior slide, we have:

$$\begin{aligned}\frac{\partial J(w, b)}{\partial b} &= \frac{1}{2m} \left(\frac{\partial J_1(w, b)}{\partial b} + \frac{\partial J_2(w, b)}{\partial b} + \dots + \frac{\partial J_m(w, b)}{\partial b} \right) \\ &= \frac{1}{2m} \left(2(wx^{(1)} + b - y^{(1)}) + 2(wx^{(2)} + b - y^{(2)}) + \dots \right. \\ &\quad \left. + 2(wx^{(m)} + b - y^{(m)}) \right)\end{aligned}$$

Gradient Descent - Introduction

From the prior slide, we have:

$$\begin{aligned}\frac{\partial J(w, b)}{\partial b} &= \frac{1}{2m} \left(\frac{\partial J_1(w, b)}{\partial b} + \frac{\partial J_2(w, b)}{\partial b} + \dots + \frac{\partial J_m(w, b)}{\partial b} \right) \\ &= \frac{1}{2m} \left(2(wx^{(1)} + b - y^{(1)}) + 2(wx^{(2)} + b - y^{(1)}) + \dots \right. \\ &\quad \left. + 2(wx^{(m)} + b - y^{(m)}) \right) \\ &= \frac{2}{2m} \left((wx^{(1)} + b - y^{(1)}) + (wx^{(2)} + b - y^{(1)}) + \dots \right. \\ &\quad \left. + (wx^{(m)} + b - y^{(m)}) \right)\end{aligned}$$

Gradient Descent - Introduction

From the prior slide, we have:

$$\begin{aligned}\frac{\partial J(w, b)}{\partial b} &= \frac{1}{2m} \left(\frac{\partial J_1(w, b)}{\partial b} + \frac{\partial J_2(w, b)}{\partial b} + \dots + \frac{\partial J_m(w, b)}{\partial b} \right) \\ &= \frac{1}{2m} \left(2(wx^{(1)} + b - y^{(1)}) + 2(wx^{(2)} + b - y^{(1)}) + \dots \right. \\ &\quad \left. + 2(wx^{(m)} + b - y^{(m)}) \right) \\ &= \frac{2}{2m} \left((wx^{(1)} + b - y^{(1)}) + (wx^{(2)} + b - y^{(1)}) + \dots \right. \\ &\quad \left. + (wx^{(m)} + b - y^{(m)}) \right) \\ &= \frac{1}{m} \left((wx^{(1)} + b - y^{(1)}) + (wx^{(2)} + b - y^{(1)}) + \dots \right. \\ &\quad \left. + (wx^{(m)} + b - y^{(m)}) \right)\end{aligned}$$

Gradient Descent - Introduction

From the prior slide, we have:

Gradient Descent - Introduction

From the prior slide, we have:

$$\frac{\partial J(w, b)}{\partial b} = \frac{1}{m} \left((wx^{(1)} + b - y^{(1)}) + (wx^{(2)} + b - y^{(2)}) + \dots + (wx^{(m)} + b - y^{(m)}) \right)$$

Gradient Descent - Introduction

From the prior slide, we have:

$$\frac{\partial J(w, b)}{\partial b} = \frac{1}{m} \left((wx^{(1)} + b - y^{(1)}) + (wx^{(2)} + b - y^{(2)}) + \dots + (wx^{(m)} + b - y^{(m)}) \right)$$

$$\therefore \frac{\partial J(w, b)}{\partial b} = \frac{1}{m} \sum_{i=1}^m (wx^{(i)} + b - y^{(i)})$$

Gradient Descent - Full Model

So, now, we have our Gradient Descent for $f_{w,b}(x)$, substituting our partial derivative expressions for J with respect to w and b :

Gradient Descent - Full Model

So, now, we have our Gradient Descent for $f_{w,b}(x)$, substituting our partial derivative expressions for J with respect to w and b :

- ① Initialize w, b to random values.

Gradient Descent - Full Model

So, now, we have our Gradient Descent for $f_{w,b}(x)$, substituting our partial derivative expressions for J with respect to w and b :

- ➊ Initialize w, b to random values.
- ➋ Repeat until convergence:

Gradient Descent - Full Model

So, now, we have our Gradient Descent for $f_{w,b}(x)$, substituting our partial derivative expressions for J with respect to w and b :

- ➊ Initialize w, b to random values.
- ➋ Repeat until convergence:

$$w := w - \alpha \frac{\partial J(w, b)}{\partial w}$$

$$b := b - \alpha \frac{\partial J(w, b)}{\partial b}$$

Gradient Descent - Full Model

So, now, we have our Gradient Descent for $f_{w,b}(x)$, substituting our partial derivative expressions for J with respect to w and b :

- ➊ Initialize w, b to random values.
- ➋ Repeat until convergence:

$$w := w - \alpha \frac{1}{m} \sum_{i=1}^m (wx^{(i)} + b - y^{(i)})x^{(i)}$$

$$b := b - \alpha \frac{1}{m} \sum_{i=1}^m (wx^{(i)} + b - y^{(i)})$$

Gradient Descent - Full Model

So, now, we have our Gradient Descent for $f_{w,b}(x)$, substituting our partial derivative expressions for J with respect to w and b :

- ➊ Initialize w, b to random values.
- ➋ Repeat until convergence:

$$w := w - \alpha \frac{1}{m} \sum_{i=1}^m (f_{w,b}(x^{(i)}) - y^{(i)}) x^{(i)}$$

$$b := b - \alpha \frac{1}{m} \sum_{i=1}^m (f_{w,b}(x^{(i)}) - y^{(i)})$$

Gradient Descent - Full Model

- ➊ Initialize w, b to random values.
- ➋ Repeat until convergence:

$$w := w - \alpha \frac{1}{m} \sum_{i=1}^m (f_{w,b}(x^{(i)}) - y^{(i)}) x^{(i)}$$
$$b := b - \alpha \frac{1}{m} \sum_{i=1}^m (f_{w,b}(x^{(i)}) - y^{(i)})$$

Gradient Descent - Full Model

- ➊ Initialize w, b to random values.
- ➋ Repeat until convergence:

$$w := w - \alpha \frac{1}{m} \sum_{i=1}^m (f_{w,b}(x^{(i)}) - y^{(i)}) x^{(i)}$$

$$b := b - \alpha \frac{1}{m} \sum_{i=1}^m (f_{w,b}(x^{(i)}) - y^{(i)})$$

Observations:

Gradient Descent - Full Model

- ➊ Initialize w, b to random values.
- ➋ Repeat until convergence:

$$w := w - \alpha \frac{1}{m} \sum_{i=1}^m (f_{w,b}(x^{(i)}) - y^{(i)}) x^{(i)}$$
$$b := b - \alpha \frac{1}{m} \sum_{i=1}^m (f_{w,b}(x^{(i)}) - y^{(i)})$$

Observations:

- Adjustments to parameters are a function of the size of the residuals, $f_{w,b}(x^{(i)}) - y^{(i)}$.

Gradient Descent - Full Model

- ➊ Initialize w, b to random values.
- ➋ Repeat until convergence:

$$w := w - \alpha \frac{1}{m} \sum_{i=1}^m (f_{w,b}(x^{(i)}) - y^{(i)}) x^{(i)}$$

$$b := b - \alpha \frac{1}{m} \sum_{i=1}^m (f_{w,b}(x^{(i)}) - y^{(i)})$$

Observations:

- Adjustments to parameters are a function of the size of the residuals, $f_{w,b}(x^{(i)}) - y^{(i)}$. In general, bigger residuals lead to bigger parameter adjustments,

Gradient Descent - Full Model

- ➊ Initialize w, b to random values.
- ➋ Repeat until convergence:

$$w := w - \alpha \frac{1}{m} \sum_{i=1}^m (f_{w,b}(x^{(i)}) - y^{(i)}) x^{(i)}$$

$$b := b - \alpha \frac{1}{m} \sum_{i=1}^m (f_{w,b}(x^{(i)}) - y^{(i)})$$

Observations:

- Adjustments to parameters are a function of the size of the residuals, $f_{w,b}(x^{(i)}) - y^{(i)}$. In general, bigger residuals lead to bigger parameter adjustments, and vice versa.

Gradient Descent - Full Model

- ➊ Initialize w, b to random values.
- ➋ Repeat until convergence:

$$w := w - \alpha \frac{1}{m} \sum_{i=1}^m (f_{w,b}(x^{(i)}) - y^{(i)}) x^{(i)}$$
$$b := b - \alpha \frac{1}{m} \sum_{i=1}^m (f_{w,b}(x^{(i)}) - y^{(i)})$$

Observations:

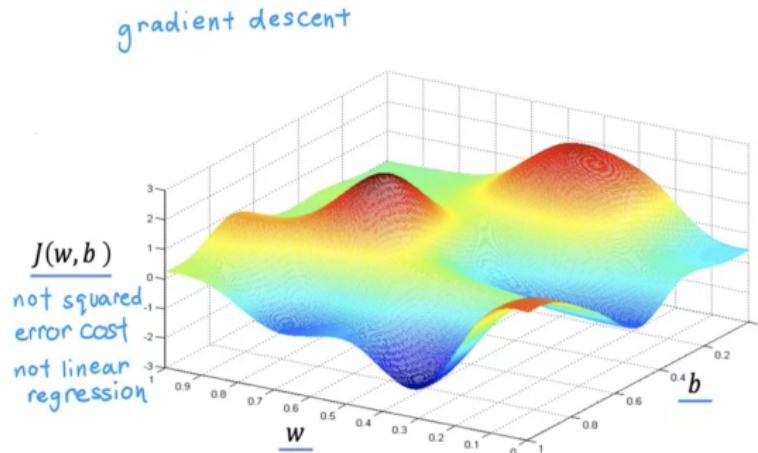
- Adjustments to parameters are a function of the size of the residuals, $f_{w,b}(x^{(i)}) - y^{(i)}$. In general, bigger residuals lead to bigger parameter adjustments, and vice versa.
- For each step, updates to w and b should be made at the same time.

Gradient Descent - Convex Cost Function

Cost Function Surface (not squared error function):

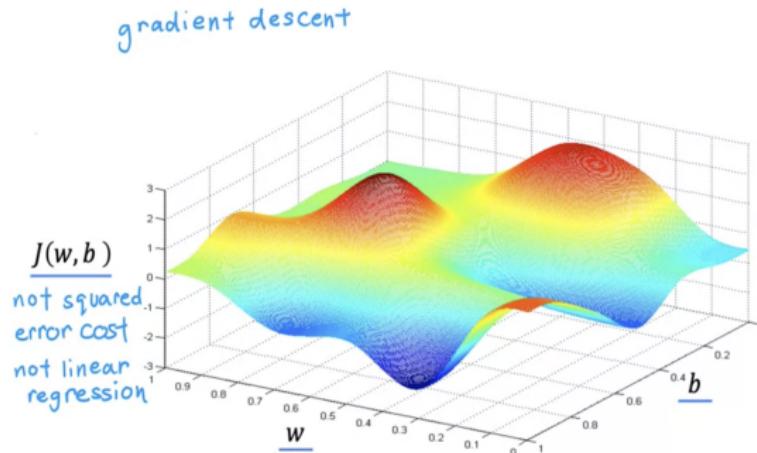
Gradient Descent - Convex Cost Function

Cost Function Surface (not squared error function):



Gradient Descent - Convex Cost Function

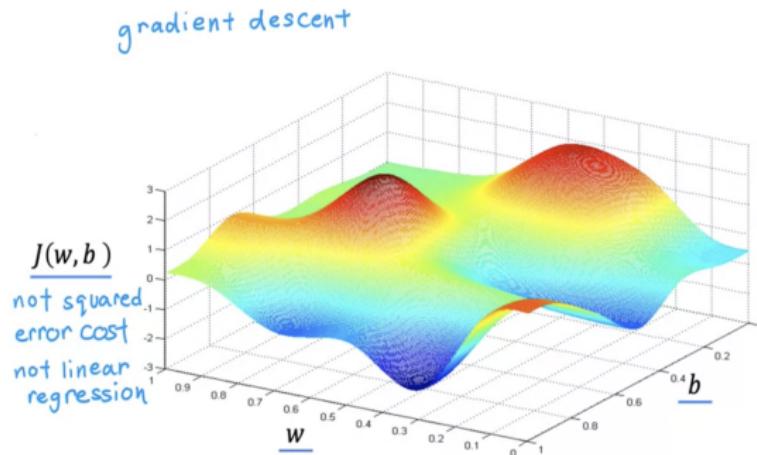
Cost Function Surface (not squared error function):



Depending on where we start on this surface (random placement),

Gradient Descent - Convex Cost Function

Cost Function Surface (not squared error function):

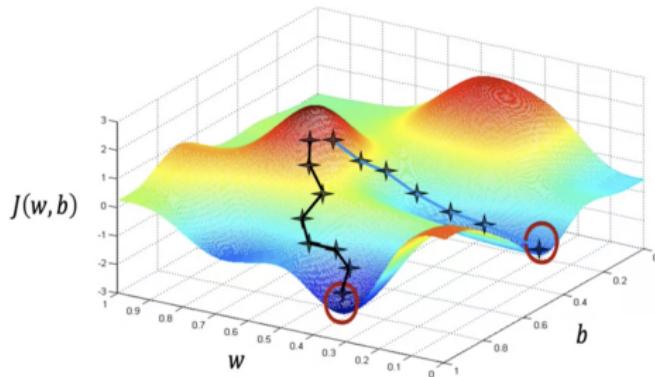


Depending on where we start on this surface (random placement), we may end up at a local min (not global).

Gradient Descent - Convex Cost Function

Cost Function Surface Search for Min:

More than one local minimum



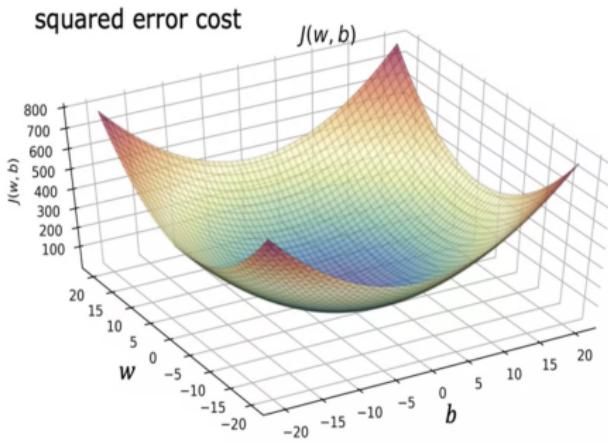
Depending on where we start on this surface (random placement), we may end up at a local min (not global).

Gradient Descent - Convex Cost Function

Cost Function Surface (squared error function):

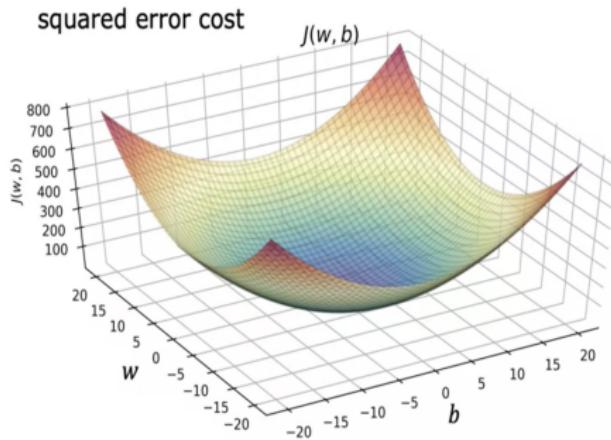
Gradient Descent - Convex Cost Function

Cost Function Surface (squared error function):



Gradient Descent - Convex Cost Function

Cost Function Surface (squared error function):



A squared error cost function insures a convex surface for which we'll converge to the global minimum.

Table of Contents

1 Overview of Machine Learning

- Definition of ML
- Types of ML
- ML Architecture

2 Example - House Price Prediction

- House Price Prediction - Training Set and Notation

3 Model Development - Part 1 - Model, Cost Function Selection

- Methodology
- Model Selection
- Cost Function Selection

4 Model Development - Part 2 - Model Training

- Review of The Chain Rule
- Gradient Descent - Simplified Model
- Gradient Descent - Full Model
- Training the House Price Prediction Model

Gradient Descent - House Prices Example

Now, we begin the third step in our machine learning approach:

Gradient Descent - House Prices Example

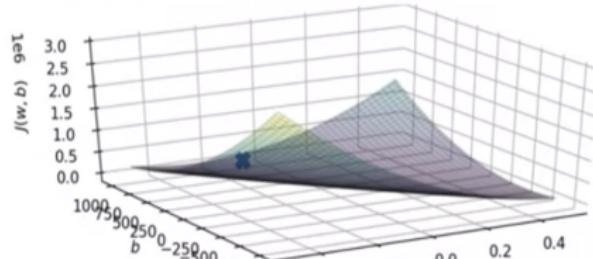
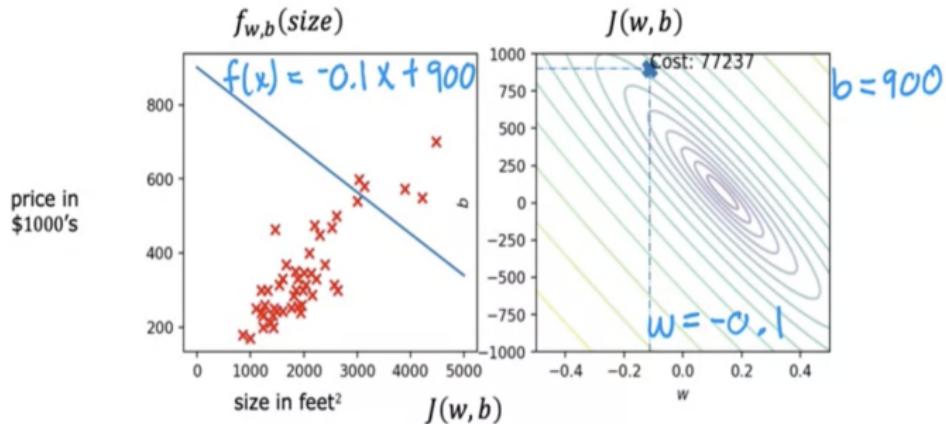
Now, we begin the third step in our machine learning approach:

- ③ Execute Gradient Descent:

Gradient Descent - House Prices Example

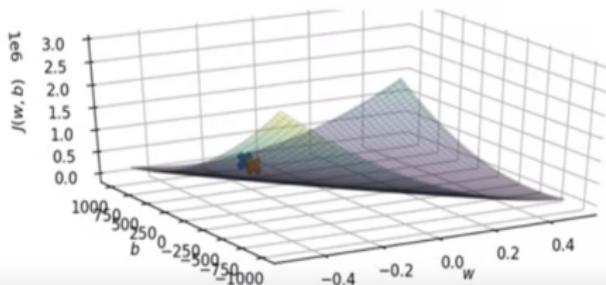
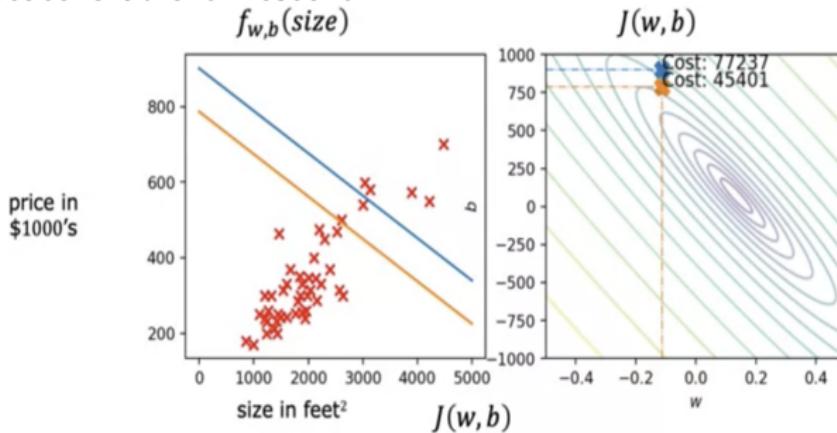
Now, we begin the third step in our machine learning approach:

- Execute Gradient Descent:



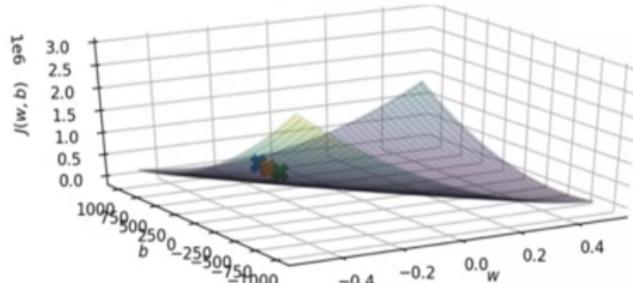
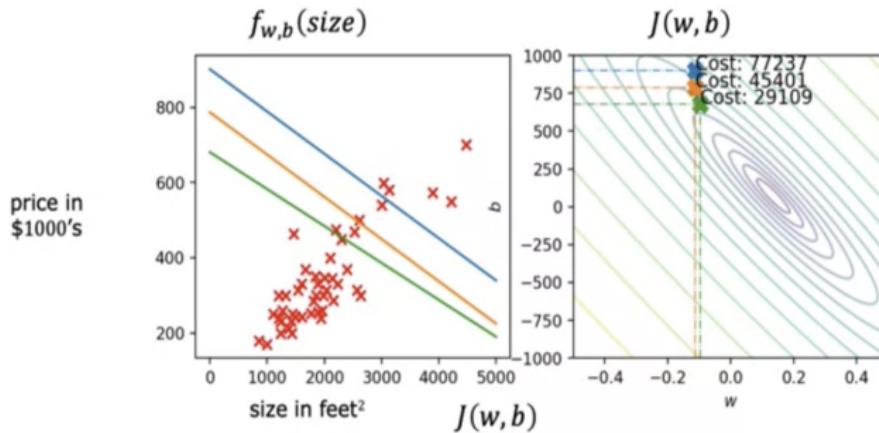
Gradient Descent - House Prices Example

③ Execute Gradient Descent:



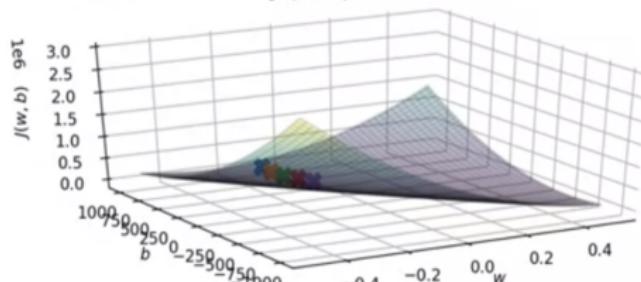
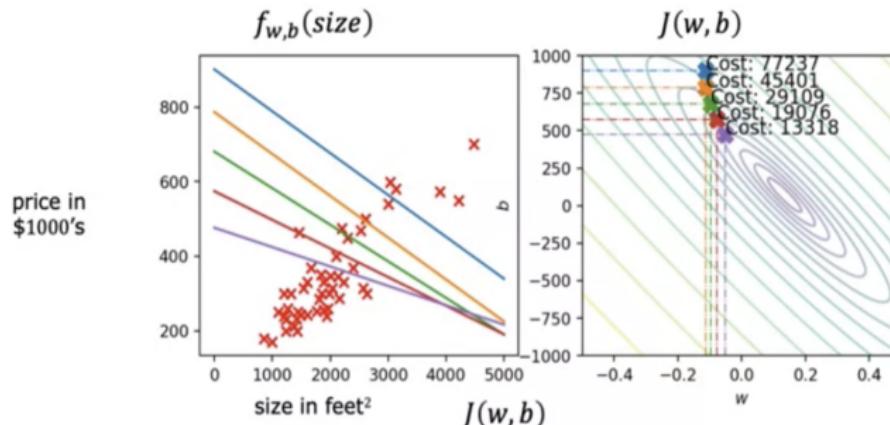
Gradient Descent - House Prices Example

③ Execute Gradient Descent:



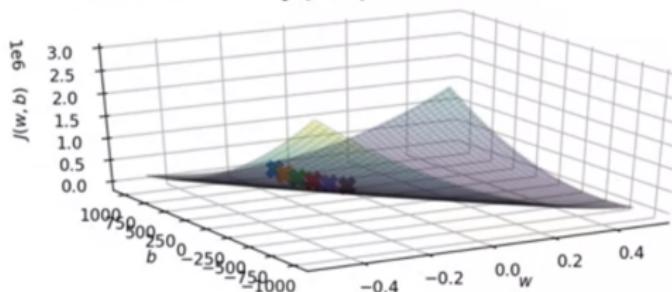
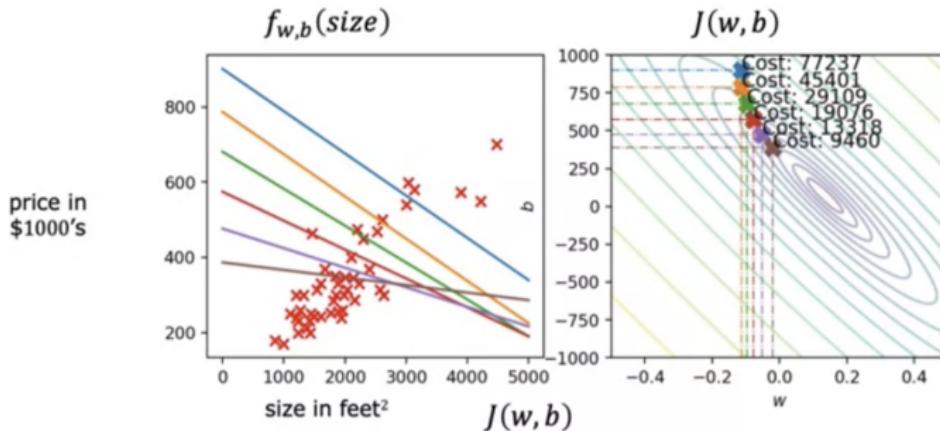
Gradient Descent - House Prices Example

③ Execute Gradient Descent:



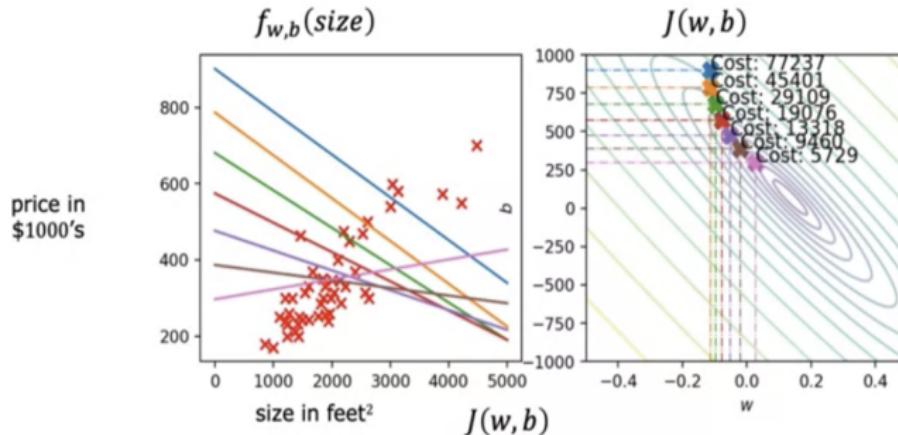
Gradient Descent - House Prices Example

③ Execute Gradient Descent:



Gradient Descent - House Prices Example

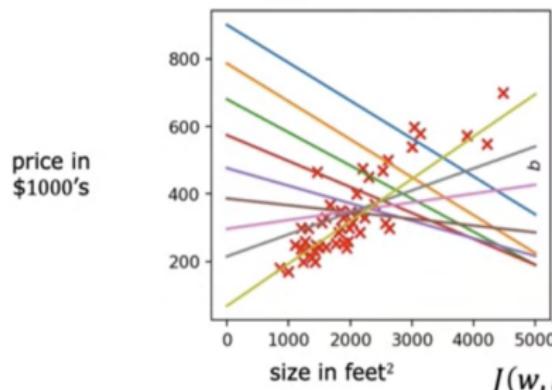
③ Execute Gradient Descent:



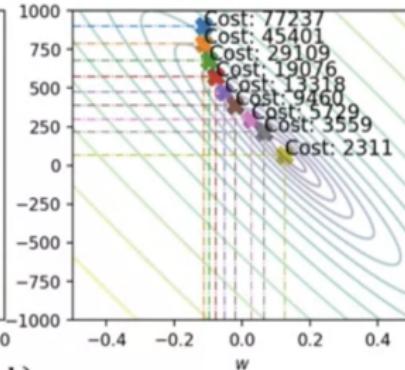
Gradient Descent - House Prices Example

③ Execute Gradient Descent:

$$f_{w,b}(\text{size})$$



$$J(w, b)$$



$$J(w, b)$$

