Introduction to Machine Learning

Introduction

Andres Mendez-Vazquez

February 4, 2019

Outline

Outline

Data is being produced in great quantities

After all our business is about

Big Data

Definition

Big data is a term for data sets that are so large or complex that traditional data processing application software is inadequate to deal with them.

Data is being produced in great quantities

After all our business is about

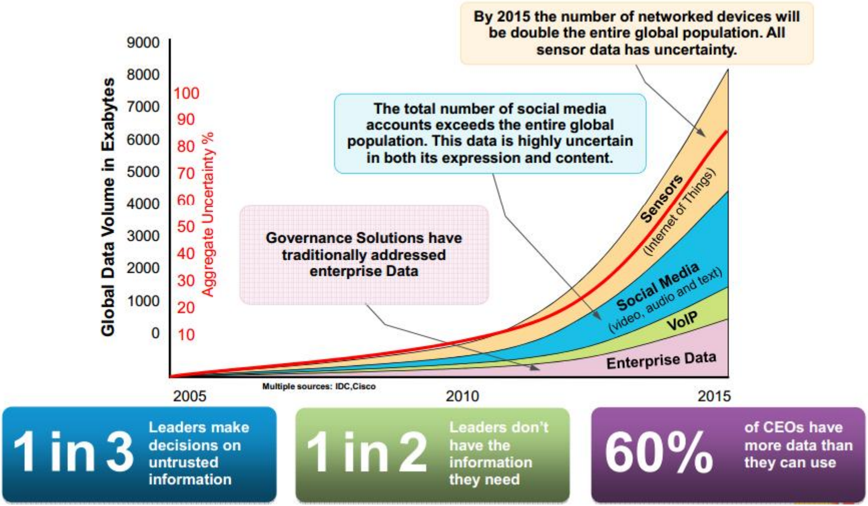
Big Data

Definition

Big data is a term for data sets that are so large or complex that traditional data processing application software is inadequate to deal with them.

Outline

VOLUME


By 2015 the number of networked devices will be double the entire global population. All sensor data has uncertainty.
The total number of social media accounts exceeds the entire global population. This data is highly uncertain in both its expression and content.
Governance Solutions have traditionally addressed enterprise Data
Global Data Volume in Exabytes
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
Aggregate Uncertainty %
100
90
80
70
60
50
40
30
20
10
Sensors
(Internet of Things)
Social Media
(video, audio and text)
VoIP
Enterprise Data
Multiple sources: IDC,Cisco
2005
2010
2015
1 in 3
Leaders make decisions on untrusted information
1 in 2
Leaders don't have the information they need
60%
of CEOs have more data than they can use

VOLUME

VOLUMES of Information

- Terabyte(10^{12} bits),
- Petabyte(10^{15} bits),
- UP!!!

VOLUME

VOLUMES of Information

- Terabyte(10^{12} bits),
- Petabyte(10^{15} bits),
- UP!!!

Examples of these Volumes are

1. Records
2. Transactions
3. Web Searches
4. etc

VOLUME

VOLUMES of Information

- Terabyte(10^{12} bits),
- Petabyte(10^{15} bits),
- UP!!!

Examples of these Volumes are

1. Records
2. Transactions
3. Web Searches
4. etc

VOLUME

VOLUMES of Information

- Terabyte(10^{12} bits),
- Petabyte(10^{15} bits),
- UP!!!

Examples of these Volumes are

1. Records
2. Transactions
3. Web Searches
4. etc

VOLUME

VOLUMES of Information

- Terabyte(10^{12} bits),
- Petabyte(10^{15} bits),
- UP!!!

Examples of these Volumes are

1. Records
2. Transactions
3. Web Searches
4. etc

VOLUME

VOLUMES of Information

- Terabyte(10^{12} bits),
- Petabyte(10^{15} bits),
- UP!!!

Examples of these Volumes are

1. Records
2. Transactions
3. Web Searches
4. etc

VOLUME

VOLUMES of Information

- Terabyte(10^{12} bits),
- Petabyte(10^{15} bits),
- UP!!!

Examples of these Volumes are

1. Records
2. Transactions
3. Web Searches
4. etc

However

Cautionary Tale

What constitutes truly “high” volume varies by industry and even geography!!!

- Simply look at the DNA data for a cellular cycle.

However

Cautionary Tale

What constitutes truly “high” volume varies by industry and even geography!!!

- Simply look at the DNA data for a cellular cycle.

Example

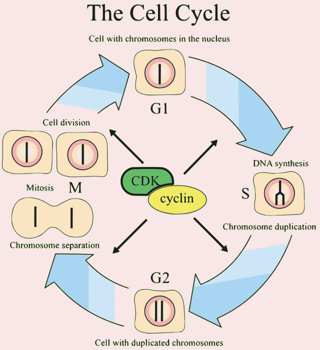
However

Cautionary Tale

What constitutes truly "high" volume varies by industry and even geography!!!

- Simply look at the DNA data for a cellular cycle.

Example



VARIETY

VARIETY

When looking at the Structure of the Information, we have:

- **Variety** like there is not tomorrow:
  - It is structured, semi-structured and unstructured

VARIETY

When looking at the Structure of the Information, we have:

- **Variety** like there is not tomorrow:
  - It is structured, semi-structured and unstructured

Question

- Do you have some examples of such structures in Information?

VARIETY

When looking at the Structure of the Information, we have:

- **Variety** like there is not tomorrow:
  - It is structured, semi-structured and unstructured

Question

- Do you have some examples of such structures in Information?

VELOCITY

VELOCITY

It refers to

- The **SPEED** at which the data is being generated.
- The **SPEED** at which the data moves around

VELOCITY

It refers to

- The **SPEED** at which the data is being generated.
- The **SPEED** at which the data moves around

Problem: Latency

- There is a **LAG TIME** between capture or generation, and when it is available!!!

- Detecting when sale and buy shares
- etc

VELOCITY

It refers to

- The **SPEED** at which the data is being generated.
- The **SPEED** at which the data moves around

Problem: Latency

- There is a **LAG TIME** between capture or generation, and when it is available!!!

It can be catastrophic

- Detecting fraudulent activities
- Detecting when sale and buy shares
- etc

VELOCITY

It refers to

- The **SPEED** at which the data is being generated.
- The **SPEED** at which the data moves around

Problem: Latency

- There is a **LAG TIME** between capture or generation, and when it is available!!!

It can be catastrophic

- Detecting fraudulent activities
- Detecting when sale and buy shares
- etc

VELOCITY

It refers to

- The **SPEED** at which the data is being generated.
- The **SPEED** at which the data moves around

Problem: Latency

- There is a **LAG TIME** between capture or generation, and when it is available!!!

It can be catastrophic

- Detecting fraudulent activities
- Detecting when sale and buy shares
- etc

VELOCITY

It refers to

- The **SPEED** at which the data is being generated.
- The **SPEED** at which the data moves around

Problem: Latency

- There is a **LAG TIME** between capture or generation, and when it is available!!!

It can be catastrophic

- Detecting fraudulent activities
- Detecting when sale and buy shares
- etc

For Example

Imagine that I have a stream of $m = 10^{30}$ integers with Ranges in $[1, ..., 10^8]$

Now, somebody ask you to find the most frequent item!!!

For Example

Imagine that I have a stream of $m = 10^{30}$ integers with Ranges in $[1, ..., 10^8]$

Now, somebody ask you to find the most frequent item!!!

A naive algorithm

1. Take hash table with a counter.
2. Then, put numbers in the hash table.

For Example

Imagine that I have a stream of $m = 10^{30}$ integers with Ranges in $[1, ..., 10^8]$

Now, somebody ask you to find the most frequent item!!!

A naive algorithm

1. Take hash table with a counter.
2. Then, put numbers in the hash table.

Problems

Which problems we have?

For Example

Imagine that I have a stream of $m = 10^{30}$ integers with Ranges in $[1, ..., 10^8]$

Now, somebody ask you to find the most frequent item!!!

A naive algorithm

1. Take hash table with a counter.
2. Then, put numbers in the hash table.

Problems

Which problems we have?

However

There is the

Count-Min Sketch Algorithm

Invented by

Charikar, Chen and Farch-Colton in 2004

With Properties

Space Used	Error Probability	Error
$O\left(\frac{1}{\epsilon}\log\left(\frac{1}{\delta}\right)\cdot(\log m+\log n)\right)$	δ	ϵ

However

There is the

Count-Min Sketch Algorithm

Invented by

Charikar, Chen and Farch-Colton in 2004

With Properties

Space Used	Error Probability	Error
$O\left(\frac{1}{\epsilon}\log\left(\frac{1}{\delta}\right)\cdot(\log m+\log n)\right)$	δ	ϵ

However

There is the

Count-Min Sketch Algorithm

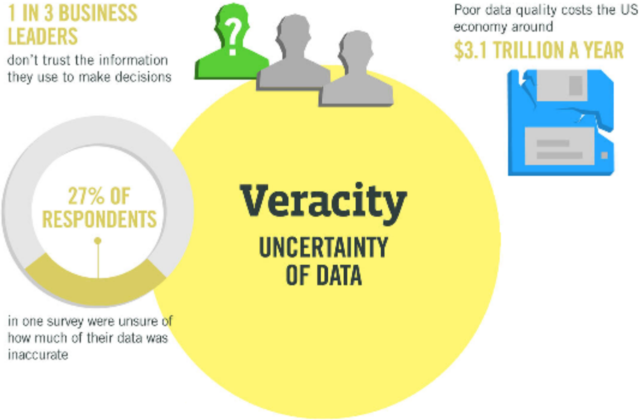
Invented by

Charikar, Chen and Farch-Colton in 2004

With Properties

Space Used	Error Probability	Error
$O\left(\frac{1}{\epsilon}\log\left(\frac{1}{\delta}\right)\cdot(\log m + \log n)\right)$	δ	ϵ

VERACITY


1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions
27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate
Veracity
UNCERTAINTY OF DATA
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR

VERACITY

It refers to

Messiness or Trustworthiness of the data

- The Completeness on the Data
- The Representation of such Data
- etc!!!

VERACITY

It refers to

Messiness or Trustworthiness of the data

However, some people claim

“**The volumes often make up for the lack of quality or accuracy**”

- The Completeness on the Data
- The Representation of such Data
- etc!!!

VERACITY

It refers to

Messiness or Trustworthiness of the data

However, some people claim

“**The volumes often make up for the lack of quality or accuracy**”

I say, it depends on

- The Noise on the Data
- The Completeness on the Data
- The Representation of such Data
- etc!!!

VERACITY

It refers to

Messiness or Trustworthiness of the data

However, some people claim

“**The volumes often make up for the lack of quality or accuracy**”

I say, it depends on

- The Noise on the Data
- The Completeness on the Data
- The Representation of such Data
- etc!!!

VERACITY

It refers to

Messiness or Trustworthiness of the data

However, some people claim

“**The volumes often make up for the lack of quality or accuracy**”

I say, it depends on

- The Noise on the Data
- The Completeness on the Data
- The Representation of such Data
- etc!!!

VERACITY

It refers to

Messiness or Trustworthiness of the data

However, some people claim

"**The volumes often make up for the lack of quality or accuracy**"

I say, it depends on

- The Noise on the Data
- The Completeness on the Data
- The Representation of such Data
- etc!!!

Outline

Therefore

Given all these things

- It is necessary to correlate and share data across entities.
- It is necessary to link, match and transform data.

Therefore

Given all these things

- It is necessary to correlate and share data across entities.
- It is necessary to link, match and transform data.

With this...

Complexity goes through the roof!!!

Therefore

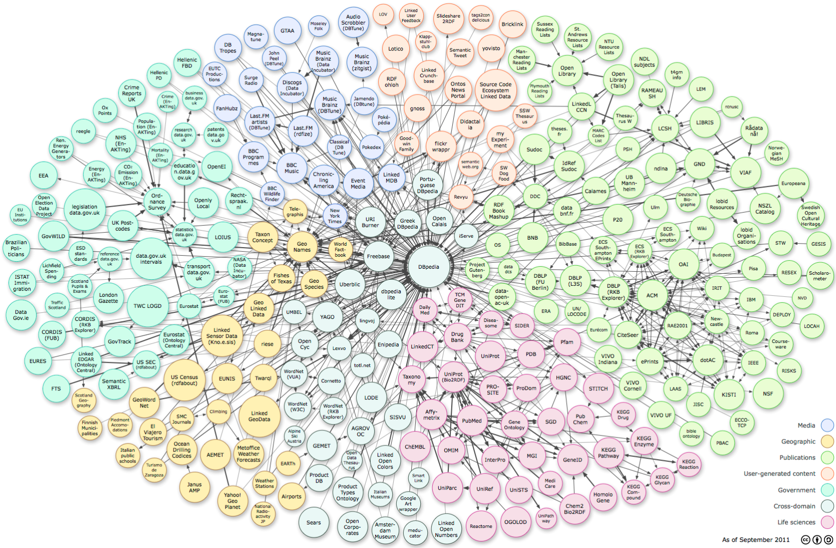
Given all these things

- It is necessary to correlate and share data across entities.
- It is necessary to link, match and transform data.

With this...

Complexity goes through the roof!!!

Example: Linking open-data community project



Cautionary Tale about Complexity

Something Notable

- In 1880 the USA made a Census of the Population in different aspects:
    - Population
    - Mortality
    - Agriculture
    - Manufacturing

Cautionary Tale about Complexity

Something Notable

- In 1880 the USA made a Census of the Population in different aspects:
  - Population
  - Mortality
  - Agriculture
  - Manufacturing

However

Once data was collected it took 7 years to say something!!!

Cautionary Tale about Complexity

Something Notable

- In 1880 the USA made a Census of the Population in different aspects:
  - Population
  - Mortality
  - Agriculture
  - Manufacturing

However

Once data was collected it took 7 years to say something!!!

Cautionary Tale about Complexity

Something Notable

- In 1880 the USA made a Census of the Population in different aspects:
  - Population
  - Mortality
  - Agriculture
  - Manufacturing

However

Once data was collected it took 7 years to say something!!!

Cautionary Tale about Complexity

Something Notable

- In 1880 the USA made a Census of the Population in different aspects:
  - Population
  - Mortality
  - Agriculture
  - Manufacturing

However

Once data was collected it took 7 years to say something!!!

Cautionary Tale about Complexity

Something Notable

- In 1880 the USA made a Census of the Population in different aspects:
  - Population
  - Mortality
  - Agriculture
  - Manufacturing

However

Once data was collected it took 7 years to say something!!!

Cautionary Tale about Complexity

Something Notable

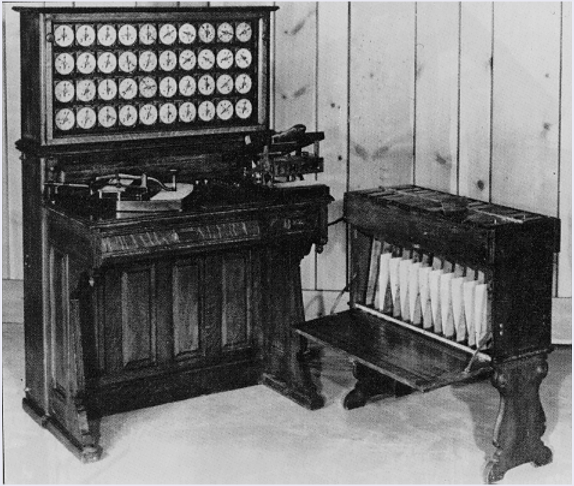
- In 1880 the USA made a Census of the Population in different aspects:
  - Population
  - Mortality
  - Agriculture
  - Manufacturing

However

Once data was collected it took 7 years to say something!!!

The Tabulator Machine

Thus, Hollering came with the following machine (Circa 1890)!!!



Hollering Tabulating Machine

It was basically a sorter and counter

- Using punching cards as memories.
- And Mercury Sensors.

Hollering Tabulating Machine

It was basically a sorter and counter

- Using punching cards as memories.
- And Mercury Sensors.

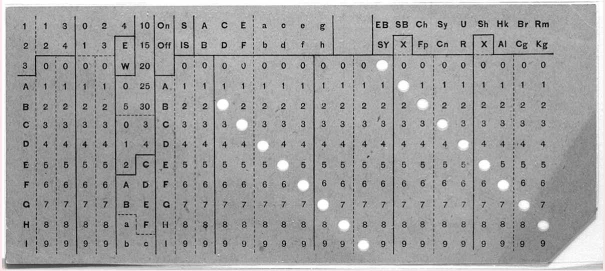
Example

Hollering Tabulating Machine

It was basically a sorter and counter

- Using punching cards as memories.
- And Mercury Sensors.

Example



It was FAST!!!

It took only!!!

2 years!!!

It was FAST!!!

It took only!!!

2 years!!!

Nevertheless in 1837

Babbage's Difference engine was

- The First General Computer!!!
- Turing-complete!!!
- **Way more complex than the tabulator!!! 53 years earlier!!!**

It was FAST!!!

It took only!!!

2 years!!!

Nevertheless in 1837

Babbage's Difference engine was

- The First General Computer!!!
- Turing-complete!!!
- Way more complex than the tabulator!!! 53 years earlier!!!

It was FAST!!!

It took only!!!

2 years!!!

Nevertheless in 1837

Babbage's Difference engine was

- The First General Computer!!!
- Turing-complete!!!
- Way more complex than the tabulator!!! 53 years earlier!!!

It was FAST!!!

It took only!!!

2 years!!!

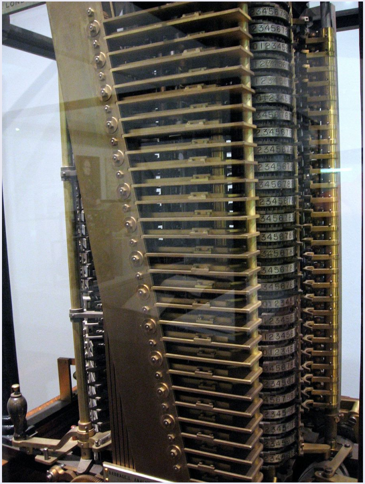
Nevertheless in 1837

Babbage's Difference engine was

- The First General Computer!!!
- Turing-complete!!!
- **Way more complex than the tabulator!!! 53 years earlier!!!**

Therefore

Complexity is highly **DEPENDANT** on the way data is handled and represented



Outline

Data is Everywhere!

Lots of data is being collected and warehoused

- Web data, e-commerce
- Purchases at department/ grocery stores
- Bank/Credit Card transactions
- Social Network
- etc

Data is Everywhere!

Lots of data is being collected and warehoused

- Web data, e-commerce
- Purchases at department/ grocery stores
- Bank/Credit Card transactions
- Social Network
- etc

Data is Everywhere!

Lots of data is being collected and warehoused

- Web data, e-commerce
- Purchases at department/ grocery stores
- Bank/Credit Card transactions
- Social Network
- etc

Data is Everywhere!

Lots of data is being collected and warehoused

- Web data, e-commerce
- Purchases at department/ grocery stores
- Bank/Credit Card transactions
- Social Network
- etc

Data is Everywhere!

Lots of data is being collected and warehoused

- Web data, e-commerce
- Purchases at department/ grocery stores
- Bank/Credit Card transactions
- Social Network
- etc

You have Staggering Numbers

A Ocean of Data

- How many data in the world?
  - 800 Terabytes, 2000
  - 160 Exabytes, 2006
  - 500 Exabytes (Internet), 2009
  - 2.7 Zettabytes, 2012
  - 35 Zettabytes by 2020

- 7 TB, Twitter
- 10 TB, Facebook

You have Staggering Numbers

A Ocean of Data

- How many data in the world?
  - 800 Terabytes, 2000
  - 160 Exabytes, 2006
  - 500 Exabytes (Internet), 2009
  - 2.7 Zettabytes, 2012
  - 35 Zettabytes by 2020

How much?

- How many data generated ONE day?
  - 7 TB, Twitter
  - 10 TB, Facebook

You have Staggering Numbers

A Ocean of Data

- How many data in the world?
  - 800 Terabytes, 2000
  - 160 Exabytes, 2006
  - 500 Exabytes (Internet), 2009
  - 2.7 Zettabytes, 2012
  - 35 Zettabytes by 2020

How much?

- How many data generated ONE day?
  - 7 TB, Twitter
  - 10 TB, Facebook

You have Staggering Numbers

A Ocean of Data

- How many data in the world?
    - 800 Terabytes, 2000
    - 160 Exabytes, 2006
    - 500 Exabytes (Internet), 2009
    - 2.7 Zettabytes, 2012
    - 35 Zettabytes by 2020

How much?

- How many data generated ONE day?
    - 7 TB, Twitter
    - 10 TB, Facebook

You have Staggering Numbers

A Ocean of Data

- How many data in the world?
  - 800 Terabytes, 2000
  - 160 Exabytes, 2006
  - 500 Exabytes (Internet), 2009
  - 2.7 Zettabytes, 2012
  - 35 Zettabytes by 2020

How much?

- How many data generated ONE day?
  - 7 TB, Twitter
  - 10 TB, Facebook

You have Staggering Numbers

A Ocean of Data

- How many data in the world?
  - 800 Terabytes, 2000
  - 160 Exabytes, 2006
  - 500 Exabytes (Internet), 2009
  - 2.7 Zettabytes, 2012
  - 35 Zettabytes by 2020

How much?

- How many data generated ONE day?
  - 7 TB, Twitter
  - 10 TB, Facebook

You have Staggering Numbers

A Ocean of Data

- How many data in the world?
  - 800 Terabytes, 2000
  - 160 Exabytes, 2006
  - 500 Exabytes (Internet), 2009
  - 2.7 Zettabytes, 2012
  - 35 Zettabytes by 2020

How much?

- How many data generated ONE day?
  - 7 TB, Twitter
  - 10 TB, Facebook

You have Staggering Numbers

A Ocean of Data

- How many data in the world?
  - 800 Terabytes, 2000
  - 160 Exabytes, 2006
  - 500 Exabytes (Internet), 2009
  - 2.7 Zettabytes, 2012
  - 35 Zettabytes by 2020

How much?

- How many data generated ONE day?
  - 7 TB, Twitter
  - 10 TB, Facebook

You have Staggering Numbers

A Ocean of Data

- How many data in the world?
  - 800 Terabytes, 2000
  - 160 Exabytes, 2006
  - 500 Exabytes (Internet), 2009
  - 2.7 Zettabytes, 2012
  - 35 Zettabytes by 2020

How much?

- How many data generated ONE day?
  - 7 TB, Twitter
  - 10 TB, Facebook

So Big Data without Analytic Tools is basically...

"A Great!!! I am storing a bunch of data, so what?"

You require to have some way to get insights on such data sets

You need

- Algorithms to find those insights that are useful
- You need to apply them in the Large Data Set context

Here, it comes the Darling

Machine Learning!!! The Darling of Computer Science!!!

So Big Data without Analytic Tools is basically...

"A Great!!! I am storing a bunch of data, so what?"

You require to have some way to get insights on such data sets

You need

- Algorithms to find those insights that are useful
- You need to apply them in the Large Data Set context

Here, it comes the Darling

Machine Learning!!! The Darling of Computer Science!!!

So Big Data without Analytic Tools is basically...

"A Great!!! I am storing a bunch of data, so what?"

You require to have some way to get insights on such data sets

You need

- Algorithms to find those insights that are useful
- You need to apply them in the Large Data Set context

Here, it comes the Darling

Machine Learning!!! The Darling of Computer Science!!!

Machine Learning

Definition

Machine learning is the subfield of computer science that, according to Arthur Samuel, gives
"***computers the ability to learn without being explicitly programmed.***"

Machine Learning

Definition

Machine learning is the subfield of computer science that, according to Arthur Samuel, gives
"***computers the ability to learn without being explicitly programmed.***"

Therefore

- Learning Algorithms are devised to learn the input data

- Support Vector Machine (SVM)
- Expectation-Maximization (EM)

Machine Learning

Definition

Machine learning is the subfield of computer science that, according to Arthur Samuel, gives
"***computers the ability to learn without being explicitly programmed.***"

Therefore

- Learning Algorithms are devised to learn the input data

Examples

- Artificial Neural Network (ANN)
- Support Vector Machine (SVM)
- Expectation-Maximization (EM)

Machine Learning

Definition

Machine learning is the subfield of computer science that, according to Arthur Samuel, gives
"***computers the ability to learn without being explicitly programmed.***"

Therefore

- Learning Algorithms are devised to learn the input data

Examples

- Artificial Neural Network (ANN)
- Support Vector Machine (SVM)
- Expectation-Maximization (EM)

Machine Learning

Definition

Machine learning is the subfield of computer science that, according to Arthur Samuel, gives
"***computers the ability to learn without being explicitly programmed.***"

Therefore

- Learning Algorithms are devised to learn the input data

Examples

- Artificial Neural Network (ANN)
- Support Vector Machine (SVM)
- Expectation-Maximization (EM)

Outline

Outline

Supervised Learning

Basically

- Inputs and their desired outputs are given $\left\{(x_i, y_i)_{i=1}^{N}\right\}$
- Then, the goal is to learn a general rule that maps inputs to outputs

$$f : X \to Y \text{ with } f(x_i) = y_i + \epsilon$$

Supervised Learning

Basically

- Inputs and their desired outputs are given $\left\{(x_i, y_i)_{i=1}^{N}\right\}$
- Then, the goal is to learn a general rule that maps inputs to outputs

$$f : X \rightarrow Y \text{ with } f(x_i) = y_i + \epsilon$$

Supervised Learning

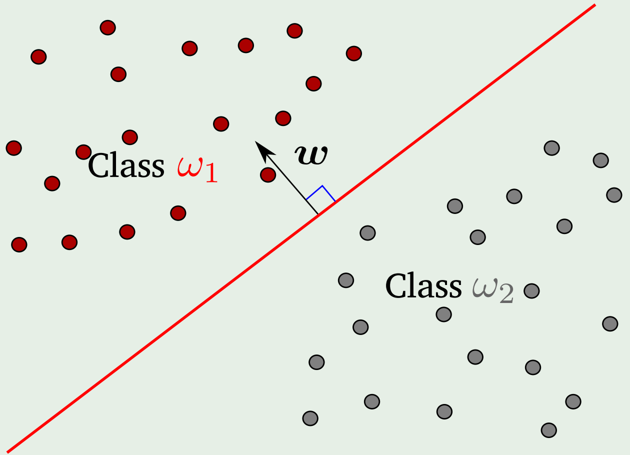
Basically

- Inputs and their desired outputs are given $\left\{(x_i, y_i)_{i=1}^{N}\right\}$
- Then, the goal is to learn a general rule that maps inputs to outputs

$$f : X \to Y \text{ with } f(x_i) = y_i + \epsilon$$

Classifying two classes in $\mathbb{R}^2$

Using a simple straight line



Outline

Unsupervised Learning

Basically

No labels are given to the learning algorithm, leaving it on its own to find structure in its input.

$$SSE = \sum_{k=1}^{K} \sum_{x_i \in C_k} dist(x_i, c_k)^2$$

- Similarities

$$dist(x, y) = \|x - y\|_2 = \sqrt{(x - y)^T (x - y)}$$

- etc

Unsupervised Learning

Basically

No labels are given to the learning algorithm, leaving it on its own to find structure in its input.

Therefore, it is necessary to find the clusters of Data

- Using Cost Functions

$$SSE = \sum_{k=1}^{K} \sum_{x \in c_k} dist\left(\mathbf{x}, \mathbf{v}_k\right)^2$$

- Similarities

$$dist(\mathbf{x}, \mathbf{y}) = ||\mathbf{x} - \mathbf{y}||_A = \sqrt{(\mathbf{x} - \mathbf{y})^T A (\mathbf{x} - \mathbf{y})}$$

- etc

Unsupervised Learning

Basically

No labels are given to the learning algorithm, leaving it on its own to find structure in its input.

Therefore, it is necessary to find the clusters of Data

- Using Cost Functions

$$SSE = \sum_{k=1}^{K} \sum_{x \in c_k} dist\left(\mathbf{x}, \mathbf{v}_k\right)^2$$

- Similarities

$$dist(\mathbf{x}, \mathbf{y}) = ||\mathbf{x} - \mathbf{y}||_A = \sqrt{(\mathbf{x} - \mathbf{y})^T A (\mathbf{x} - \mathbf{y})}$$

- etc

Unsupervised Learning

Basically

No labels are given to the learning algorithm, leaving it on its own to find structure in its input.

Therefore, it is necessary to find the clusters of Data

- Using Cost Functions

$$SSE = \sum_{k=1}^{K} \sum_{x \in c_k} dist\left(\mathbf{x}, \mathbf{v}_k\right)^2$$

- Similarities

$$dist(\mathbf{x}, \mathbf{y}) = ||\mathbf{x} - \mathbf{y}||_A = \sqrt{(\mathbf{x} - \mathbf{y})^T A (\mathbf{x} - \mathbf{y})}$$

- etc

Unsupervised Learning

Basically

No labels are given to the learning algorithm, leaving it on its own to find structure in its input.

Therefore, it is necessary to find the clusters of Data

- Using Cost Functions

$$SSE = \sum_{k=1}^{K} \sum_{x \in c_k} dist\left(\mathbf{x}, \mathbf{v}_k\right)^2$$

- Similarities

$$dist(\mathbf{x}, \mathbf{y}) = ||\mathbf{x} - \mathbf{y}||_A = \sqrt{(\mathbf{x} - \mathbf{y})^T A (\mathbf{x} - \mathbf{y})}$$

- etc

Unsupervised Learning

Basically

No labels are given to the learning algorithm, leaving it on its own to find structure in its input.

Therefore, it is necessary to find the clusters of Data

- Using Cost Functions

$$SSE = \sum_{k=1}^{K} \sum_{x \in c_k} dist\left(\mathbf{x}, \mathbf{v}_k\right)^2$$

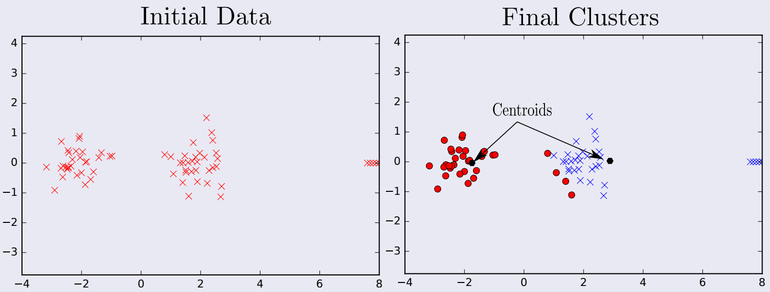
- Similarities

$$dist(\mathbf{x}, \mathbf{y}) = ||\mathbf{x} - \mathbf{y}||_A = \sqrt{(\mathbf{x} - \mathbf{y})^T A (\mathbf{x} - \mathbf{y})}$$

- etc

Example

K-Means



Outline

Other Main Areas are...

Semi-Supervised Learning

There is a training set with some (often many) of the target outputs missing.

- The Program receives rewards and punishments given its actions
- Those inputs allows the Program to learn by reinforcement

- Experience is gained by exploiting meta-knowledge extracted
  - Thus, Learning Bias must be chosen dynamically

Other Main Areas are...

Semi-Supervised Learning

There is a training set with some (often many) of the target outputs missing.

Reinforcement Learning

- The Program interacts with a dynamic environment to perform a certain goal.
- The Program receives rewards and punishments given its actions.
- Those inputs allows the Program to learn by reinforcement.

- Experience is gained by exploiting meta-knowledge extracted
  - Thus, Learning Bias must be chosen dynamically!

Other Main Areas are...

Semi-Supervised Learning

There is a training set with some (often many) of the target outputs missing.

Reinforcement Learning

- The Program interacts with a dynamic environment to perform a certain goal.
- The Program receives rewards and punishments given its actions.
- Those inputs allows the Program to learn by reinforcement.

Meta Learning

- The system must include a learning subsystem, which adapts with experience.
- Experience is gained by exploiting meta knowledge extracted.
  - Thus, Learning Bias must be chosen dynamically.

Other Main Areas are...

Semi-Supervised Learning

There is a training set with some (often many) of the target outputs missing.

Reinforcement Learning

- The Program interacts with a dynamic environment to perform a certain goal.
- The Program receives rewards and punishments given its actions.
- Those inputs allows the Program to learn by reinforcement.

Meta Learning

- The system must include a learning subsystem, which adapts with experience.
- Experience is gained by exploiting meta knowledge extracted.
  - Thus, Learning Bias must be chosen dynamically.

Other Main Areas are...

Semi-Supervised Learning

There is a training set with some (often many) of the target outputs missing.

Reinforcement Learning

- The Program interacts with a dynamic environment to perform a certain goal.
- The Program receives rewards and punishments given its actions.
- Those inputs allows the Program to learn by reinforcement.

Meta Learning

- The system must include a learning subsystem, which adapts with experience.
- Experience is gained by exploiting meta knowledge extracted.
  - Thus, Learning Bias must be chosen dynamically.

Other Main Areas are...

Semi-Supervised Learning

There is a training set with some (often many) of the target outputs missing.

Reinforcement Learning

- The Program interacts with a dynamic environment to perform a certain goal.
- The Program receives rewards and punishments given its actions.
- Those inputs allows the Program to learn by reinforcement.

Meta Learning

- The system must include a learning subsystem, which adapts with experience.
- Experience is gained by exploiting meta knowledge extracted.
  - Thus, Learning Bias must be chosen dynamically.

Other Main Areas are...

Semi-Supervised Learning

There is a training set with some (often many) of the target outputs missing.

Reinforcement Learning

- The Program interacts with a dynamic environment to perform a certain goal.
- The Program receives rewards and punishments given its actions.
- Those inputs allows the Program to learn by reinforcement.

Meta Learning

- The system must include a learning subsystem, which adapts with experience.
- Experience is gained by exploiting meta knowledge extracted.
  - Thus, Learning Bias must be chosen dynamically.

Outline

Machine Learning Process

Process

1. ***Pre-Processing***
2. *Feature Extraction/Feature Generation*
3. *Clustering $\approx$ Class Identification $\approx$ Unsupervised Learning*
4. *Classification $\approx$ Supervised Learning*

Machine Learning Process

Process

1. ***Pre-Processing***
2. ***Feature Extraction/Feature Generation***
3. *Clustering* $\approx$ *Class Identification* $\approx$ *Unsupervised Learning*
4. *Classification* $\approx$ *Supervised Learning*

Then

We need to process a lot of data...!!!

Machine Learning Process

Process

1. ***Pre-Processing***
2. ***Feature Extraction/Feature Generation***
3. ***Clustering*** $\approx$ ***Class Identification*** $\approx$ ***Unsupervised Learning***
4. ***Classification*** $\approx$ ***Supervised Learning***

Then

We need to process a lot of data...!!!

Machine Learning Process

Process

1. ***Pre-Processing***
2. ***Feature Extraction/Feature Generation***
3. ***Clustering*** $\approx$ ***Class Identification*** $\approx$ ***Unsupervised Learning***
4. ***Classification*** $\approx$ ***Supervised Learning***

Then

We need to process a lot of data...!!!

Machine Learning Process

Process

1. ***Pre-Processing***
2. ***Feature Extraction/Feature Generation***
3. ***Clustering*** $\approx$ ***Class Identification*** $\approx$ ***Unsupervised Learning***
4. ***Classification*** $\approx$ ***Supervised Learning***

Then

We need to process a lot of data...!!!

Outline

Feature Generation

Feature Generation

Given a set of measurements, the goal is to discover compact and informative representations of the obtained data.

Feature Generation

Feature Generation

Given a set of measurements, the goal is to discover compact and informative representations of the obtained data.

Examples

1. The Karhunen–Loève transform $\approx$ Principal Component Analysis
   1. Popular for feature generation and Dimensionality Reduction
2. The Singular Value Decomposition
   1. Used for Dimensionality Reduction

Feature Generation

Feature Generation

Given a set of measurements, the goal is to discover compact and informative representations of the obtained data.

Examples

1. The Karhunen–Loève transform $\approx$ Principal Component Analysis
   1. Popular for feature generation and Dimensionality Reduction
2. The Singular Value Decomposition
   1. Used for Dimensionality Reduction

Feature Generation

Feature Generation

Given a set of measurements, the goal is to discover compact and informative representations of the obtained data.

Examples

1. The Karhunen–Loève transform $\approx$ Principal Component Analysis
    1. Popular for feature generation and Dimensionality Reduction
2. The Singular Value Decomposition
    1. Used for Dimensionality Reduction

Feature Generation

Feature Generation

Given a set of measurements, the goal is to discover compact and informative representations of the obtained data.

Examples

1. The Karhunen–Loève transform $\approx$ Principal Component Analysis
   1. Popular for feature generation and Dimensionality Reduction
2. The Singular Value Decomposition
   1. Used for Dimensionality Reduction

Feature Extraction

Definition

Process to transform high-dimensional data into low-dimensional ones for improving accuracy, understanding, or removing noises.

Why?

Curse of dimensionality: Complexity grows exponentially in volume by adding extra dimensions.

Feature Extraction

Definition

Process to transform high-dimensional data into low-dimensional ones for improving accuracy, understanding, or removing noises.

Why?

Curse of dimensionality: Complexity grows exponentially in volume by adding extra dimensions.

Outline

Curse of Dimensionality

Question

Which **features** should be used for the classifier?

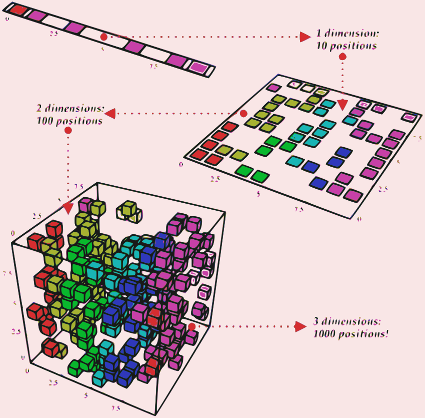
The Curse of Dimensionality!!!

Curse of Dimensionality

Question

Which **features** should be used for the classifier?

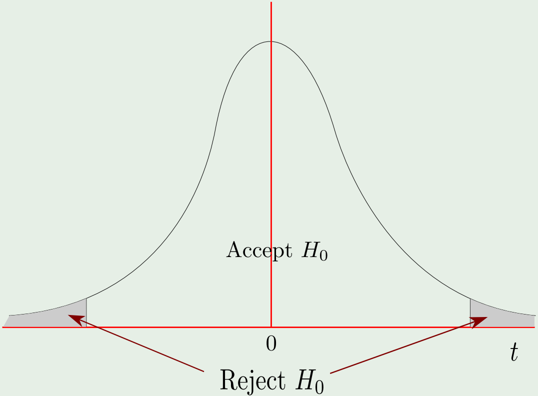
The Curse of Dimensionality!!!



What can be done?

Hypothesis Testing to discriminate good features

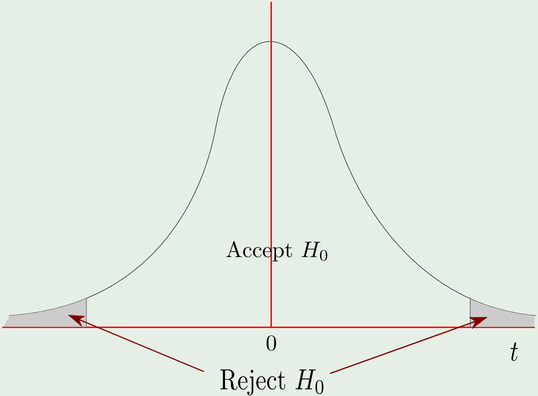
$$
\begin{aligned}
H_1 &: \Delta\mu = \mu_1 - \mu_2 \neq 0 \\
H_0 &: \Delta\mu = \mu_1 - \mu_2 = 0
\end{aligned}
$$



What can be done?

Hypothesis Testing to discriminate good features

$$
\begin{aligned}
H_1 &: \Delta\mu = \mu_1 - \mu_2 \neq 0 \\
H_0 &: \Delta\mu = \mu_1 - \mu_2 = 0
\end{aligned}
$$



What can be done?

Using Measures for Class Separability

- Between-class scatter matrix:

$$S_b = \sum_{i=1}^{M} P_i \left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_0\right)\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_0\right)^T \tag{1}$$

What can be done?

Using Measures for Class Separability

- Between-class scatter matrix:

$$S_b = \sum_{i=1}^{M} P_i \left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_0\right)\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_0\right)^T \tag{1}$$

Where

- $\boldsymbol{\mu}_0$ is the global mean vector, $\boldsymbol{\mu}_0 = \sum_{i=1}^{M} P_i \boldsymbol{\mu}_i$.
- $\boldsymbol{\mu}_i$ the median of class ω_i.
- $P_i \cong \frac{n_i}{N}$.

What can be done?

Using Measures for Class Separability

- Between-class scatter matrix:

$$S_b = \sum_{i=1}^{M} P_i \left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_0\right)\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_0\right)^T \tag{1}$$

Where

- $\boldsymbol{\mu}_0$ is the global mean vector, $\boldsymbol{\mu}_0 = \sum_{i=1}^{M} P_i \boldsymbol{\mu}_i$.
- $\boldsymbol{\mu}_i$ the median of class ω_i.
- $P_i \cong \frac{n_i}{N}$.

What can be done?

Using Measures for Class Separability

- Between-class scatter matrix:

$$S_b = \sum_{i=1}^{M} P_i \left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_0\right)\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_0\right)^T \quad (1)$$

Where

- $\boldsymbol{\mu}_0$ is the global mean vector, $\boldsymbol{\mu}_0 = \sum_{i=1}^{M} P_i \boldsymbol{\mu}_i$.
- $\boldsymbol{\mu}_i$ the median of class ω_i.
- $P_i \cong \frac{n_i}{N}$.

What can be done?

Feature Subset Selection

- Examples:
  - Filter Approach
    - All combinations of features are used together with a separability measure.
  - Wrapper Approach:
    - Use the decided classifier itself to find the best set.

What can be done?

Feature Subset Selection

- Examples:
  - Filter Approach
    - All combinations of features are used together with a separability measure.
  - Wrapper Approach:
    - Use the decided classifier itself to find the best set.

What can be done?

Feature Subset Selection

- Examples:
  - Filter Approach
    - ★ All combinations of features are used together with a separability measure.
  - Wrapper Approach:
    - ★ Use the decided classifier itself to find the best set.

What can be done?

Feature Subset Selection

- Examples:
  - Filter Approach
    - ★ All combinations of features are used together with a separability measure.
  - Wrapper Approach:
    - ★ Use the decided classifier itself to find the best set.

What can be done?

Feature Subset Selection

- Examples:
  - Filter Approach
    - All combinations of features are used together with a separability measure.
  - Wrapper Approach:
    - Use the decided classifier itself to find the best set.

Outline

Clustering

Definition

Grouping unlabeled data into clusters, for the purpose of inference of hidden structures or information.

Clustering

Definition

Grouping unlabeled data into clusters, for the purpose of inference of hidden structures or information.

Using, for example

Dissimilarity measures

- Angle : Inner product, ...
- Non-metric : Rank, Intensity, ...
- Distance : Euclidean (l_2), Manhattan(l_1), ...

Clustering

Definition

Grouping unlabeled data into clusters, for the purpose of inference of hidden structures or information.

Using, for example

Dissimilarity measures

- Angle : Inner product, ...
- Non-metric : Rank, Intensity, ...
- Distance : Euclidean (l_2), Manhattan(l_1), ...

Clustering

Definition

Grouping unlabeled data into clusters, for the purpose of inference of hidden structures or information.

Using, for example

Dissimilarity measures

- Angle : Inner product, . . .
- Non-metric : Rank, Intensity, . . .
- Distance : Euclidean (l_2), Manhattan(l_1), . . .

Clustering

Definition

Grouping unlabeled data into clusters, for the purpose of inference of hidden structures or information.

Using, for example

Dissimilarity measures

- Angle : Inner product, ...
- Non-metric : Rank, Intensity, ...
- Distance : Euclidean (l_2), Manhattan(l_1), ...

Examples of Clustering Algorithms

Clustering

1. Basic Clustering Algorithms
   1. K-means
2. Clustering Based in Cost Functions
   1. Fuzzy C-means
   2. Possibilistic
3. Hierarchical Clustering
   1. Entropy based
4. Clustering Based in Graph Theory

Examples of Clustering Algorithms

Clustering

1. Basic Clustering Algorithms
   1. K-means
2. Clustering Based in Cost Functions
   1. Fuzzy C-means
   2. Possibilistic
3. Hierarchical Clustering
   1. Entropy based
4. Clustering Based in Graph Theory

Examples of Clustering Algorithms

Clustering

1. Basic Clustering Algorithms
   1. K-means
2. Clustering Based in Cost Functions
   1. Fuzzy C-means
   2. Possibilistic
3. Hierarchical Clustering
   1. Entropy based
4. Clustering Based in Graph Theory

Examples of Clustering Algorithms

Clustering

1. Basic Clustering Algorithms
   1. K-means
2. Clustering Based in Cost Functions
   1. Fuzzy C-means
   2. Possibilistic
3. Hierarchical Clustering
   1. Entropy based
4. Clustering Based in Graph Theory

Examples of Clustering Algorithms

Clustering

1. Basic Clustering Algorithms
   1. K-means
2. Clustering Based in Cost Functions
   1. Fuzzy C-means
   2. Possibilistic
3. Hierarchical Clustering
   1. Entropy based
4. Clustering Based in Graph Theory

Examples of Clustering Algorithms

Clustering

1. Basic Clustering Algorithms
   1. K-means
2. Clustering Based in Cost Functions
   1. Fuzzy C-means
   2. Possibilistic
3. Hierarchical Clustering
   1. Entropy based
4. Clustering Based in Graph Theory

Examples of Clustering Algorithms

Clustering

1. Basic Clustering Algorithms
   1. K-means
2. Clustering Based in Cost Functions
   1. Fuzzy C-means
   2. Possibilistic
3. Hierarchical Clustering
   1. Entropy based
4. Clustering Based in Graph Theory

Examples of Clustering Algorithms

Clustering

1. Basic Clustering Algorithms
   1. K-means
2. Clustering Based in Cost Functions
   1. Fuzzy C-means
   2. Possibilistic
3. Hierarchical Clustering
   1. Entropy based
4. Clustering Based in Graph Theory

Outline

Classification

Definition

A procedure dividing data into the given set of categories based on the training set in a supervised way.

Classification

Definition

A procedure dividing data into the given set of categories based on the training set in a supervised way.

What do we want from classification?

1. To Learn the pattern that relates $f(\boldsymbol{x}) \Longleftrightarrow y$ from the training set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$.
2. To generalize new samples i.e. given a new sample $\boldsymbol{x}'$, $f(\boldsymbol{x}')$ gets the correct label.

Outline

The previous Controlled Over-fitting

We have a problem

Bias–Variance Trade-Off

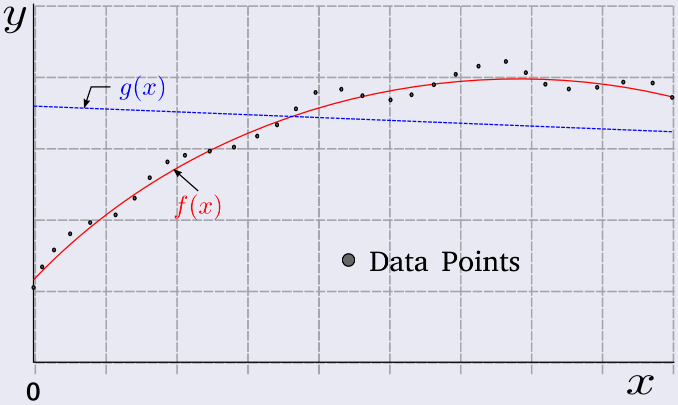
Intuition - Bias

The previous Controlled Over-fitting

We have a problem

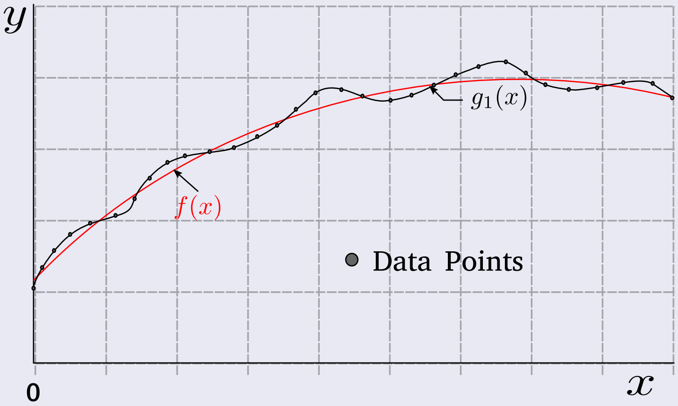
Bias–Variance Trade-Off

Intuition - Bias



A the Other Hand

The Variance



Possible Solution

Validation Error and Training Error

- Two Data Sets are used:
  - One for validation
  - The other for training

Possible Solution

Validation Error and Training Error

- Two Data Sets are used:
  - One for validation
  - The other for training

Possible Solution

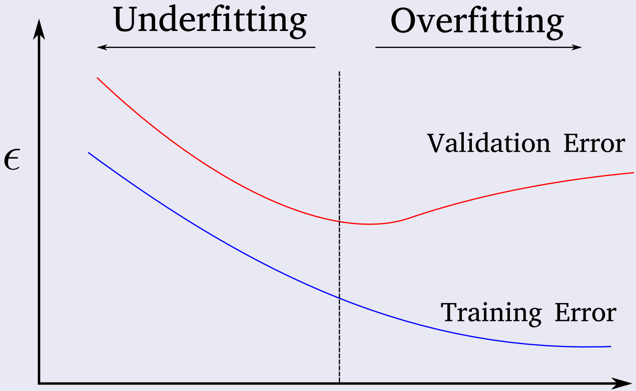
Validation Error and Training Error

- Two Data Sets are used:
  - One for validation
  - The other for training

Possible Solution

Validation Error and Training Error

- Two Data Sets are used:
  - One for validation
  - The other for training



Outline

Examples of Classification Algorithms

Many Possible Algorithms

- Linear Classifiers: Perceptron
- Probability Classifiers: Naive Bayes
- Kernel Methods Classifiers : Support Vector Machines
- Non-Linear Classifiers: Artificial Neural Networks
- Graph Model Classifiers:

Examples of Classification Algorithms

Many Possible Algorithms

- Linear Classifiers: Perceptron
- Probability Classifiers: Naive Bayes
- Kernel Methods Classifiers : Support Vector Machines
- Non-Linear Classifiers: Artificial Neural Networks
- Graph Model Classifiers:

Examples of Classification Algorithms

Many Possible Algorithms

- Linear Classifiers: Perceptron
- Probability Classifiers: Naive Bayes
- Kernel Methods Classifiers : Support Vector Machines
- Non-Linear Classifiers: Artificial Neural Networks
- Graph Model Classifiers:

Examples of Classification Algorithms

Many Possible Algorithms

- Linear Classifiers: Perceptron
- Probability Classifiers: Naive Bayes
- Kernel Methods Classifiers : Support Vector Machines
- Non-Linear Classifiers: Artificial Neural Networks
- Graph Model Classifiers:

Examples of Classification Algorithms

Many Possible Algorithms

- Linear Classifiers: Perceptron
- Probability Classifiers: Naive Bayes
- Kernel Methods Classifiers : Support Vector Machines
- Non-Linear Classifiers: Artificial Neural Networks
- Graph Model Classifiers:

Examples of Classification Algorithms

Many Possible Algorithms

- Linear Classifiers: Perceptron
- Probability Classifiers: Naive Bayes
- Kernel Methods Classifiers : Support Vector Machines
- Non-Linear Classifiers: Artificial Neural Networks
- Graph Model Classifiers: