Outline

Introduction

ML Motivation

Maximum Likelihood Regression

Regression Examples

Bayesian Perspective

Bayesian Regression
Outline

Introduction

ML Motivation

Maximum Likelihood Regression

Regression Examples

Bayesian Perspective

Bayesian Regression
What is Machine Learning?
Equipping Computers with Human Like Capabilities.

- Endow computers with the ability to “learn” from “data”.
- Present data from sensors, the internet, experiments.
- Expect computer to make “sensible” decisions.
- Traditionally categorized as:
  - **Supervised learning**: classification, regression.
  - **Unsupervised learning**: dimensionality reduction, clustering.
  - **Reinforcement learning**: learning from delayed feedback.
    - Planning. Difficult stuff!
What is Machine Learning?
Equipping Computers with Human Like Capabilities.

- Endow computers with the ability to “learn” from “data”.
- Present data from sensors, the internet, experiments.
- Expect computer to make “sensible” decisions.
- Traditionally categorized as:
  - **Supervised learning**: classification, regression.
  - **Unsupervised learning**: dimensionality reduction, clustering.
  - **Reinforcement learning**: learning from delayed feedback.
    - Planning. Difficult stuff!
What is Machine Learning?
Equipping Computers with Human Like Capabilities.

- Endow computers with the ability to “learn” from “data”.
- Present data from sensors, the internet, experiments.
- Expect computer to make “sensible” decisions.
- Traditionally categorized as:
  - Supervised learning: classification, regression.
  - Unsupervised learning: dimensionality reduction, clustering.
  - Reinforcement learning: learning from delayed feedback.
  - Planning. Difficult stuff!
What is Machine Learning?
Equipping Computers with Human Like Capabilities.

- Endow computers with the ability to “learn” from “data”.
- Present data from sensors, the internet, experiments.
- Expect computer to make “sensible” decisions.
- Traditionally categorized as:
  - **Supervised learning**: classification, regression.
  - **Unsupervised learning**: dimensionality reduction, clustering.
  - **Reinforcement learning**: learning from delayed feedback.
  - Planning. Difficult stuff!
What is Machine Learning?
Equipping Computers with Human Like Capabilities.

- Endow computers with the ability to “learn” from “data”.
- Present data from sensors, the internet, experiments.
- Expect computer to make “sensible” decisions.
- Traditionally categorized as:
  - Supervised learning: classification, regression.
  - Unsupervised learning: dimensionality reduction, clustering.
  - Reinforcement learning: learning from delayed feedback.
  - Planning. Difficult stuff!
What is Machine Learning?
Equipping Computers with Human Like Capabilities.

▶ Endow computers with the ability to “learn” from “data”.
▶ Present data from sensors, the internet, experiments.
▶ Expect computer to make “sensible” decisions.
▶ Traditionally categorized as:
  
  * **Supervised learning:** classification, regression.
  * **Unsupervised learning:** dimensionality reduction, clustering.
  * **Reinforcement learning:** learning from delayed feedback.
  
  Planning. Difficult stuff!
Applications of Machine Learning


**Collaborative Filtering** : Prediction of user preferences for items given purchase history. For example the Netflix Prize [http://www.netflixprize.com/](http://www.netflixprize.com/).


History of Machine Learning (personal)
Rosenblatt to Vapnik

- Arises from the Connectionist movement in AI.
- Early Connectionist research focused on models of the brain.
Arises from the Connectionist movement in AI. 
http://en.wikipedia.org/wiki/Connectionism

Early Connectionist research focused on models of the brain.
Rosenblatt’s perceptron (Rosenblatt, 1962) based on simple model of a neuron (McCulloch and Pitts, 1943) and a learning algorithm.

Figure: Frank Rosenblatt in 1950 (source: Cornell University Library)
Later machine learning research focused on theoretical foundations of such models and their capacity to learn (Vapnik, 1998).

Machine learning benefited greatly by incorporating ideas from psychology, but not being afraid to incorporate rigorous theory.
Early machine learning viewed with scepticism by statisticians.

Modern machine learning and statistics interact to both communities benefits.

*Personal view:* statistics and machine learning are fundamentally different. Statistics aims to provide a human with the tools to analyze data. Machine learning wants to replace the human in the processing of data.
Early machine learning viewed with scepticism by statisticians.

Modern machine learning and statistics interact to both communities benefits.

*Personal view:* statistics and machine learning are fundamentally different. Statistics aims to provide a human with the tools to analyze data. Machine learning wants to replace the human in the processing of data.
Early machine learning viewed with scepticism by statisticians.

Modern machine learning and statistics interact to both communities benefits.

*Personal view:* statistics and machine learning are fundamentally different. Statistics aims to provide a human with the tools to analyze data. Machine learning wants to replace the human in the processing of data.
For the moment the two overlap strongly. But they are not the same field!

Machine learning also has overlap with Cognitive Science.

Mathematical formalisms of a problem are helpful, but they can hide facts: i.e. the fallacy that “aerodynamically a bumble bee can’t fly”. Clearly a limitation of the model rather than fact.

Mathematical foundations are still very important though: they help us understand the capabilities of our algorithms.

But we mustn’t restrict our ambitions to the limitations of current mathematical formalisms. That is where humans give inspiration.
For the moment the two overlap strongly. But they are not the same field!

Machine learning also has overlap with Cognitive Science.

Mathematical formalisms of a problem are helpful, but they can hide facts: i.e. the fallacy that “aerodynamically a bumble bee can’t fly”. Clearly a limitation of the model rather than fact.

Mathematical foundations are still very important though: they help us understand the capabilities of our algorithms.

But we mustn’t restrict our ambitions to the limitations of current mathematical formalisms. That is where humans give inspiration.
For the moment the two overlap strongly. But they are not the same field!

Machine learning also has overlap with Cognitive Science.

Mathematical formalisms of a problem are helpful, but they can hide facts: i.e. the fallacy that “aerodynamically a bumble bee can’t fly”. Clearly a limitation of the model rather than fact.

Mathematical foundations are still very important though: they help us understand the capabilities of our algorithms.

But we mustn’t restrict our ambitions to the limitations of current mathematical formalisms. That is where humans give inspiration.
For the moment the two overlap strongly. But they are not the same field!

Machine learning also has overlap with Cognitive Science.

Mathematical formalisms of a problem are helpful, but they can hide facts: i.e. the fallacy that “aerodynamically a bumble bee can’t fly”. Clearly a limitation of the model rather than fact.

Mathematical foundations are still very important though: they help us understand the capabilities of our algorithms.

But we mustn’t restrict our ambitions to the limitations of current mathematical formalisms. That is where humans give inspiration.
For the moment the two overlap strongly. But they are not the same field!

Machine learning also has overlap with Cognitive Science.

Mathematical formalisms of a problem are helpful, but they can hide facts: i.e. the fallacy that “aerodynamically a bumble bee can’t fly”. Clearly a limitation of the model rather than fact.

Mathematical foundations are still very important though: they help us understand the capabilities of our algorithms.

But we mustn’t restrict our ambitions to the limitations of current mathematical formalisms. That is where humans give inspiration.
Early statistics had great success with the idea of statistical proof.

**Question:** I computed the mean of these two tables of numbers (a statistic). They are different. Does this “prove” anything?

**Answer:** it depends on how the numbers are generated, how many there are and how big the difference. Randomization is important.

- Hypothesis testing: questions you can ask about your data are quite limiting.
- This can have the affect of limiting science too.
- Many successes: crop fertilization, clinical trials, brewing, polling.
- Many open questions: e.g. causality.
Early statistics had great success with the idea of statistical proof.

**Question:** I computed the mean of these two tables of numbers (a statistic). They are different. Does this “prove” anything?

**Answer:** it depends on how the numbers are generated, how many there are and how big the difference. Randomization is important.

Hypothesis testing: questions you can ask about your data are quite limiting.

This can have the affect of limiting science too.

Many successes: crop fertilization, clinical trials, brewing, polling.

Many open questions: e.g. causality.
Early statistics had great success with the idea of statistical proof.

**Question:** I computed the mean of these two tables of numbers (a statistic). They are different. Does this “prove” anything?

**Answer:** It depends on how the numbers are generated, how many there are and how big the difference. Randomization is important.

Hypothesis testing: questions you can ask about your data are quite limiting.

This can have the affect of limiting science too.

Many successes: crop fertilization, clinical trials, brewing, polling.

Many open questions: e.g. causality.
Early statistics had great success with the idea of statistical proof.

**Question:** I computed the mean of these two tables of numbers (a statistic). They are different. Does this “prove” anything?

**Answer:** it depends on how the numbers are generated, how many there are and how big the difference. Randomization is important.

Hypothesis testing: questions you can ask about your data are quite limiting.

This can have the affect of limiting science too.

Many successes: crop fertilization, clinical trials, brewing, polling.

Many open questions: e.g. causality.
Early statistics had great success with the idea of statistical proof.

**Question:** I computed the mean of these two tables of numbers (a statistic). They are different. Does this “prove” anything?

**Answer:** it depends on how the numbers are generated, how many there are and how big the difference. Randomization is important.

Hypothesis testing: questions you can ask about your data are quite limiting.

This can have the affect of limiting science too.

Many successes: crop fertilization, clinical trials, brewing, polling.

Many open questions: e.g. causality.
Early statistics had great success with the idea of statistical proof.

**Question:** I computed the mean of these two tables of numbers (a statistic). They are different. Does this “prove” anything?

**Answer:** it depends on how the numbers are generated, how many there are and how big the difference. Randomization is important.

- Hypothesis testing: questions you can ask about your data are quite limiting.
- This can have the affect of limiting science too.
- Many successes: crop fertilization, clinical trials, brewing, polling.
- Many open questions: e.g. causality.
Many statisticians were Edwardian English gentleman.

Figure: William Sealy Gosset in 1908
Cricket and Baseball are two games with a lot of “statistics”.

The study of the meaning behind these numbers is “mathematical statistics” often abbreviated to “statistics”.
▶ Cricket and Baseball are two games with a lot of “statistics”.
▶ The study of the meaning behind these numbers is “mathematical statistics” often abbreviated to “statistics”.
The world is an *uncertain* place.

Epistemic uncertainty: uncertainty arising through lack of knowledge. (What colour socks is that person wearing?)

Aleatoric uncertainty: uncertainty arising through an underlying stochastic system. (Where will a sheet of paper fall if I drop it?)
The world is an *uncertain* place.

**Epistemic uncertainty**: uncertainty arising through lack of knowledge. *(What colour socks is that person wearing?)*

**Aleatoric uncertainty**: uncertainty arising through an underlying stochastic system. *(Where will a sheet of paper fall if I drop it?)*
The world is an *uncertain* place.

**Epistemic uncertainty:** uncertainty arising through lack of knowledge. (What colour socks is that person wearing?)

**Aleatoric uncertainty:** uncertainty arising through an underlying stochastic system. (Where will a sheet of paper fall if I drop it?)
We need a framework to characterise the uncertainty.

In this course we make use of probability theory to characterise uncertainty.
We need a framework to characterise the uncertainty.

In this course we make use of probability theory to characterise uncertainty.
Richard Price

- Welsh philosopher and essay writer.
- Edited Thomas Bayes’s essay which contained foundations of Bayesian philosophy.

Figure: Richard Price, 1723–1791. (source Wikipedia)
Laplace

- French Mathematician and Astronomer.

Figure: Pierre-Simon Laplace, 1749–1827. (source Wikipedia)
Supervised Learning
We are given data set containing “inputs”, \( X \), and “targets”, \( y \).

Each data point consists of an input vector \( x_i \) and a class label, \( y_i \).

For binary classification assume \( y_i \) should be either 1 (yes) or \(-1\) (no).

Input vector can be thought of as features.
Classification Examples

- Classifying handwritten digits from binary images (automatic zip code reading).
- Detecting faces in images (e.g. digital cameras).
- Who a detected face belongs to (e.g. Picasa).
- Classifying type of cancer given gene expression data.
- Categorization of document types (different types of news article on the internet).
Developed in 1957 by Rosenblatt.

Take a data point at, $x_i$.

Predict it belongs to a class, $y_i = 1$ if $\sum_j w_j x_{i,j} + b > 0$ i.e. $\mathbf{w}^\top \mathbf{x}_i + b > 0$. Otherwise assume $y_i = -1$. 
Perceptron-like Algorithm

1. Select a random data point \( i \).
2. Ensure \( i \) is correctly classified by setting \( \mathbf{w} = y_i \mathbf{x}_i \).
   ▶ i.e. \( \text{sign} (\mathbf{w}^\top \mathbf{x}_i) = \text{sign} (y_i \mathbf{x}_i^\top \mathbf{x}_i) = \text{sign} (y_i) = y_i \)
3. Iterate: increment \( k \) and select a misclassified point, \( i \).
4. Set \( \mathbf{w} \leftarrow \mathbf{w} + \eta y_i \mathbf{x}_i \).
   ▶ If \( \eta \) is large enough this will guarantee this point becomes correctly classified.
5. Repeat until there are no misclassified points.
Perceptron Algorithm

- Iteration 1 data no 29

Simple Dataset
Perceptron Algorithm

- Iteration 1 data no 29
- \( w_1 = 0, w_2 = 0 \)
- Iteration 1 data no 29
- \( w_1 = 0, \ w_2 = 0 \)
- First Iteration

**Simple Dataset**

![Graph showing simple dataset with data points in 2D space.](image-url)
Perceptron Algorithm

- Iteration 1 data no 29
- \( w_1 = 0, \ w_2 = 0 \)
- First Iteration
- Set weight vector to data point.
Perceptron Algorithm

- Iteration 1 data no 29
- $w_1 = 0, w_2 = 0$
- First Iteration
- Set weight vector to data point.
- $\mathbf{w} = y_{29} \mathbf{x}_{29}$
Perceptron Algorithm

- Iteration 1 data no 29
- $w_1 = 0, w_2 = 0$
- First Iteration
- Set weight vector to data point.
- $\mathbf{w} = y_{29} \mathbf{x}_{29}$
- Select new incorrectly classified data point.
Perceptron Algorithm

- **Iteration 2 data no 16**

Simple Dataset

```
\[
\begin{align*}
\mathbf{w}_1 &= 0.3519, \\
\mathbf{w}_2 &= -0.6787.
\end{align*}
\]

Incorrect classification

Adjust weight vector with new data point.

\[
\mathbf{w} \leftarrow \mathbf{w} + \eta y_{16} \mathbf{x}_{16},
\]

Select new incorrectly classified data point.

![Simple Dataset Diagram](image-url)
Perceptron Algorithm

- Iteration 2 data no 16
- $w_1 = 0.3519$, $w_2 = -0.6787$
Perceptron Algorithm

- Iteration 2 data no 16
- $w_1 = 0.3519$, $w_2 = -0.6787$
- Incorrect classification
Perceptron Algorithm

- Iteration 2 data no 16
- \( w_1 = 0.3519, \)
  \( w_2 = -0.6787 \)
- Incorrect classification
- Adjust weight vector with new data point.
Perceptron Algorithm

- Iteration 2 data no 16
- $w_1 = 0.3519, \quad w_2 = -0.6787$
- Incorrect classification
- Adjust weight vector with new data point.
- $\mathbf{w} \leftarrow \mathbf{w} + \eta y_{16} \mathbf{x}_{16}$,

![Simple Dataset](image)
Perceptron Algorithm

- Iteration 2 data no 16
- $w_1 = 0.3519,$ $w_2 = -0.6787$
- Incorrect classification
- Adjust weight vector with new data point.
- $\mathbf{w} \leftarrow \mathbf{w} + \eta y_{16} \mathbf{x}_{16}$
- Select new incorrectly classified data point.
- Perceptron Algorithm

- Iteration 3 data no 58

- $w_1 = -1.2143$, $w_2 = -1.0217$

- Incorrect classification

- Adjust weight vector with new data point.

- $w \leftarrow w + \eta y_{58} x_{58}$

- All data correctly classified.

- Simple Dataset

- Graph showing data points and decision boundary.
Perceptron Algorithm

- Iteration 3 data no 58
- \( w_1 = -1.2143, \quad w_2 = -1.0217 \)
Perceptron Algorithm

- Iteration 3 data no 58
- \( w_1 = -1.2143, \)
- \( w_2 = -1.0217 \)
- Incorrect classification
Perceptron Algorithm

- Iteration 3 data no 58
- \( w_1 = -1.2143, \)
  \( w_2 = -1.0217 \)
- Incorrect classification
- Adjust weight vector with new data point.
Perceptron Algorithm

▶ Iteration 3 data no 58
▶ \( w_1 = -1.2143, \)
  \( w_2 = -1.0217 \)
▶ Incorrect classification
▶ Adjust weight vector with new data point.
▶ \( \mathbf{w} \leftarrow \mathbf{w} + \eta y_{58} x_{58}, \)
Perceptron Algorithm

- Iteration 3 data no 58
- \( w_1 = -1.2143, \)
  \( w_2 = -1.0217 \)
- Incorrect classification
- Adjust weight vector with new data point.
- \( \mathbf{w} \leftarrow \mathbf{w} + \eta y_{58} \mathbf{x}_{58}, \)
- All data correctly classified.
Regression Examples

- Predict a real value, \( y_i \) given some inputs \( x_i \).
- Predict quality of meat given spectral measurements (Tecator data).
- Radiocarbon dating, the C14 calibration curve: predict age given quantity of C14 isotope.
- Predict quality of different Go or Backgammon moves given expert rated training data.
Linear Regression

Is there an equivalent learning rule for regression?

- Predict a real value \( y \) given \( x \).
- We can also construct a learning rule for regression.
  - Define our prediction
    \[
    f(x) = mx + c.
    \]
  - Define an error
    \[
    \Delta y_i = y_i - f(x_i).
    \]
Updating Bias/Intercept

- $c$ represents bias. Add portion of error to bias.

\[ c \rightarrow c + \eta \Delta y_i. \]

\[ \Delta y_i = y_i - mx_i - c. \]

1. For +ve error, $c$ and therefore $f(x_i)$ become larger and error magnitude becomes smaller.
2. For -ve error, $c$ and therefore $f(x_i)$ become smaller and error magnitude becomes smaller.
Updating Slope

- \( m \) represents Slope. Add portion of error \( \times \) input to slope.

\[
m \rightarrow m + \eta \Delta y_i x_i.
\]

\[
\Delta y_i = y_i - mx_i - c.
\]

1. For +ve error and +ve input, \( m \) becomes larger and \( f(x_i) \) becomes larger: error magnitude becomes smaller.
2. For +ve error and -ve input, \( m \) becomes smaller and \( f(x_i) \) becomes larger: error magnitude becomes smaller.
3. For -ve error and -ve slope, \( m \) becomes larger and \( f(x_i) \) becomes smaller: error magnitude becomes smaller.
4. For -ve error and +ve input, \( m \) becomes smaller and \( f(x_i) \) becomes smaller: error magnitude becomes smaller.
Iteration 1  \( \hat{m} = -0.3 \)
\( \hat{c} = 1 \)
Linear Regression Example

- Iteration 1 \( \hat{m} = -0.3 \)
  \( \hat{c} = 1 \)
- Present data point 4

\[ \Delta y_4 = (y_4 - \hat{m}x_4 - \hat{c}) \]

\[ \hat{m} \leftarrow \hat{m} + \eta x_4 \Delta y_4 \]
\[ \hat{c} \leftarrow \hat{c} + \eta \Delta y_4 \]

Updated values \( \hat{m} = -0.2593 \)
\( \hat{c} = 1.0175 \)
Iteration 1  \( \hat{m} = -0.3 \)
\( \hat{c} = 1 \)

- Present data point 4
- \( \Delta y_4 = (y_4 - \hat{m}x_4 - \hat{c}) \)
**Linear Regression Example**

- **Iteration 1**  
  \( \hat{m} = -0.3 \)  
  \( \hat{c} = 1 \)

- Present data point 4
- \( \Delta y_4 = (y_4 - \hat{m}x_4 - \hat{c}) \)
- Adjust \( \hat{m} \) and \( \hat{c} \)
  \( \hat{m} \leftarrow \hat{m} + \eta x_4 \Delta y_4 \)  
  \( \hat{c} \leftarrow \hat{c} + \eta \Delta y_4 \)
Iteration 1 \( \hat{m} = -0.3 \)
\( \hat{c} = 1 \)

- Present data point 4
- \( \Delta y_4 = (y_4 - \hat{m}x_4 - \hat{c}) \)
- Adjust \( \hat{m} \) and \( \hat{c} \)
  \( \hat{m} \leftarrow \hat{m} + \eta x_4 \Delta y_4 \)
  \( \hat{c} \leftarrow \hat{c} + \eta \Delta y_4 \)

Updated values
\( \hat{m} = -0.25593 \)
\( \hat{c} = 1.0175 \)
Iteration 2

\[ \hat{m} = -0.25593 \]
\[ \hat{c} = 1.0175 \]
Iteration 2
\[ \hat{m} = -0.25593 \]
\[ \hat{c} = 1.0175 \]

Present data point 7

\[ \Delta y_7 = (y_7 - \hat{m}x_7 - \hat{c}) \]

Adjust \( \hat{m} \) and \( \hat{c} \)

\[ \hat{m} \leftarrow \hat{m} + \eta x_7 \Delta y_7 \]
\[ \hat{c} \leftarrow \hat{c} + \eta \Delta y_7 \]

Updated values

\[ \hat{m} = -0.20693 \]
\[ \hat{c} = 1.0358 \]
Iteration 2
\[ \hat{m} = -0.25593 \]
\[ \hat{c} = 1.0175 \]

- Present data point 7
- \( \Delta y_7 = (y_7 - \hat{m}x_7 - \hat{c}) \)
**Linear Regression Example**

- **Iteration 2**
  - $\hat{m} = -0.25593$
  - $\hat{c} = 1.0175$
- Present data point 7
- $\Delta y_7 = (y_7 - \hat{m}x_7 - \hat{c})$
- Adjust $\hat{m}$ and $\hat{c}$
  - $\hat{m} \leftarrow \hat{m} + \eta x_7 \Delta y_7$
  - $\hat{c} \leftarrow \hat{c} + \eta \Delta y_7$
Iteration 2
\[ \hat{m} = -0.25593 \]
\[ \hat{c} = 1.0175 \]

- Present data point 7
- \[ \Delta y_7 = (y_7 - \hat{m}x_7 - \hat{c}) \]
- Adjust \( \hat{m} \) and \( \hat{c} \)
  \[ \hat{m} \leftarrow \hat{m} + \eta x_7 \Delta y_7 \]
  \[ \hat{c} \leftarrow \hat{c} + \eta \Delta y_7 \]

Updated values
\[ \hat{m} = -0.20693 \]
\[ \hat{c} = 1.0358 \]
Iteration 3
\[ \hat{m} = -0.20693 \]
\[ \hat{c} = 1.0358 \]
Iteration 3
\( \hat{m} = -0.20693 \)
\( \hat{c} = 1.0358 \)

Present data point 10
Iteration 3
\[ \hat{m} = -0.20693 \]
\[ \hat{c} = 1.0358 \]
- Present data point 10
- \[ \Delta y_{10} = (y_{10} - \hat{m}x_{10} - \hat{c}) \]
Iteration 3
\( \hat{m} = -0.20693 \)
\( \hat{c} = 1.0358 \)

- Present data point 10
- \( \Delta y_{10} = (y_{10} - \hat{m}x_{10} - \hat{c}) \)
- Adjust \( \hat{m} \) and \( \hat{c} \)
  \( \hat{m} \leftarrow \hat{m} + \eta x_{10} \Delta y_{10} \)
  \( \hat{c} \leftarrow \hat{c} + \eta \Delta y_{10} \)
Linear Regression Example

- **Iteration 3**
  \[ \hat{m} = -0.20693 \]
  \[ \hat{c} = 1.0358 \]
  - Present data point 10
    - \[ \Delta y_{10} = (y_{10} - \hat{m}x_{10} - \hat{c}) \]
  - Adjust \( \hat{m} \) and \( \hat{c} \)
    \[ \hat{m} \leftarrow \hat{m} + \eta x_{10} \Delta y_{10} \]
    \[ \hat{c} \leftarrow \hat{c} + \eta \Delta y_{10} \]
- **Updated values**
  \[ \hat{m} = -0.085591 \]
  \[ \hat{c} = 1.0617 \]
Iteration 4
\[ \hat{m} = -0.085591 \]
\[ \hat{c} = 1.0617 \]
Iteration 4
\[ \hat{m} = -0.085591 \]
\[ \hat{c} = 1.0617 \]
- Present data point 7

\[ \Delta y_7 = (y_7 - \hat{m}x_7 - \hat{c}) \]

\[ \hat{m} \leftarrow \hat{m} + \eta x_7 \Delta y_7 \]
\[ \hat{c} \leftarrow \hat{c} + \eta \Delta y_7 \]

Updated values
\[ \hat{m} = -0.050355 \]
\[ \hat{c} = 1.0749 \]
Linear Regression Example

- **Iteration 4**
  - $\hat{m} = -0.085591$
  - $\hat{c} = 1.0617$
  
  - Present data point 7
  - $\Delta y_7 = (y_7 - \hat{m}x_7 - \hat{c})$
**Linear Regression Example**

- **Iteration 4**
  \[ \hat{m} = -0.085591 \]
  \[ \hat{c} = 1.0617 \]
  - Present data point 7
  - \[ \Delta y_7 = (y_7 - \hat{m}x_7 - \hat{c}) \]
  - Adjust \( \hat{m} \) and \( \hat{c} \)
  \[ \hat{m} \leftarrow \hat{m} + \eta x_7 \Delta y_7 \]
  \[ \hat{c} \leftarrow \hat{c} + \eta \Delta y_7 \]
Linear Regression Example

- Iteration 4
  \[ \hat{m} = -0.085591 \]
  \[ \hat{c} = 1.0617 \]
  - Present data point 7
  - \[ \Delta y_7 = (y_7 - \hat{m}x_7 - \hat{c}) \]
  - Adjust \( \hat{m} \) and \( \hat{c} \)
    - \( \hat{m} \leftarrow \hat{m} + \eta x_7 \Delta y_7 \)
    - \( \hat{c} \leftarrow \hat{c} + \eta \Delta y_7 \)

- Updated values
  \[ \hat{m} = -0.050355 \]
  \[ \hat{c} = 1.0749 \]
Iteration 5
\[ \hat{m} = -0.050355 \]
\[ \hat{c} = 1.0749 \]
Linear Regression Example

- Iteration 5
  \( \hat{m} = -0.050355 \)
  \( \hat{c} = 1.0749 \)
- Present data point 4
Iteration 5
\[
\hat{m} = -0.050355 \\
\hat{c} = 1.0749
\]

- Present data point 4
- \( \Delta y_4 = (y_4 - \hat{m}x_4 - \hat{c}) \)
Iteration 5
\[ \hat{m} = -0.050355 \]
\[ \hat{c} = 1.0749 \]

- Present data point 4
- \[ \Delta y_4 = (y_4 - \hat{m}x_4 - \hat{c}) \]
- Adjust \( \hat{m} \) and \( \hat{c} \)
  \[ \hat{m} \leftarrow \hat{m} + \eta x_4 \Delta y_4 \]
  \[ \hat{c} \leftarrow \hat{c} + \eta \Delta y_4 \]
Iteration 5

\( \hat{m} = -0.050355 \)

\( \hat{c} = 1.0749 \)

- Present data point 4
- \( \Delta y_4 = (y_4 - \hat{m}x_4 - \hat{c}) \)
- Adjust \( \hat{m} \) and \( \hat{c} \)
  \( \hat{m} \leftarrow \hat{m} + \eta x_4 \Delta y_4 \)
  \( \hat{c} \leftarrow \hat{c} + \eta \Delta y_4 \)

Updated values

\( \hat{m} = -0.024925 \)

\( \hat{c} = 1.0849 \)
Iteration 6
\[ \hat{m} = -0.024925 \]
\[ \hat{c} = 1.0849 \]
Linear Regression Example

- Iteration 6
  - $\hat{m} = -0.024925$
  - $\hat{c} = 1.0849$
    - Present data point 5

![Graph showing linear regression](image-url)
Iteration 6
\[ \hat{m} = -0.024925 \]
\[ \hat{c} = 1.0849 \]

- Present data point 5
- \[ \Delta y_5 = (y_5 - \hat{m}x_5 - \hat{c}) \]
Linear Regression Example

- **Iteration 6**
  \[ \hat{m} = -0.024925 \]
  \[ \hat{c} = 1.0849 \]

  - Present data point 5
  - \[ \Delta y_5 = (y_5 - \hat{m}x_5 - \hat{c}) \]
  - Adjust \( \hat{m} \) and \( \hat{c} \)
    \[ \hat{m} \leftarrow \hat{m} + \eta x_5 \Delta y_5 \]
    \[ \hat{c} \leftarrow \hat{c} + \eta \Delta y_5 \]
Linear Regression Example

- **Iteration 6**
  - \( \hat{m} = -0.024925 \)
  - \( \hat{c} = 1.0849 \)
  - Present data point 5
  - \( \Delta y_5 = (y_5 - \hat{m}x_5 - \hat{c}) \)
  - Adjust \( \hat{m} \) and \( \hat{c} \)
    - \( \hat{m} \leftarrow \hat{m} + \eta x_5 \Delta y_5 \)
    - \( \hat{c} \leftarrow \hat{c} + \eta \Delta y_5 \)
- **Updated values**
  - \( \hat{m} = 0.00098511 \) \( \hat{c} = 1.0949 \)
Iteration 7

\[ \hat{m} = 0.00098511 \]

\[ \hat{c} = 1.0949 \]
Iteration 7
\[ \hat{m} = 0.00098511 \]
\[ \hat{c} = 1.0949 \]

- Present data point 10

- Updated values
  \[ \hat{m} = 0.072529 \]
  \[ \hat{c} = 1.1101 \]
Iteration 7
\[ \hat{m} = 0.00098511 \]
\[ \hat{c} = 1.0949 \]
- Present data point 10
- \[ \Delta y_{10} = (y_{10} - \hat{m}x_{10} - \hat{c}) \]
- **Iteration 7**
  \[
  \hat{m} = 0.00098511 \\
  \hat{c} = 1.0949
  \]
- Present data point 10
- \[
  \Delta y_{10} = (y_{10} - \hat{m} x_{10} - \hat{c})
  \]
- Adjust \( \hat{m} \) and \( \hat{c} \)
  \[
  \hat{m} \leftarrow \hat{m} + \eta x_{10} \Delta y_{10} \\
  \hat{c} \leftarrow \hat{c} + \eta \Delta y_{10}
  \]
Linear Regression Example

- Iteration 7
  - $\hat{m} = 0.00098511$
  - $\hat{c} = 1.0949$
  - Present data point 10
  - $\Delta y_{10} = (y_{10} - \hat{m}x_{10} - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    - $\hat{m} \leftarrow \hat{m} + \eta x_{10} \Delta y_{10}$
    - $\hat{c} \leftarrow \hat{c} + \eta \Delta y_{10}$

- Updated values
  - $\hat{m} = 0.072529$  $\hat{c} = 1.1101$
Linear Regression Example

Iteration 8  \( \hat{m} = 0.072529 \)
\( \hat{c} = 1.1101 \)
Iteration 8  \( \hat{m} = 0.072529 \)
\( \hat{c} = 1.1101 \)

Present data point 10

\[
\Delta y_{10} = (y_{10} - \hat{m}x_{10} - \hat{c})
\]

\[
\hat{m} \leftarrow \hat{m} + \eta x_{10} \Delta y_{10}
\]
\[
\hat{c} \leftarrow \hat{c} + \eta \Delta y_{10}
\]

Updated values
\( \hat{m} = 0.1282 \)
\( \hat{c} = 1.122 \)
Iteration 8  \( \hat{m} = 0.072529 \)
\( \hat{c} = 1.1101 \)

- Present data point 10
- \( \Delta y_{10} = (y_{10} - \hat{m}x_{10} - \hat{c}) \)
Iteration 8  \( \hat{m} = 0.072529 \)
\( \hat{c} = 1.1101 \)

- Present data point 10
- \( \Delta y_{10} = (y_{10} - \hat{m}x_{10} - \hat{c}) \)
- Adjust \( \hat{m} \) and \( \hat{c} \)
  \( \hat{m} \leftarrow \hat{m} + \eta x_{10} \Delta y_{10} \)
  \( \hat{c} \leftarrow \hat{c} + \eta \Delta y_{10} \)
Linear Regression Example

- **Iteration 8** \( \hat{m} = 0.072529 \)  
  \( \hat{c} = 1.1101 \)
  - Present data point 10
  - \( \Delta y_{10} = (y_{10} - \hat{m}x_{10} - \hat{c}) \)
  - Adjust \( \hat{m} \) and \( \hat{c} \)
    \( \hat{m} \leftarrow \hat{m} + \eta x_{10} \Delta y_{10} \)
    \( \hat{c} \leftarrow \hat{c} + \eta \Delta y_{10} \)

- **Updated values**  
  \( \hat{m} = 0.1282 \)  \( \hat{c} = 1.122 \)
Linear Regression Example

Iteration 9  \( \hat{m} = 0.1282 \)
\( \hat{c} = 1.122 \)
Iteration 9  $\hat{m} = 0.1282$
$\hat{c} = 1.122$

Present data point 7

$\Delta y_7 = (y_7 - \hat{m}x_7 - \hat{c})$

Adjust $\hat{m}$ and $\hat{c}$:

$\hat{m} \leftarrow \hat{m} + \eta x_7 \Delta y_7$
$\hat{c} \leftarrow \hat{c} + \eta \Delta y_7$

Updated values:

$\hat{m} = 0.14634$
$\hat{c} = 1.1288$
Iteration 9 \(\hat{m} = 0.1282\)  
\(\hat{c} = 1.122\)

- Present data point 7
- \(\Delta y_7 = (y_7 - \hat{m}x_7 - \hat{c})\)
Iteration 9  \( \hat{m} = 0.1282 \)
\( \hat{c} = 1.122 \)

- Present data point 7
- \( \Delta y_7 = (y_7 - \hat{m}x_7 - \hat{c}) \)
- Adjust \( \hat{m} \) and \( \hat{c} \)
  \( \hat{m} \leftarrow \hat{m} + \eta x_7 \Delta y_7 \)
  \( \hat{c} \leftarrow \hat{c} + \eta \Delta y_7 \)
Linear Regression Example

- **Iteration 9**  \( \hat{m} = 0.1282 \)
  \( \hat{c} = 1.122 \)
  - Present data point 7
  - \( \Delta y_7 = (y_7 - \hat{m}x_7 - \hat{c}) \)
  - Adjust \( \hat{m} \) and \( \hat{c} \)
    \( \hat{m} \leftarrow \hat{m} + \eta x_7 \Delta y_7 \)
    \( \hat{c} \leftarrow \hat{c} + \eta \Delta y_7 \)
  
- **Updated values**
  \( \hat{m} = 0.14634 \)  \( \hat{c} = 1.1288 \)
Iteration 10  \( \hat{m} = 0.14634 \)
\( \hat{c} = 1.1288 \)

- Present data point 10
- \( \Delta y_{10} = (y_{10} - \hat{m}x_{10} - \hat{c}) \)
- Adjust \( \hat{m} \) and \( \hat{c} \)
  \( \hat{m} \leftarrow \hat{m} + \eta x_{10} \Delta y_{10} \)
  \( \hat{c} \leftarrow \hat{c} + \eta \Delta y_{10} \)
Iteration 10  \( \hat{m} = 0.14634 \)  
\( \hat{c} = 1.1288 \)

- Present data point 10
- \( \Delta y_{10} = (y_{10} - \hat{m}x_{10} - \hat{c}) \)
- Adjust \( \hat{m} \) and \( \hat{c} \)
  \[ \hat{m} \leftarrow \hat{m} + \eta x_{10} \Delta y_{10} \]
  \[ \hat{c} \leftarrow \hat{c} + \eta \Delta y_{10} \]

Updated values
\( \hat{m} = 0.18547 \)  \( \hat{c} = 1.1372 \)
Iteration 20  \( \hat{m} = 0.27764 \)
\( \hat{c} = 1.1621 \)

- Present data point 6
- \( \Delta y_6 = (y_6 - \hat{m}x_6 - \hat{c}) \)
- Adjust \( \hat{m} \) and \( \hat{c} \)
  \( \hat{m} \leftarrow \hat{m} + \eta x_6 \Delta y_6 \)
  \( \hat{c} \leftarrow \hat{c} + \eta \Delta y_6 \)
Iteration 20  \( \hat{m} = 0.27764 \)
\( \hat{c} = 1.1621 \)

- Present data point 6
- \( \Delta y_6 = (y_6 - \hat{m}x_6 - \hat{c}) \)
- Adjust \( \hat{m} \) and \( \hat{c} \)
  \( \hat{m} \leftarrow \hat{m} + \eta x_6 \Delta y_6 \)
  \( \hat{c} \leftarrow \hat{c} + \eta \Delta y_6 \)

Updated values
\( \hat{m} = 0.28135 \)  \( \hat{c} = 1.1635 \)
Iteration 30  \( \hat{m} = 0.30249 \)
\( \hat{c} = 1.1673 \)

- Present data point 9
- \( \Delta y_9 = (y_9 - \hat{m}x_9 - \hat{c}) \)
- Adjust \( \hat{m} \) and \( \hat{c} \)
  \( \hat{m} \leftarrow \hat{m} + \eta x_9 \Delta y_9 \)
  \( \hat{c} \leftarrow \hat{c} + \eta \Delta y_9 \)
Iteration 30  \( \hat{m} = 0.30249 \)
\( \hat{c} = 1.1673 \)

- Present data point 9
- \( \Delta y_9 = (y_9 - \hat{m}x_9 - \hat{c}) \)
- Adjust \( \hat{m} \) and \( \hat{c} \):
  \( \hat{m} \leftarrow \hat{m} + \eta x_9 \Delta y_9 \)
  \( \hat{c} \leftarrow \hat{c} + \eta \Delta y_9 \)

Updated values
\( \hat{m} = 0.31119 \)
\( \hat{c} = 1.1693 \)
Linear Regression Example

- Iteration 40  \( \hat{m} = 0.33551 \)
  \( \hat{c} = 1.1754 \)
  - Present data point 10
  - \( \Delta y_{10} = (y_{10} - \hat{m}x_{10} - \hat{c}) \)
  - Adjust \( \hat{m} \) and \( \hat{c} \)
    \( \hat{m} \leftarrow \hat{m} + \eta x_{10} \Delta y_{10} \)
    \( \hat{c} \leftarrow \hat{c} + \eta \Delta y_{10} \)
Iteration 40  \( \hat{m} = 0.33551 \)
\( \hat{c} = 1.1754 \)

- Present data point 10
- \( \Delta y_{10} = (y_{10} - \hat{m}x_{10} - \hat{c}) \)
- Adjust \( \hat{m} \) and \( \hat{c} \)
  \( \hat{m} \leftarrow \hat{m} + \eta x_{10} \Delta y_{10} \)
  \( \hat{c} \leftarrow \hat{c} + \eta \Delta y_{10} \)

- Updated values
  \( \hat{m} = 0.33503 \)  \( \hat{c} = 1.1753 \)
Linear Regression Example

- **Iteration 50**  \( \hat{m} = 0.34126 \)
  \( \hat{c} = 1.1763 \)

  - Present data point 8
  - \( \Delta y_8 = (y_8 - \hat{m}x_8 - \hat{c}) \)
  - Adjust \( \hat{m} \) and \( \hat{c} \)
    \( \hat{m} \leftarrow \hat{m} + \eta x_8 \Delta y_8 \)
    \( \hat{c} \leftarrow \hat{c} + \eta \Delta y_8 \)
Linear Regression Example

- **Iteration 50**  \( \hat{m} = 0.34126 \)
  \( \hat{c} = 1.1763 \)

- Present data point 8
- \( \Delta y_8 = (y_8 - \hat{m}x_8 - \hat{c}) \)
- Adjust \( \hat{m} \) and \( \hat{c} \)
  \( \hat{m} \leftarrow \hat{m} + \eta x_8 \Delta y_8 \)
  \( \hat{c} \leftarrow \hat{c} + \eta \Delta y_8 \)

- **Updated values**
  \( \hat{m} = 0.3439 \)  \( \hat{c} = 1.177 \)
Linear Regression Example

- Iteration 60  \( \hat{m} = 0.34877 \)  
  \( \hat{c} = 1.1775 \)
  
  - Present data point 2
  - \( \Delta y_2 = (y_2 - \hat{m}x_2 - \hat{c}) \)
  - Adjust \( \hat{m} \) and \( \hat{c} \)
    \( \hat{m} \leftarrow \hat{m} + \eta x_2 \Delta y_2 \)
    \( \hat{c} \leftarrow \hat{c} + \eta \Delta y_2 \)
Linear Regression Example

Iteration 60  \( \hat{m} = 0.34877 \)
\( \hat{c} = 1.1775 \)

- Present data point 2
- \( \Delta y_2 = (y_2 - \hat{m}x_2 - \hat{c}) \)
- Adjust \( \hat{m} \) and \( \hat{c} \)
  \( \hat{m} \leftarrow \hat{m} + \eta x_2 \Delta y_2 \)
  \( \hat{c} \leftarrow \hat{c} + \eta \Delta y_2 \)

- Updated values
  \( \hat{m} = 0.34621 \)  \( \hat{c} = 1.1757 \)
Linear Regression Example

- **Iteration 70** \( \hat{m} = 0.34207 \)
  \( \hat{c} = 1.1734 \)
- Present data point 10
- \( \Delta y_{10} = (y_{10} - \hat{m}x_{10} - \hat{c}) \)
- Adjust \( \hat{m} \) and \( \hat{c} \)
  \( \hat{m} \leftarrow \hat{m} + \eta x_{10} \Delta y_{10} \)
  \( \hat{c} \leftarrow \hat{c} + \eta \Delta y_{10} \)
Linear Regression Example

- Iteration 70  \( \hat{m} = 0.34207 \)
  \( \hat{c} = 1.1734 \)
- Present data point 10
- \( \Delta y_{10} = (y_{10} - \hat{m}x_{10} - \hat{c}) \)
- Adjust \( \hat{m} \) and \( \hat{c} \)
  \( \hat{m} \leftarrow \hat{m} + \eta x_{10} \Delta y_{10} \)
  \( \hat{c} \leftarrow \hat{c} + \eta \Delta y_{10} \)
- Updated values
  \( \hat{m} = 0.34088 \)  \( \hat{c} = 1.1732 \)
Problem with Linear Regression—\( x \) may not be linearly related to \( y \).

Potential solution: create a feature space: define \( \phi(x) \) where \( \phi(\cdot) \) is a nonlinear function of \( x \).

Model for target is a linear combination of these nonlinear functions

\[
f(x) = \sum_{j=1}^{K} w_j \phi_j(x) \tag{1}
\]
Basis functions can be global. E.g. quadratic basis:

\[ [1, x, x^2] \]

Figure: A quadratic basis.
Basis functions can be global. E.g. quadratic basis:

\[ [1, x, x^2] \]

**Figure:** A quadratic basis.
Basis functions can be global. E.g. quadratic basis:

\[ [1, x, x^2] \]
Functions Derived from Quadratic Basis

\[ f(x) = w_1 + w_2 x + w_3 x^2 \]

Figure: Function from quadratic basis with weights \( w_1 = 0.87466 \), \( w_2 = -0.38835 \), \( w_3 = -2.0058 \).
Functions Derived from Quadratic Basis

\[ f(x) = w_1 + w_2 x + w_3 x^2 \]

**Figure:** Function from quadratic basis with weights \( w_1 = -0.35908 \), \( w_2 = 1.2274 \), \( w_3 = -0.32825 \).
Functions Derived from Quadratic Basis

\[ f(x) = w_1 + w_2x + w_3x^2 \]

Figure: Function from quadratic basis with weights $w_1 = -1.5638$, $w_2 = -0.73577$, $w_3 = 1.6861$. 
Radial Basis Functions

- Or they can be local. E.g. radial (or Gaussian) basis

\[ \phi_j(x) = \exp \left( -\frac{(x-\mu_j)^2}{\ell^2} \right) \]

**Figure:** Radial basis functions.
Radial Basis Functions

- Or they can be local. E.g. radial (or Gaussian) basis

\[ \phi_j(x) = \exp \left( - \frac{(x - \mu_j)^2}{\ell^2} \right) \]

\[ \phi_1(x) = e^{-2(x+1)^2} \]

\[ \phi_2(x) = e^{-2x^2} \]

**Figure:** Radial basis functions.
Radial Basis Functions

- Or they can be local. E.g. radial (or Gaussian) basis

\[ \phi_j(x) = \exp \left( -\frac{(x-\mu_j)^2}{\ell^2} \right) \]

**Figure:** Radial basis functions.
Functions Derived from Radial Basis

$$f(x) = w_1 e^{-2(x+1)^2} + w_2 e^{-2x^2} + w_3 e^{-2(x-1)^2}$$

Figure: Function from radial basis with weights $w_1 = -0.47518$, $w_2 = -0.18924$, $w_3 = -1.8183$. 
Functions Derived from Radial Basis

\[ f(x) = w_1 e^{-2(x+1)^2} + w_2 e^{-2x^2} + w_3 e^{-2(x-1)^2} \]

Figure: Function from radial basis with weights \( w_1 = 0.50596, \ w_2 = -0.046315, \ w_3 = 0.26813 \).
Functions Derived from Radial Basis

\[ f(x) = w_1 e^{-2(x+1)^2} + w_2 e^{-2x^2} + w_3 e^{-2(x-1)^2} \]

\[ \begin{align*}
  w_1 &= 0.07179, \\
  w_2 &= 1.3591, \\
  w_3 &= 0.50604.
\end{align*} \]

Figure: Function from radial basis with weights $w_1 = 0.07179$, $w_2 = 1.3591$, $w_3 = 0.50604$. 
Nonlinear Regression Example

Iteration 1
- $w_1 = 0.13018,$
- $w_2 = -0.11355,$
- $w_3 = 0.15448$
- Present data point 4
Nonlinear Regression Example

- **Iteration 1**
  - $w_1 = 0.13018,$
  - $w_2 = -0.11355,$
  - $w_3 = 0.15448$
- Present data point 4
- $\Delta y_4 = y_4 - \phi_4^T w$
Nonlinear Regression Example

- Iteration 1
  - \( w_1 = 0.13018 \),
  - \( w_2 = -0.11355 \),
  - \( w_3 = 0.15448 \)
- Present data point 4
- \( \Delta y_4 = y_4 - \phi_4^T w \)
- Adjust \( \hat{w} \)
Nonlinear Regression Example

▶ Iteration 1
  ▶ $w_1 = 0.13018$, $w_2 = -0.11355$, $w_3 = 0.15448$
  ▶ Present data point 4
  ▶ $\Delta y_4 = y_4 - \phi_4^\top w$
  ▶ Adjust $\hat{w}$

▶ Updated values
$\hat{w} \leftarrow \hat{w} + \eta \phi_4 \Delta y_4$
Nonlinear Regression Example

- **Iteration 2**
  - $w_1 = 0.33696$, $w_2 = 0.11481$, $w_3 = 0.1591$
  - Present data point 7
Nonlinear Regression Example

- **Iteration 2**
  - \( w_1 = 0.33696, \)
  - \( w_2 = 0.11481, \)
  - \( w_3 = 0.1591 \)
  - Present data point 7
  - \( \Delta y_7 = y_7 - \phi_7^\top w \)
Nonlinear Regression Example

- **Iteration 2**
  - $w_1 = 0.33696, \ w_2 = 0.11481, \ w_3 = 0.1591$
  - Present data point 7
  - $\Delta y_7 = y_7 - \phi_7^\top w$
  - Adjust $\hat{w}$

![Graph showing data points and a nonlinear regression curve]

The graph illustrates the data points plotted along with the nonlinear regression curve. The equation $\Delta y_7 = y_7 - \phi_7^\top w$ represents the adjustment of the estimated weights $\hat{w}$ for the present data point 7, where $\phi_7$ is the matrix of basis functions evaluated at the data point $x_7$, and $\Delta y_7$ represents the correction made to the predicted value $\hat{y}_7$.
Nonlinear Regression Example

- Iteration 2
  - $w_1 = 0.33696,$
  - $w_2 = 0.11481,$
  - $w_3 = 0.1591$
  - Present data point 7
  - $\Delta y_7 = y_7 - \phi_7^T w$
  - Adjust $\hat{w}$

- Updated values
  $\hat{w} \leftarrow \hat{w} + \eta \phi_7 \Delta y_7$
Nonlinear Regression Example

- **Iteration 3**
  - $w_1 = 0.18076,$
  - $w_2 = -0.4266,$
  - $w_3 = 0.12473$
  - Present data point 10
Nonlinear Regression Example

- **Iteration 3**
  - $w_1 = 0.18076,$
  - $w_2 = -0.4266,$
  - $w_3 = 0.12473$
  - Present data point 10
  - $\Delta y_{10} = y_{10} - \phi_{10}^\top w$

![Graph showing nonlinear regression with data points and curve]
Iteration 3

- \( w_1 = 0.18076, \)
- \( w_2 = -0.4266, \)
- \( w_3 = 0.12473 \)

- Present data point 10
- \( \Delta y_{10} = y_{10} - \phi_{10}^\top w \)
- Adjust \( \hat{w} \)
Nonlinear Regression Example

- **Iteration 3**
  - \( w_1 = 0.18076, \)
  - \( w_2 = -0.4266, \)
  - \( w_3 = 0.12473 \)
  - Present data point 10
  - \( \Delta y_{10} = y_{10} - \phi_{10}^\top w \)
  - Adjust \( \hat{w} \)

- **Updated values**
  \[ \hat{w} \leftarrow \hat{w} + \eta \phi_{10} \Delta y_{10} \]
Nonlinear Regression Example

- Iteration 4
  - $w_1 = 0.18076,$
  - $w_2 = -0.42893,$
  - $w_3 = -0.14306$
  - Present data point 7
Nonlinear Regression Example

- **Iteration 4**
  - $w_1 = 0.18076,$
  - $w_2 = -0.42893,$
  - $w_3 = -0.14306$
  - **Present data point 7**
  - $\Delta y_7 = y_7 - \phi_7^\top \mathbf{w}$
Nonlinear Regression Example

- Iteration 4
  - $w_1 = 0.18076,$
  - $w_2 = -0.42893,$
  - $w_3 = -0.14306$
  - Present data point 7
  - $\Delta y_7 = y_7 - \phi_7^\top w$
  - Adjust $\hat{w}$
Nonlinear Regression Example

- **Iteration 4**
  - $w_1 = 0.18076$, $w_2 = -0.42893$, $w_3 = -0.14306$
  - Present data point 7
  - $\Delta y_7 = y_7 - \phi_7^T w$
  - Adjust $\hat{w}$

- **Updated values**
  - $\hat{w} \leftarrow \hat{w} + \eta \phi_7 \Delta y_7$
Nonlinear Regression Example

- Iteration 5
  - $w_1 = 0.17372,$
  - $w_2 = -0.45335,$
  - $w_3 = -0.14461$
  - Present data point 4

![Graph showing nonlinear regression example with iteration 5 parameters and data points.](attachment:graph.png)
Nonlinear Regression Example

- **Iteration 5**
  - $w_1 = 0.17372,$
  - $w_2 = -0.45335,$
  - $w_3 = -0.14461$
  - Present data point 4
  - $\Delta y_4 = y_4 - \phi_4^\top w$

![Graph showing nonlinear regression example]
Nonlinear Regression Example

- **Iteration 5**
  - $w_1 = 0.17372,$
  - $w_2 = -0.45335,$
  - $w_3 = -0.14461$
  - Present data point 4
  - $\Delta y_4 = y_4 - \phi_4^\top w$
  - Adjust $\hat{w}$
Iteration 5

- $w_1 = 0.17372$
- $w_2 = -0.45335$
- $w_3 = -0.14461$

Present data point 4

- $\Delta y_4 = y_4 - \phi_4 \hat{w}$

Adjust $\hat{w}$

Updated values

$\hat{w} \leftarrow \hat{w} + \eta \phi_4 \Delta y_4$
Nonlinear Regression Example

- Iteration 6
  - $w_1 = 0.47971,$
  - $w_2 = -0.11541,$
  - $w_3 = -0.13778$
- Present data point 5
- $\Delta y_5 = y_5 - \phi_5^T w$
- Adjust $\hat{w}$
- Updated values
  - $\hat{w} \leftarrow \hat{w} + \eta \phi_5 \Delta y_5$
Nonlinear Regression Example

- Iteration 6
  - \( w_1 = 0.47971, \)
  - \( w_2 = -0.11541, \)
  - \( w_3 = -0.13778 \)
- Present data point 5
- \( \Delta y_5 = y_5 - \phi_5^T w \)
- Adjust \( \hat{w} \)
- Updated values
  - \( \hat{w} \leftarrow \hat{w} + \eta \phi_5 \Delta y_5 \)
Nonlinear Regression Example

- **Iteration 7**
  - \( w_1 = 0.46599, \)
  - \( w_2 = -0.13952, \)
  - \( w_3 = -0.13855 \)
  - Present data point 10
  - \( \Delta y_{10} = y_{10} - \phi_{10}^\top w \)
  - Adjust \( \hat{w} \)

- Updated values
  - \( \hat{w} \leftarrow \hat{w} + \eta \phi_{10} \Delta y_{10} \)
Nonlinear Regression Example

- Iteration 7
  - \( w_1 = 0.46599, \)
  - \( w_2 = -0.13952, \)
  - \( w_3 = -0.13855 \)
- Present data point 10
  - \( \Delta y_{10} = y_{10} - \phi_{10}^\top w \)
- Adjust \( \hat{w} \)
- Updated values
  - \( \hat{w} \leftarrow \hat{w} + \eta \phi_{10} \Delta y_{10} \)
Nonlinear Regression Example

- **Iteration 8**
  - $w_1 = 0.46599$
  - $w_2 = -0.14144$
  - $w_3 = -0.35924$
  - Present data point 10
  - $\Delta y_{10} = y_{10} - \phi_{10}^\top w$
  - Adjust $\hat{w}$

- Updated values
  - $\hat{w} \leftarrow \hat{w} + \eta \phi_{10} \Delta y_{10}$
Nonlinear Regression Example

- **Iteration 8**
  - $w_1 = 0.46599$,
  - $w_2 = -0.14144$,
  - $w_3 = -0.35924$

- Present data point 10

- $\Delta y_{10} = y_{10} - \phi_{10}^\top \hat{w}$

- Adjust $\hat{w}$

- Updated values

  $\hat{w} \leftarrow \hat{w} + \eta \phi_{10} \Delta y_{10}$
Nonlinear Regression Example

- **Iteration 9**
  - $w_1 = 0.46599,$
  - $w_2 = -0.14307,$
  - $w_3 = -0.54679$
  - Present data point 7
  - $\Delta y_7 = y_7 - \phi_7^T w$
  - Adjust $\hat{w}$

- **Updated values**
  - $\hat{w} \leftarrow \hat{w} + \eta \phi_7 \Delta y_7$
Nonlinear Regression Example

- **Iteration 9**
  - $w_1 = 0.46599$
  - $w_2 = -0.14307$
  - $w_3 = -0.54679$
  - Present data point 7
    - $\Delta y_7 = y_7 - \phi_7^\top w$
  - Adjust $\hat{w}$

- **Updated values**
  - $\hat{w} \leftarrow \hat{w} + \eta \phi_7 \Delta y_7$
Nonlinear Regression Example

- **Iteration 10**
  - $w_1 = 0.38071$
  - $w_2 = -0.43867$
  - $w_3 = -0.56556$
  - **Present data point 10**
  - $\Delta y_{10} = y_{10} - \phi_{10}^\top w$
  - **Adjust $\hat{w}$**

- **Updated values**
  - $\hat{w} \leftarrow \hat{w} + \eta \phi_{10} \Delta y_{10}$
Nonlinear Regression Example

- **Iteration 10**
  - $w_1 = 0.38071,$
  - $w_2 = -0.43867,$
  - $w_3 = -0.56556$
  - Present data point 10
  - $\Delta y_{10} = y_{10} - \phi_{10}^\top w$
  - Adjust $\hat{w}$
- **Updated values**
  - $\hat{w} \leftarrow \hat{w} + \eta \phi_{10} \Delta y_{10}$
Nonlinear Regression Example

- **Iteration 11**
  - $w_1 = 0.38071,$
  - $w_2 = -0.44002,$
  - $w_3 = -0.7208$
- Present data point 8
  - $\Delta y_8 = y_8 - \phi_8^T w$
- Adjust $\hat{w}$

- **Updated values**
  - $\hat{w} \leftarrow \hat{w} + \eta \phi_8 \Delta y_8$
Nonlinear Regression Example

- **Iteration 11**
  - $w_1 = 0.38071,$
  - $w_2 = -0.44002,$
  - $w_3 = -0.7208$
  - Present data point 8
  - $\Delta y_8 = y_8 - \phi_8^\top w$
  - Adjust $\hat{w}$

- **Updated values**
  $\hat{w} \leftarrow \hat{w} + \eta \phi_8 \Delta y_8$
Nonlinear Regression Example

- **Iteration 12**
  - \( w_1 = 0.37237, \)
  - \( w_2 = -0.90666, \)
  - \( w_3 = -1.1987 \)
  - **Present data point 5**
  - \( \Delta y_5 = y_5 - \phi_5^\top w \)
  - Adjust \( \hat{w} \)

- **Updated values**
  - \( \hat{w} \leftarrow \hat{w} + \eta \phi_5 \Delta y_5 \)
Nonlinear Regression Example

- **Iteration 12**
  - $w_1 = 0.37237$,
  - $w_2 = -0.90666$,
  - $w_3 = -1.1987$
- **Present data point 5**
  - $\Delta y_5 = y_5 - \phi_5^T w$
- **Adjust $\hat{w}$**

- **Updated values**
  - $\hat{w} \leftarrow \hat{w} + \eta \phi_5 \Delta y_5$
Nonlinear Regression Example

- **Iteration 13**
  - $w_1 = 0.62833$, $w_2 = -0.45691$, $w_3 = -1.1842$
  - Present data point 10
  - $\Delta y_{10} = y_{10} - \phi_{10}^\top w$
  - Adjust $\hat{w}$
- Updated values
  - $\hat{w} \leftarrow \hat{w} + \eta \phi_{10} \Delta y_{10}$
Nonlinear Regression Example

- Iteration 13
  - $w_1 = 0.62833,$
  - $w_2 = -0.45691,$
  - $w_3 = -1.1842$
- Present data point 10
- $\Delta y_{10} = y_{10} - \phi_1^\top w$
- Adjust $\hat{w}$

- Updated values
  - $\hat{w} \leftarrow \hat{w} + \eta \phi_{10} \Delta y_{10}$
Iteration 14

\[ w_1 = 0.62833, \]
\[ w_2 = -0.4575, \]
\[ w_3 = -1.252 \]

Present data point 2

\[ \Delta y_2 = y_2 - \phi_2^\top w \]

Adjust \( \hat{w} \)

Updated values

\[ \hat{w} \leftarrow \hat{w} + \eta \phi_2 \Delta y_2 \]
Iteration 14

- $w_1 = 0.62833$
- $w_2 = -0.4575$
- $w_3 = -1.252$

Present data point 2

- $\Delta y_2 = y_2 - \phi_2^T w$

Adjust $\hat{w}$

Updated values

$\hat{w} \leftarrow \hat{w} + \eta \phi_2 \Delta y_2$
Nonlinear Regression Example

- **Iteration 15**
  - $w_1 = 0.7016$, $w_2 = -0.45646$, $w_3 = -1.252$
  - Present data point 1
    - $\Delta y_1 = y_1 - \phi_1^\top w$
  - Adjust $\hat{w}$

- Updated values
  $$\hat{w} \leftarrow \hat{w} + \eta \phi_1 \Delta y_1$$
Nonlinear Regression Example

- **Iteration 15**
  - $w_1 = 0.7016,$
  - $w_2 = -0.45646,$
  - $w_3 = -1.252$
  - Present data point 1
  - $\Delta y_1 = y_1 - \phi_1^T w$
  - Adjust $\hat{w}$

- **Updated values**
  $\hat{w} \leftarrow \hat{w} + \eta \phi_1 \Delta y_1$
Nonlinear Regression Example

iteration 16

- $w_1 = 0.7109,$ 
  $w_2 = -0.45641,$ 
  $w_3 = -1.252$

- Present data point 5
  - $\Delta y_5 = y_5 - \phi_5^T w$

- Adjust $\hat{w}$

- Updated values
  - $\hat{w} \leftarrow \hat{w} + \eta \phi_5 \Delta y_5$
Nonlinear Regression Example

- **Iteration 16**
  - \( w_1 = 0.7109, \)
  - \( w_2 = -0.45641, \)
  - \( w_3 = -1.252 \)
- Present data point 5
- \( \Delta y_5 = y_5 - \phi_5^\top \mathbf{w} \)
- Adjust \( \hat{\mathbf{w}} \)
- Updated values
  - \( \hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \phi_5 \Delta y_5 \)
Nonlinear Regression Example

- **Iteration 17**
  - \( w_1 = 0.77022, \)
  - \( w_2 = -0.35219, \)
  - \( w_3 = -1.2487 \)
  - Present data point 9
  - \( \Delta y_9 = y_9 - \phi_9^\top w \)
  - Adjust \( \hat{w} \)
- **Updated values**
  \( \hat{w} \leftarrow \hat{w} + \eta \phi_9 \Delta y_9 \)
Iteration 17

- $w_1 = 0.77022$
- $w_2 = -0.35219$
- $w_3 = -1.2487$

Present data point 9

- $\Delta y_9 = y_9 - \phi_9^\top \hat{w}$

Adjust $\hat{w}$

Updated values

$\hat{w} \leftarrow \hat{w} + \eta \phi_9 \Delta y_9$
Nonlinear Regression Example

- **Iteration 18**
  - \( w_1 = 0.77019, \) \( w_2 = -0.3832, \) \( w_3 = -1.8175 \)
  - **Present data point 4**
  - \( \Delta y_4 = y_4 - \phi_4^\top w \)
  - **Adjust \( \hat{w} \)**
- **Updated values**
  - \( \hat{w} \leftarrow \hat{w} + \eta \phi_4 \Delta y_4 \)
Iteration 18

- $w_1 = 0.77019,$
- $w_2 = -0.3832,$
- $w_3 = -1.8175$

- Present data point 4
- $\Delta y_4 = y_4 - \phi_4^T w$
- Adjust $\hat{w}$

Updated values

$\hat{w} \leftarrow \hat{w} + \eta \phi_4 \Delta y_4$
Iteration 19

- \( w_1 = 0.86321 \)
- \( w_2 = -0.28046 \)
- \( w_3 = -1.8154 \)

- Present data point 7
- \( \Delta y_7 = y_7 - \phi_7^\top w \)
- Adjust \( \hat{w} \)

- Updated values
  \( \hat{w} \leftarrow \hat{w} + \eta \phi_7 \Delta y_7 \)
Nonlinear Regression Example

- **Iteration 19**
  - $w_1 = 0.86321$, 
    $w_2 = -0.28046$, 
    $w_3 = -1.8154$
  - Present data point 7
  - $\Delta y_7 = y_7 - \phi_7^T w$
  - Adjust $\hat{w}$

- **Updated values**
  $$\hat{w} \leftarrow \hat{w} + \eta \phi_7 \Delta y_7$$
Nonlinear Regression Example

- **Iteration 20**
  - \( w_1 = 0.80681 \),
  - \( w_2 = -0.47597 \),
  - \( w_3 = -1.8278 \)
- **Present data point 6**
  - \( \Delta y_6 = y_6 - \phi_6^\top w \)
  - Adjust \( \hat{w} \)
- **Updated values**
  - \( \hat{w} \leftarrow \hat{w} + \eta \phi_6 \Delta y_6 \)
Nonlinear Regression Example

- **Iteration 20**
  - \( w_1 = 0.80681 \)
  - \( w_2 = -0.47597 \)
  - \( w_3 = -1.8278 \)
- **Present data point 6**
- \( \Delta y_6 = y_6 - \phi_6^\top w \)
- **Adjust \( \hat{w} \)**
- **Updated values**
  - \( \hat{w} \leftarrow \hat{w} + \eta \phi_6 \Delta y_6 \)
Nonlinear Regression Example

- Iteration 50
  - $w_1 = 0.9777,$
  - $w_2 = -0.4076,$
  - $w_3 = -2.038$
- Present data point 8
- $\Delta y_8 = y_8 - \phi_8^T w$
- Adjust $\hat{w}$
- Updated values
  - $\hat{w} \leftarrow \hat{w} + \eta \phi_8 \Delta y_8$
Nonlinear Regression Example

- **Iteration 100**
  - $w_1 = 0.98593$, $w_2 = -0.49744$, $w_3 = -2.046$
  - Present data point 8
  - $\Delta y_8 = y_8 - \phi_8^T w$
  - Adjust $\hat{w}$

- **Updated values**
  \[ \hat{w} \leftarrow \hat{w} + \eta \phi_8 \Delta y_8 \]
Nonlinear Regression Example

- **Iteration 200**
  - $w_1 = 0.95307$, $w_2 = -0.48041$, $w_3 = -2.0553$
- Present data point 4
  - $\Delta y_4 = y_4 - \phi_4^T w$
- Adjust $\hat{w}$

**Updated values**
\[ \hat{w} \leftarrow \hat{w} + \eta \phi_4 \Delta y_4 \]
Nonlinear Regression Example

- **Iteration 300**
  - $w_1 = 0.97066,$
  - $w_2 = -0.44667,$
  - $w_3 = -2.0588$
- **Present data point 1**
- **$\Delta y_1 = y_1 - \phi_1^\top w$**
- **Adjust $\hat{w}$**
- **Updated values**
  - $\hat{w} \leftarrow \hat{w} + \eta \phi_1 \Delta y_1$
Nonlinear Regression Example

- Iteration 400
  - \( w_1 = 0.95515 \),
  - \( w_2 = -0.40611 \),
  - \( w_3 = -2.0289 \)
  - Present data point 8
  - \( \Delta y_8 = y_8 - \phi_8^\top w \)
  - Adjust \( \hat{w} \)

- Updated values
  \[ \hat{w} \leftarrow \hat{w} + \eta \phi_8 \Delta y_8 \]
Nonlinear Regression Example

- **Iteration 500**
  - $w_1 = 0.94178$, $w_2 = -0.49879$, $w_3 = -1.9209$
  - Present data point 5
  - $\Delta y_5 = y_5 - \phi_5^T \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$

- **Updated values**
  - $\mathbf{w} \leftarrow \hat{\mathbf{w}} + \eta \phi_5 \Delta y_5$
What is the mathematical interpretation?

- There is a cost function.
- It expresses mismatch between your prediction and reality.

\[ E(w) = \sum_{i=1}^{n} \left( \sum_{j=1}^{K} w_j \phi_j(x_i) - y_i \right)^2 \]

- This is known as the sum of squares error.
What is the mathematical interpretation?

- There is a cost function.
- It expresses mismatch between your prediction and reality.

\[ E(w) = \sum_{i=1}^{n} (w^\top \phi_i - y_i)^2 \]

- This is known as the sum of squares error.
- Defining \( \phi_i = [\phi_1(x_i), \ldots, \phi_K(x_i)]^\top \).
Learning is minimization of the cost function.

At the minima the gradient is zero.

Gradient of error function:

$$\frac{dE(w)}{dw} = -2 \sum_{i=1}^{n} \phi_i \left( y_i - w^\top \phi_i \right)$$
Learning is minimization of the cost function.

At the minima the gradient is zero.

Gradient of error function:

\[
\frac{dE(w)}{dw} = -2 \sum_{i=1}^{n} \phi_i \Delta y_i
\]

Where \( \Delta y_i = (y_i - w^\top \phi_i) \).
One way of minimizing is steepest descent.

Initialize algorithm with $\mathbf{w}$.

Compute gradient of error function, $\frac{dE(\mathbf{w})}{d\mathbf{w}}$.

Change $\mathbf{w}$ by moving in steepest downhill direction.

$$
\mathbf{w} \leftarrow \mathbf{w} - \eta \frac{dE(\mathbf{w})}{d\mathbf{w}}
$$
Figure: Steepest descent on a quadratic error surface.
Figure: Steepest descent on a quadratic error surface.
Figure: Steepest descent on a quadratic error surface.
Figure: Steepest descent on a quadratic error surface.
Steepest Descent

Figure: Steepest descent on a quadratic error surface.
Figure: Steepest descent on a quadratic error surface.
Figure: Steepest descent on a quadratic error surface.
Steepest Descent

Figure: Steepest descent on a quadratic error surface.
Figure: Steepest descent on a quadratic error surface.
**Steepest Descent**

**Figure:** Steepest descent on a quadratic error surface.
Figure: Steepest descent on a quadratic error surface.
Figure: Steepest descent on a quadratic error surface.
Figure: Steepest descent on a quadratic error surface.
Figure: Steepest descent on a quadratic error surface.
Figure: Steepest descent on a quadratic error surface.
Figure: Steepest descent on a quadratic error surface.
Figure: Steepest descent on a quadratic error surface.
Figure: Steepest descent on a quadratic error surface.
Figure: Steepest descent on a quadratic error surface.
Figure: Steepest descent on a quadratic error surface.
Figure: Steepest descent on a quadratic error surface.
Figure: Steepest descent on a quadratic error surface.
For regression, the learning rule can be seen as a variant of gradient descent.

This variant is known as stochastic gradient descent.

For regression steepest descent gives

\[ w \leftarrow w - \eta \frac{dE(w)}{dw} \]
For regression, the learning rule can be seen as a variant of gradient descent.

This variant is known as stochastic gradient descent.

For regression steepest descent gives

\[ \mathbf{w} \leftarrow \mathbf{w} - 2\eta \sum_{i=1}^{n} \phi_i \left( \mathbf{w}^\top \phi_i - y_i \right) \]
For regression, the learning rule can be seen as a variant of gradient descent.

This variant is known as stochastic gradient descent.

For regression steepest descent gives

\[ w \leftarrow w - \eta' \sum_{i=1}^{n} \phi_i \left( w^\top \phi_i - y_i \right) \]
For regression, the learning rule can be seen as a variant of gradient descent.

This variant is known as stochastic gradient descent.

For regression steepest descent gives

\[ w \leftarrow w - \eta' \sum_{i=1}^{n} \phi_i \left( w^\top \phi_i - y_i \right) \]
For regression, the learning rule can be seen as a variant of gradient descent.

This variant is known as stochastic gradient descent.

For regression steepest descent gives

$$w \leftarrow w - \eta' \sum_{i=1}^{n} \phi_i \Delta y_i$$
Stochastic Gradient Descent

How does this relate to learning rules we presented?

- For regression, the learning rule can be seen as a variant of gradient descent.
- This variant is known as stochastic gradient descent.
- For regression steepest descent gives

\[
\mathbf{w} \leftarrow \mathbf{w} - \eta' \sum_{i=1}^{n} \phi_i \Delta y_i
\]

- And the stochastic approximation is

\[
\mathbf{w} \leftarrow \mathbf{w} + \eta' \phi_i \Delta y_i
\]
Figure: Stochastic gradient descent on a quadratic error surface.
Figure: Stochastic gradient descent on a quadratic error surface.
Figure: Stochastic gradient descent on a quadratic error surface.
Figure: Stochastic gradient descent on a quadratic error surface.
Stochastic Gradient Descent

Figure: Stochastic gradient descent on a quadratic error surface.
Figure: Stochastic gradient descent on a quadratic error surface.
Figure: Stochastic gradient descent on a quadratic error surface.
Figure: Stochastic gradient descent on a quadratic error surface.
Figure: Stochastic gradient descent on a quadratic error surface.
Stochastic Gradient Descent

Figure: Stochastic gradient descent on a quadratic error surface.
Figure: Stochastic gradient descent on a quadratic error surface.
**Figure:** Stochastic gradient descent on a quadratic error surface.
Figure: Stochastic gradient descent on a quadratic error surface.
Stochastic Gradient Descent

Figure: Stochastic gradient descent on a quadratic error surface.
Figure: Stochastic gradient descent on a quadratic error surface.
Stochastic Gradient Descent

Figure: Stochastic gradient descent on a quadratic error surface.
Figure: Stochastic gradient descent on a quadratic error surface.
Figure: Stochastic gradient descent on a quadratic error surface.
Figure: Stochastic gradient descent on a quadratic error surface.
Stochastic Gradient Descent

Figure: Stochastic gradient descent on a quadratic error surface.
Figure: Stochastic gradient descent on a quadratic error surface.
Figure: Stochastic gradient descent on a quadratic error surface.
Figure: Stochastic gradient descent on a quadratic error surface.
Stochastic Gradient Descent

Figure: Stochastic gradient descent on a quadratic error surface.
Figure: Stochastic gradient descent on a quadratic error surface.
Figure: Stochastic gradient descent on a quadratic error surface.
Figure: Stochastic gradient descent on a quadratic error surface.
Figure: Stochastic gradient descent on a quadratic error surface.
Figure: Stochastic gradient descent on a quadratic error surface.
Figure: Stochastic gradient descent on a quadratic error surface.
Figure: Stochastic gradient descent on a quadratic error surface.
Figure: Stochastic gradient descent on a quadratic error surface.
Error function has a probabilistic interpretation (maximum likelihood).

Error function is an actual loss function that you want to minimize (empirical risk minimization).

For these interpretations probability and optimization theory become important.

Much of the last 15 years of machine learning research has focused on probabilistic interpretations or clever relaxations of difficult objective functions.
Optimization methods.
  - Second order methods, conjugate gradient, quasi-Newton and Newton.
  - Effective heuristics such as momentum.

Local vs global solutions.
Clustering

- Divide data into discrete groups according to characteristics.
  - For example different animal species.
  - Different political parties.
- Determine the allocation to the groups and (harder) number of different groups.
**K-means Clustering**

An Algorithm

- **Require:** Set of $K$ cluster centers & assignment of each point to a cluster.
  - Initialize cluster centers as data points.
  - Assign each data point to nearest cluster center.
  - Update each cluster center by setting it to the mean of assigned data points.
This minimizes the objective:

\[
\sum_{j=1}^{K} \sum_{i \text{ allocated to } j} (y_{i,:} - \mu_{j,:})^\top (y_{i,:} - \mu_{j,:})
\]

i.e. it minimizes the sum of Euclidean squared distances between points and their associated centers.

The minimum is not guaranteed to be \textit{global} or \textit{unique}.

This objective is a non-convex optimization problem.
K-means Clustering

- K-means clustering.
  - Data set to be analyzed. Initialize cluster centers.
**K-means Clustering**

- **K-means clustering.**
  - Allocate each point to the cluster with the nearest center
- $K$-means clustering.
  - Update each center by setting to the mean of the allocated points.
**K-means Clustering**

- K-means clustering.
  - Allocate each data point to the nearest cluster center.

![Iteration 1](image-url)
$K$-means Clustering

- $K$-means clustering.
  - Update each center by setting to the mean of the allocated points.
K-means Clustering

- K-means clustering.
  - Allocate each data point to the nearest cluster center.
K-means Clustering

- K-means clustering.
  - Update each center by setting to the mean of the allocated points.
- $K$-means clustering.
  - Allocate each data point to the nearest cluster center.
**K-means Clustering**

- K-means clustering.
  - Update each center by setting to the mean of the allocated points.
K-means Clustering

- K-means clustering.
  - Allocate each data point to the nearest cluster center.
- $K$-means clustering.
  - Allocation doesn’t change so stop.
Other Clustering Approaches

- Spectral clustering (Shi and Malik, 2000; Ng et al., 2002).
  - Allows clusters which aren’t convex hulls.
- Dirichlet processes
  - A probabilistic formulation for a clustering algorithm that is non-parameteric.