Natural Language Processing
Lecture 02 Machine Learning Basics; Text Classification

Qun Liu, Valentin Malykh
Huawei Noah's Ark Lab

Spring 2020
A course delivered at MIPT, Moscow

Content

Content

Content

What is machine learning?

— Wikipedia definition

- *Machine learning (ML)* is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead.
- It is seen as a subset of artificial intelligence.
- Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task.

- Machine learning algorithms are used in a wide variety of applications, such as email filtering and computer vision, where it is difficult or infeasible to develop a conventional algorithm for effectively performing the task.
- Machine learning is closely related to *computational statistics*, which focuses on making predictions using computers.
- The study of *mathematical optimization* delivers methods, theory and application domains to the field of machine learning.
- *Data mining* is a field of study within machine learning, and focuses on exploratory data analysis through unsupervised learning.
- In its application across business problems, machine learning is also referred to as *predictive analytics*.

Supervised machine learning

- (Supervised) Machine Learning techniques automatically learn a model of the relationship between a set of **descriptive features** and a **target feature** from a set of historical examples.

John Kelleher and Brian Mac Namee and Aoife D'Arcy, Fundamentals of Machine Learning for Predictive Data Analytics

Supervised machine learning

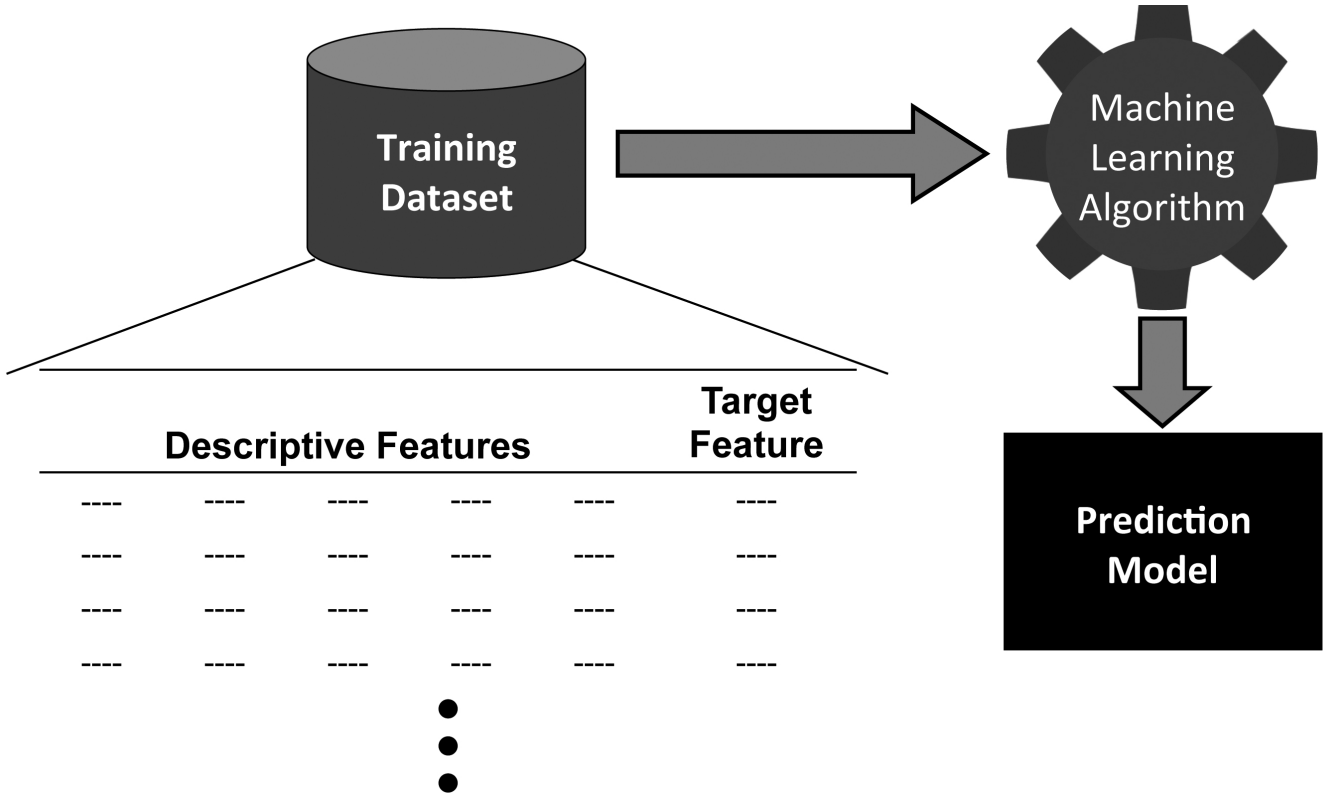


Figure: Using machine learning to induce a prediction model from a training dataset.

John Kelleher and Brian Mac Namee and Aoife D'Arcy, Fundamentals of Machine Learning for Predictive Data Analytics

Supervised machine learning

Figure: Using the model to make predictions for new query instances.

John Kelleher and Brian Mac Namee and Aoife D'Arcy, Fundamentals of Machine Learning for Predictive Data Analytics

ID	Occupation	Age	Loan-Salary Ratio	Outcome
1	industrial	34	2.96	repaid
2	professional	41	4.64	default
3	professional	36	3.22	default
4	professional	41	3.11	default
5	industrial	48	3.80	default
6	industrial	61	2.52	repaid
7	professional	37	1.50	repaid
8	professional	40	1.93	repaid
9	industrial	33	5.25	default
10	industrial	32	4.15	default

Input
Input Features
Descriptive Features
Query Instance

Output
Output Features
Target Features
Prediction

John Kelleher and Brian Mac Namee and Aoife D'Arcy, Fundamentals of Machine Learning for Predictive Data Analytics

Content

ID	Occupation	Age	Loan-Salary Ratio	Outcome
1	industrial	34	2.96	repaid
2	professional	41	4.64	default
3	professional	36	3.22	default
4	professional	41	3.11	default
5	industrial	48	3.80	default
6	industrial	61	2.52	repaid
7	professional	37	1.50	repaid
8	professional	40	1.93	repaid
9	industrial	33	5.25	default
10	industrial	32	4.15	default

- What is the relationship between the **descriptive features** (Occupation, Age, Loan-Salary Ratio) and the **target**

John Kelleher and Brian Mac Namee and Aoife D'Arcy, Fundamentals of Machine Learning for Predictive Data Analytics

Machine learning – an example

if LOAN-SALARY RATIO $>$ 3 **then**
  OUTCOME=*'default'*
else
  OUTCOME=*'repay'*
end if

John Kelleher and Brian Mac Namee and Aoife D'Arcy, Fundamentals of Machine Learning for Predictive Data Analytics

Machine learning – an example

if LOAN-SALARY RATIO $>$ 3 **then**
  OUTCOME=*'default'*
else
  OUTCOME=*'repay'*
end if

- This is an example of a **prediction model**

John Kelleher and Brian Mac Namee and Aoife D'Arcy, Fundamentals of Machine Learning for Predictive Data Analytics

Machine learning – an example

if LOAN-SALARY RATIO $>$ 3 **then**
  OUTCOME=*'default'*
else
  OUTCOME=*'repay'*
end if

- This is an example of a **prediction model**
- This is also an example of a **consistent** prediction model

John Kelleher and Brian Mac Namee and Aoife D'Arcy, Fundamentals of Machine Learning for Predictive Data Analytics

Machine learning – an example

if LOAN-SALARY RATIO $>$ 3 **then**
  OUTCOME=*'default'*
else
  OUTCOME=*'repay'*
end if

- This is an example of a **prediction model**
- This is also an example of a **consistent** prediction model
- Notice that this model does not use all the features and the feature that it uses is a derived feature (in this case a ratio): **feature design** and **feature selection** are two

John Kelleher and Brian Mac Namee and Aoife D'Arcy, Fundamentals of Machine Learning for Predictive Data Analytics

Machine learning – an example

- What is the relationship between the **descriptive features** and the **target feature** (OUTCOME) in the following dataset?

John Kelleher and Brian Mac Namee and Aoife D'Arcy, Fundamentals of Machine Learning for Predictive Data Analytics

ID	Amount	Salary	Loan-Salary Ratio	Age	Occupation	House	Type	Outcome
1	245,100	66,400	3.69	44	industrial	farm	stb	repaid
2	90,600	75,300	1.2	41	industrial	farm	stb	repaid
3	195,600	52,100	3.75	37	industrial	farm	ftb	default
4	157,800	67,600	2.33	44	industrial	apartment	ftb	repaid
5	150,800	35,800	4.21	39	professional	apartment	stb	default
6	133,000	45,300	2.94	29	industrial	farm	ftb	default
7	193,100	73,200	2.64	38	professional	house	ftb	repaid
8	215,000	77,600	2.77	17	professional	farm	ftb	repaid
9	83,000	62,500	1.33	30	professional	house	ftb	repaid
10	186,100	49,200	3.78	30	industrial	house	ftb	default
11	161,500	53,300	3.03	28	professional	apartment	stb	repaid
12	157,400	63,900	2.46	30	professional	farm	stb	repaid
13	210,000	54,200	3.87	43	professional	apartment	ftb	repaid
14	209,700	53,000	3.96	39	industrial	farm	ftb	default
15	143,200	65,300	2.19	32	industrial	apartment	ftb	default
16	203,000	64,400	3.15	44	industrial	farm	ftb	repaid
17	247,800	63,800	3.88	46	industrial	house	stb	repaid
18	162,700	77,400	2.1	37	professional	house	ftb	repaid
19	213,300	61,100	3.49	21	industrial	apartment	ftb	default
20	284,100	32,300	8.8	51	industrial	farm	ftb	default
21	154,000	48,900	3.15	49	professional	house	stb	repaid
22	112,800	79,700	1.42	41	professional	house	ftb	repaid
23	252,000	59,700	4.22	27	professional	house	stb	default
24	175,200	39,900	4.39	37	professional	apartment	stb	default
25	149,700	58,600	2.55	35	industrial	farm	stb	default

John Kelleher and Brian Mac Namee and Aoife D'Arcy, Fundamentals of Machine Learning for Predictive Data Analytics

Machine learning – an example

if LOAN-SALARY RATIO $<$ 1.5 **then**
  OUTCOME=*'repay'*
else if LOAN-SALARY RATIO $>$ 4 **then**
  OUTCOME=*'default'*
else if AGE $<$ 40 **and** OCCUPATION $=$*'industrial'* **then**
  OUTCOME=*'default'*
else
  OUTCOME=*'repay'*
end if

John Kelleher and Brian Mac Namee and Aoife D'Arcy, Fundamentals of Machine Learning for Predictive Data Analytics

Machine learning – an example

if LOAN-SALARY RATIO $<$ 1.5 **then**
  OUTCOME=*'repay'*
else if LOAN-SALARY RATIO $>$ 4 **then**
  OUTCOME=*'default'*
else if AGE $<$ 40 **and** OCCUPATION $=$*'industrial'* **then**
  OUTCOME=*'default'*
else
  OUTCOME=*'repay'*
end if

- The real value of machine learning becomes apparent in situations like this when we want to build prediction models

John Kelleher and Brian Mac Namee and Aoife D'Arcy, Fundamentals of Machine Learning for Predictive Data Analytics

Content

Model spaces and inductive bias

- Machine learning algorithms work by searching through a set of possible prediction models for the model that best captures the relationship between the descriptive features and the target feature.

John Kelleher and Brian Mac Namee and Aoife D'Arcy, Fundamentals of Machine Learning for Predictive Data Analytics

Model spaces and inductive bias

- Machine learning algorithms work by searching through a set of possible prediction models for the model that best captures the relationship between the descriptive features and the target feature.
- An obvious search criteria to drive this search is to look for models that are consistent with the data.

John Kelleher and Brian Mac Namee and Aoife D'Arcy, Fundamentals of Machine Learning for Predictive Data Analytics

Model spaces and inductive bias

- Machine learning algorithms work by searching through a set of possible prediction models for the model that best captures the relationship between the descriptive features and the target feature.
- An obvious search criteria to drive this search is to look for models that are consistent with the data.
- However, because a training dataset is only a sample ML is an ill-posed problem.

John Kelleher and Brian Mac Namee and Aoife D'Arcy, Fundamentals of Machine Learning for Predictive Data Analytics

Model spaces and inductive bias

Table: A simple retail dataset

ID	BBY	ALC	ORG	GRP
1	no	no	no	couple
2	yes	no	yes	family
3	yes	yes	no	family
4	no	no	yes	couple
5	no	yes	yes	single

John Kelleher and Brian Mac Namee and Aoife D'Arcy, Fundamentals of Machine Learning for Predictive Data Analytics

Model spaces and inductive bias

Table: A full set of potential prediction models before any training data becomes available.

BBY	ALC	ORG	GRP	$\mathbb{M}_1$	$\mathbb{M}_2$	$\mathbb{M}_3$	$\mathbb{M}_4$	$\mathbb{M}_5$	...	$\mathbb{M}_{6\,561}$
no	no	no	?	couple	couple	single	couple	couple		couple
no	no	yes	?	single	couple	single	couple	couple		single
no	yes	no	?	family	family	single	single	single		family
no	yes	yes	?	single	single	single	single	single		couple
yes	no	no	?	couple	couple	family	family	family	...	family
yes	no	yes	?	couple	family	family	family	family		couple
yes	yes	no	?	single	family	family	family	family		single
yes	yes	yes	?	single	single	family	family	couple		family

John Kelleher and Brian Mac Namee and Aoife D'Arcy, Fundamentals of Machine Learning for Predictive Data Analytics

Model spaces and inductive bias

Table: A sample of the models that are consistent with the training data

BBY	ALC	ORG	GRP	$\mathbb{M}_1$	$\mathbb{M}_2$	$\mathbb{M}_3$	$\mathbb{M}_4$	$\mathbb{M}_5$	...	$\mathbb{M}_{6\,561}$
no	no	no	couple	couple	couple	single	couple	couple		couple
no	no	yes	couple	single	couple	single	couple	couple		single
no	yes	no	?	family	family	single	single	single		family
no	yes	yes	single	single	single	single	single	single		couple
yes	no	no	?	couple	couple	family	family	family	...	family
yes	no	yes	family	couple	family	family	family	family		couple
yes	yes	no	family	single	family	family	family	family		single
yes	yes	yes	?	single	single	family	family	couple		family

John Kelleher and Brian Mac Namee and Aoife D'Arcy, Fundamentals of Machine Learning for Predictive Data Analytics

Model spaces and inductive bias

Table: A sample of the models that are consistent with the training data

BBY	ALC	ORG	GRP	$\mathbb{M}_1$	$\mathbb{M}_2$	$\mathbb{M}_3$	$\mathbb{M}_4$	$\mathbb{M}_5$	...	$\mathbb{M}_{6\,561}$
no	no	no	couple	couple	couple	single	couple	couple		couple
no	no	yes	couple	single	couple	single	couple	couple		single
no	yes	no	?	family	family	single	single	single		family
no	yes	yes	single	single	single	single	single	single		couple
yes	no	no	?	couple	couple	family	family	family	...	family
yes	no	yes	family	couple	family	family	family	family		couple
yes	yes	no	family	single	family	family	family	family		single
yes	yes	yes	?	single	single	family	family	couple		family

- Notice that there is more than one candidate model left! It is because a single consistent model cannot be found

John Kelleher and Brian Mac Namee and Aoife D'Arcy, Fundamentals of Machine Learning for Predictive Data Analytics

Model spaces and inductive bias

- Consistency $\approx$ memorizing the dataset.
- Consistency with noise in the data isn't desirable.
- Goal: a model that **generalises** beyond the dataset and that isn't influenced by the noise in the dataset.
- So what criteria should we use for choosing between models?

John Kelleher and Brian Mac Namee and Aoife D'Arcy, Fundamentals of Machine Learning for Predictive Data Analytics

Model spaces and inductive bias

- Inductive bias the set of assumptions that define the model selection criteria of an ML algorithm.
- There are two types of bias that we can use:
  1. restriction bias
  2. preference bias
- Inductive bias is necessary for learning (beyond the dataset).

John Kelleher and Brian Mac Namee and Aoife D'Arcy, Fundamentals of Machine Learning for Predictive Data Analytics

Content

Classification

Table: A simple retail dataset

ID	BBY	ALC	ORG	GRP
1	no	no	no	couple
2	yes	no	yes	family
3	yes	yes	no	family
4	no	no	yes	couple
5	no	yes	yes	single

To predict a target feature with categorical values.

John Kelleher and Brian Mac Namee and Aoife D'Arcy, Fundamentals of Machine Learning for Predictive Data Analytics

Regression

Table: The age-income dataset.

ID	AGE	INCOME
1	21	24,000
2	32	48,000
3	62	83,000
4	72	61,000
5	84	52,000

To predict a target feature with numerical values.

John Kelleher and Brian Mac Namee and Aoife D'Arcy, Fundamentals of Machine Learning for Predictive Data Analytics

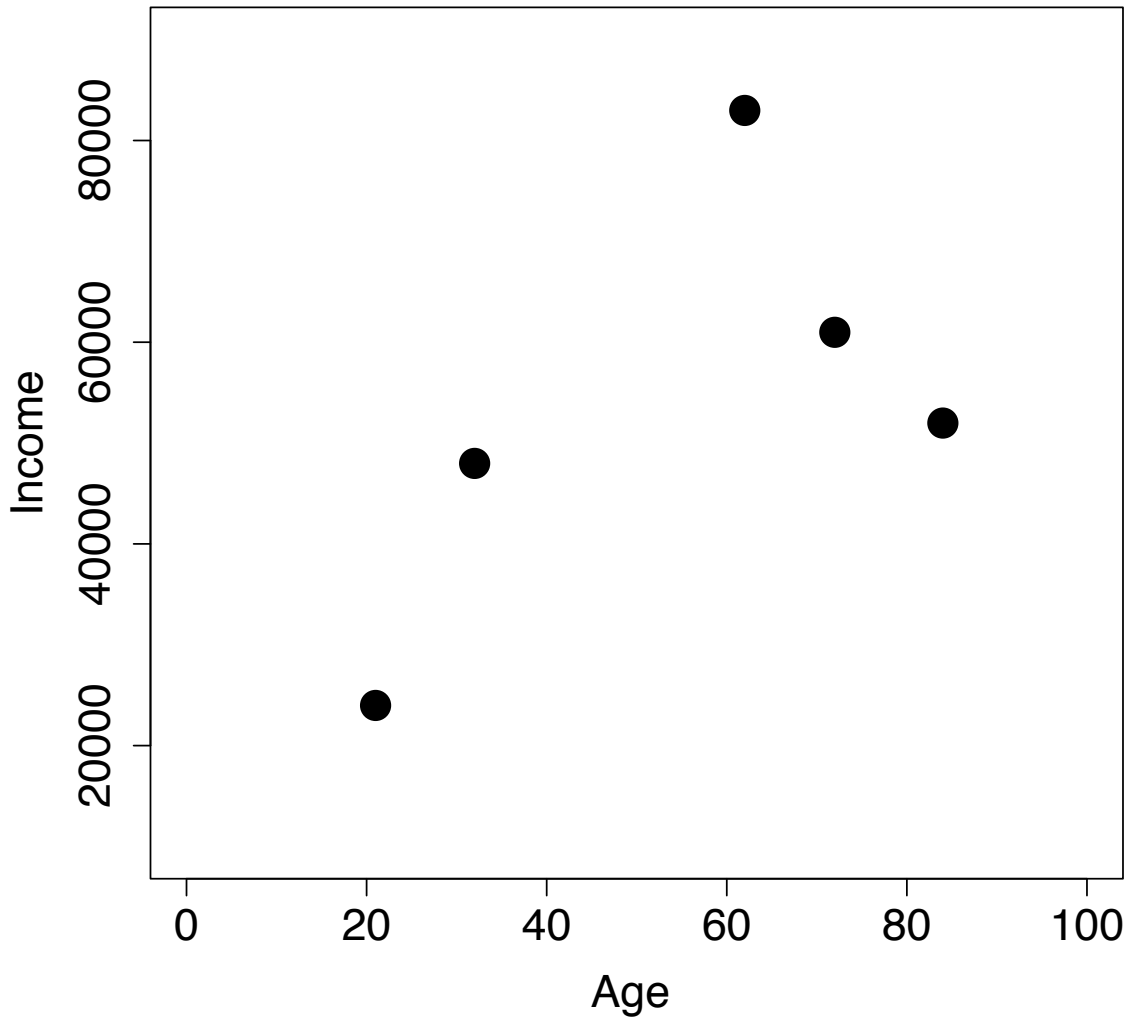
Content

Overfitting and underfitting

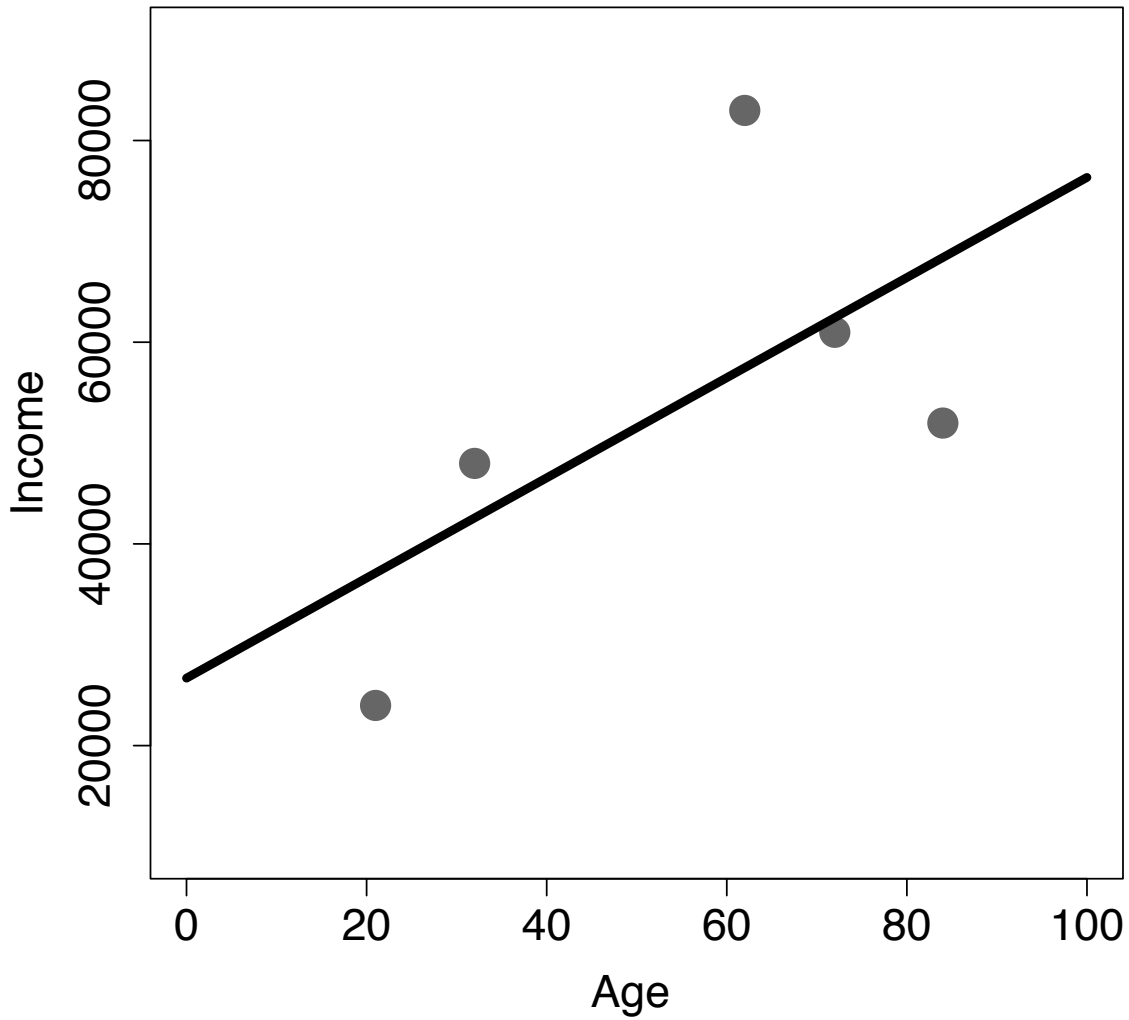
Table: The age-income dataset.

ID	AGE	INCOME
1	21	24,000
2	32	48,000
3	62	83,000
4	72	61,000
5	84	52,000

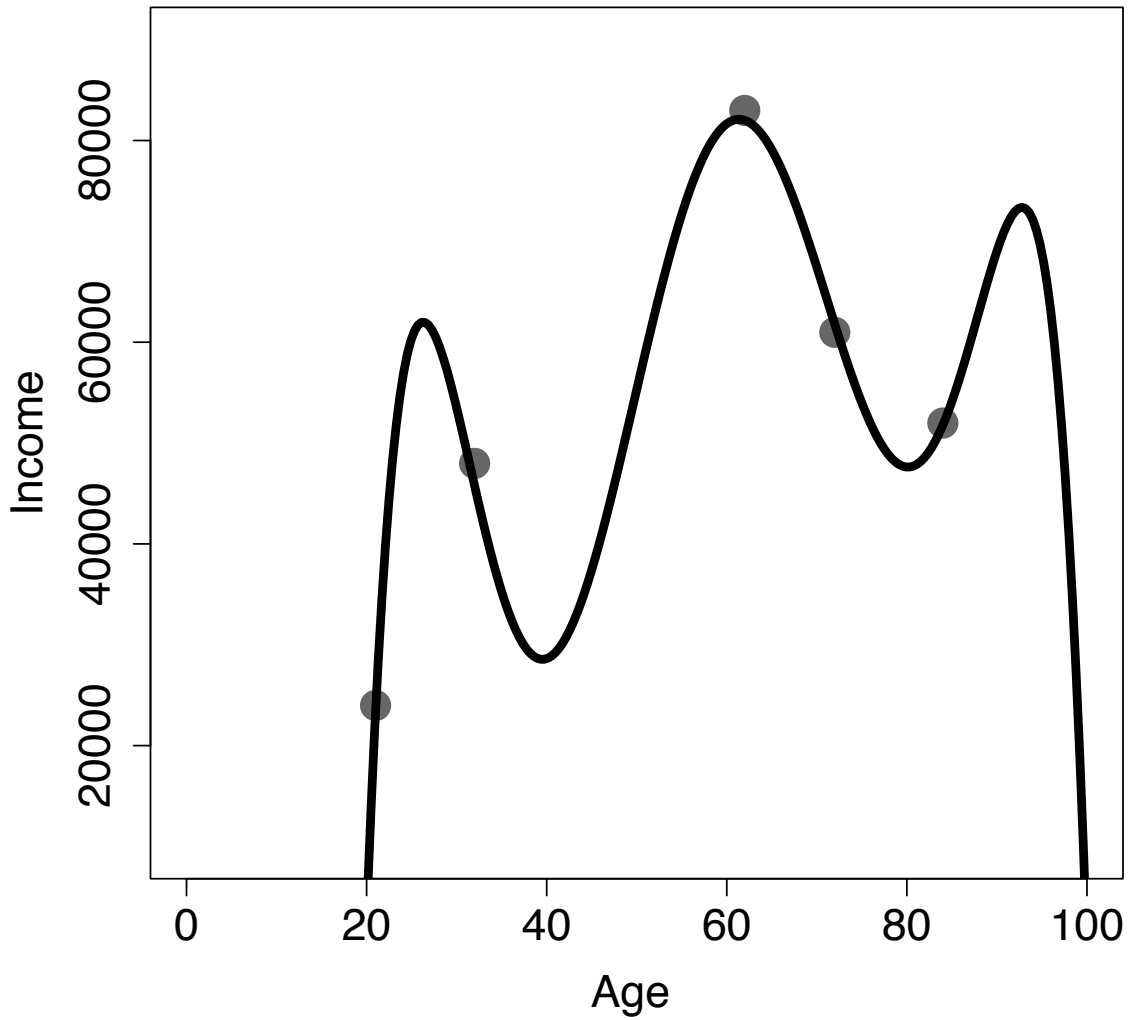
John Kelleher and Brian Mac Namee and Aoife D'Arcy, Fundamentals of Machine Learning for Predictive Data Analytics



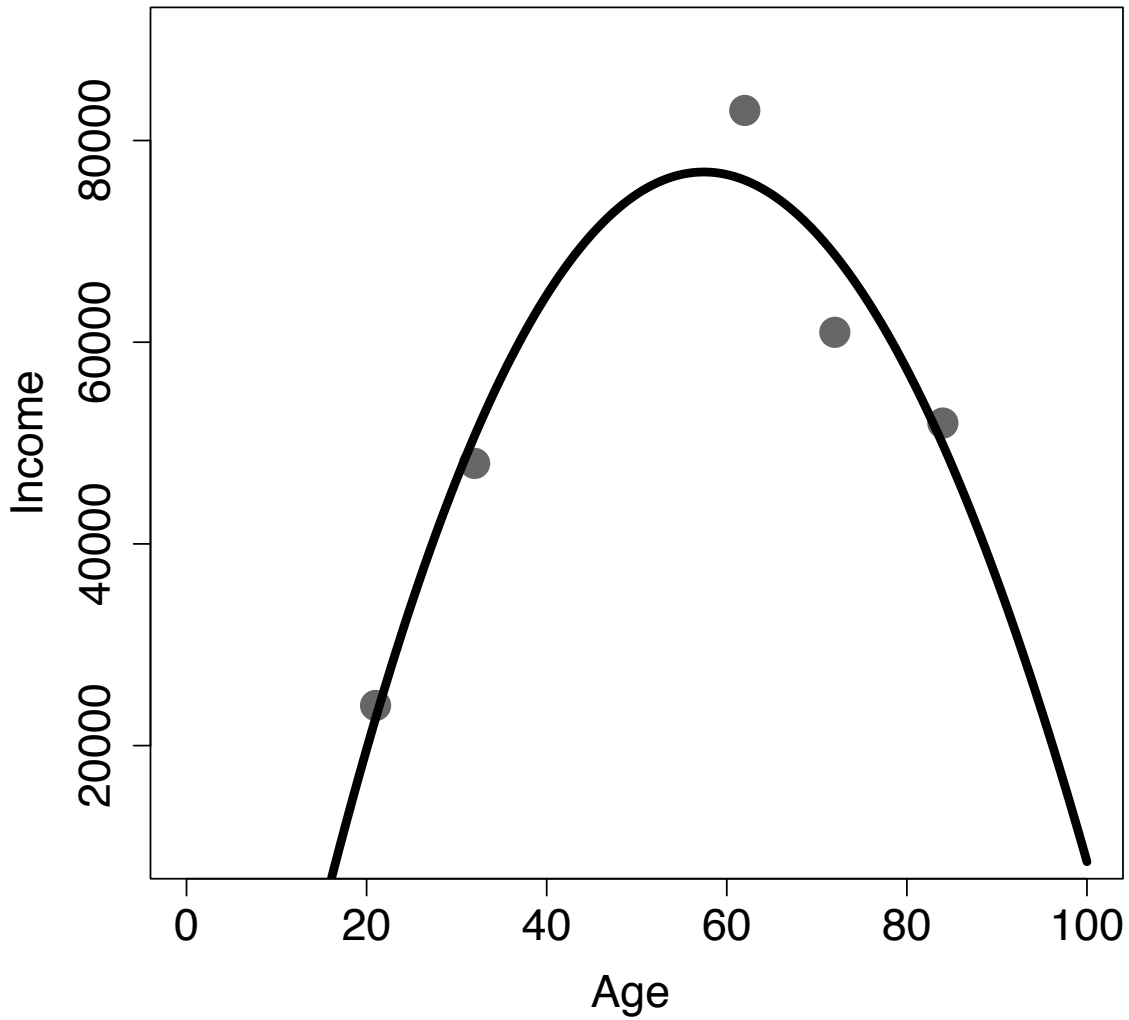
John Kelleher and Brian Mac Namee and Aoife D'Arcy, Fundamentals of Machine Learning for Predictive Data Analytics



John Kelleher and Brian Mac Namee and Aoife D'Arcy, Fundamentals of Machine Learning for Predictive Data Analytics

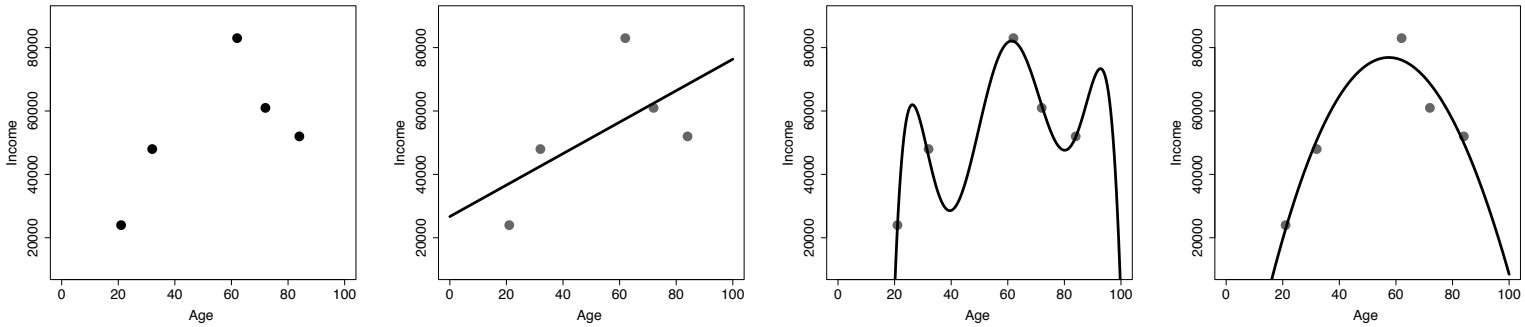


John Kelleher and Brian Mac Namee and Aoife D'Arcy, Fundamentals of Machine Learning for Predictive Data Analytics



John Kelleher and Brian Mac Namee and Aoife D'Arcy, Fundamentals of Machine Learning for Predictive Data Analytics

Overfitting and underfitting



(a) Dataset (b) Underfitting (c) Overfitting (d) Just right

Figure: Striking a balance between overfitting and underfitting when trying to predict age from income.

John Kelleher and Brian Mac Namee and Aoife D'Arcy, Fundamentals of Machine Learning for Predictive Data Analytics

Content

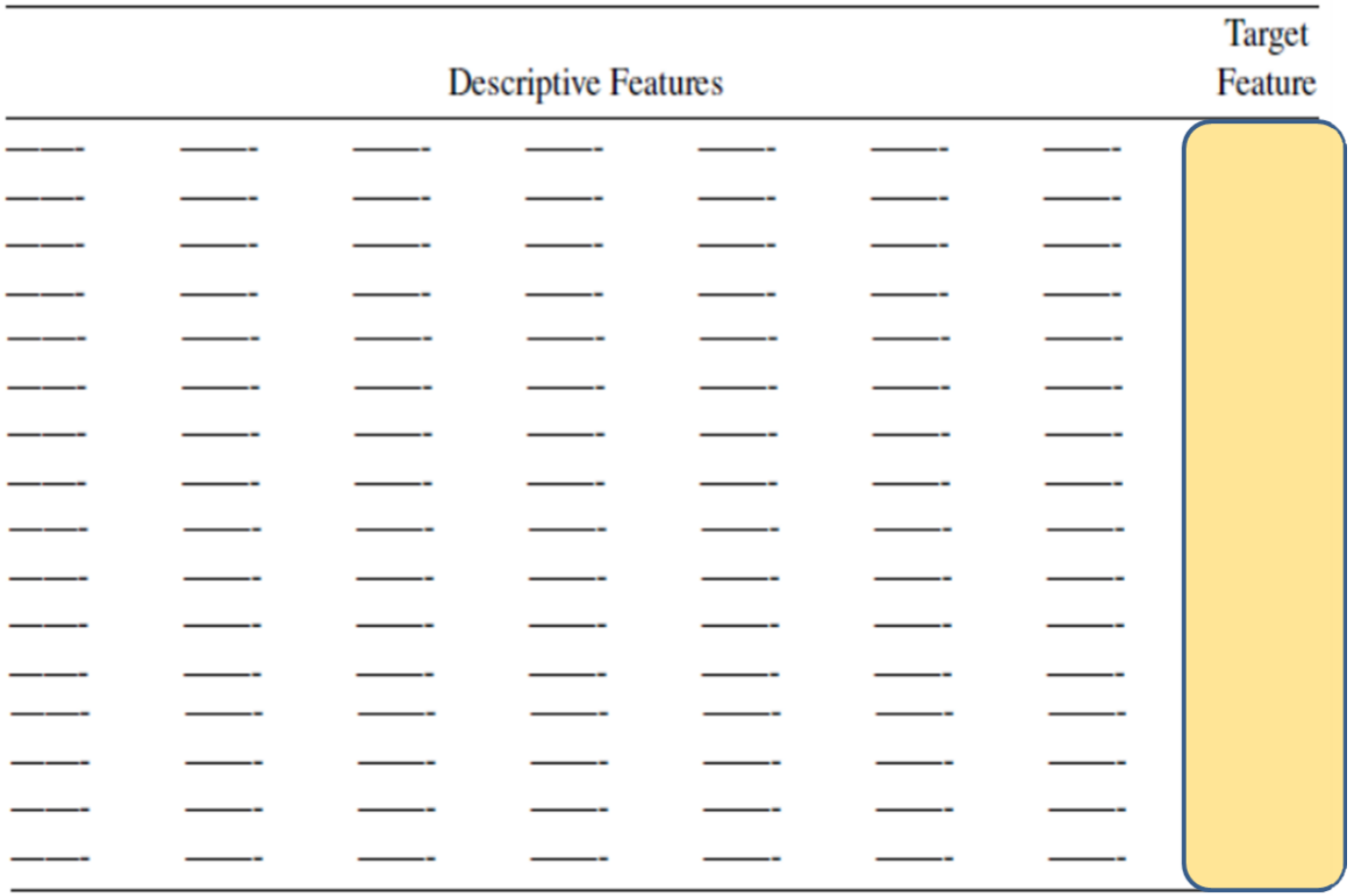
Unsupervised learning

- *Unsupervised learning* is the *machine learning* task of inferring a function to describe hidden structure from unlabeled data.
- Since the examples given to the learner are unlabeled, there is no error or reward signal to evaluate a potential solution.
- This distinguishes unsupervised learning from supervised learning.

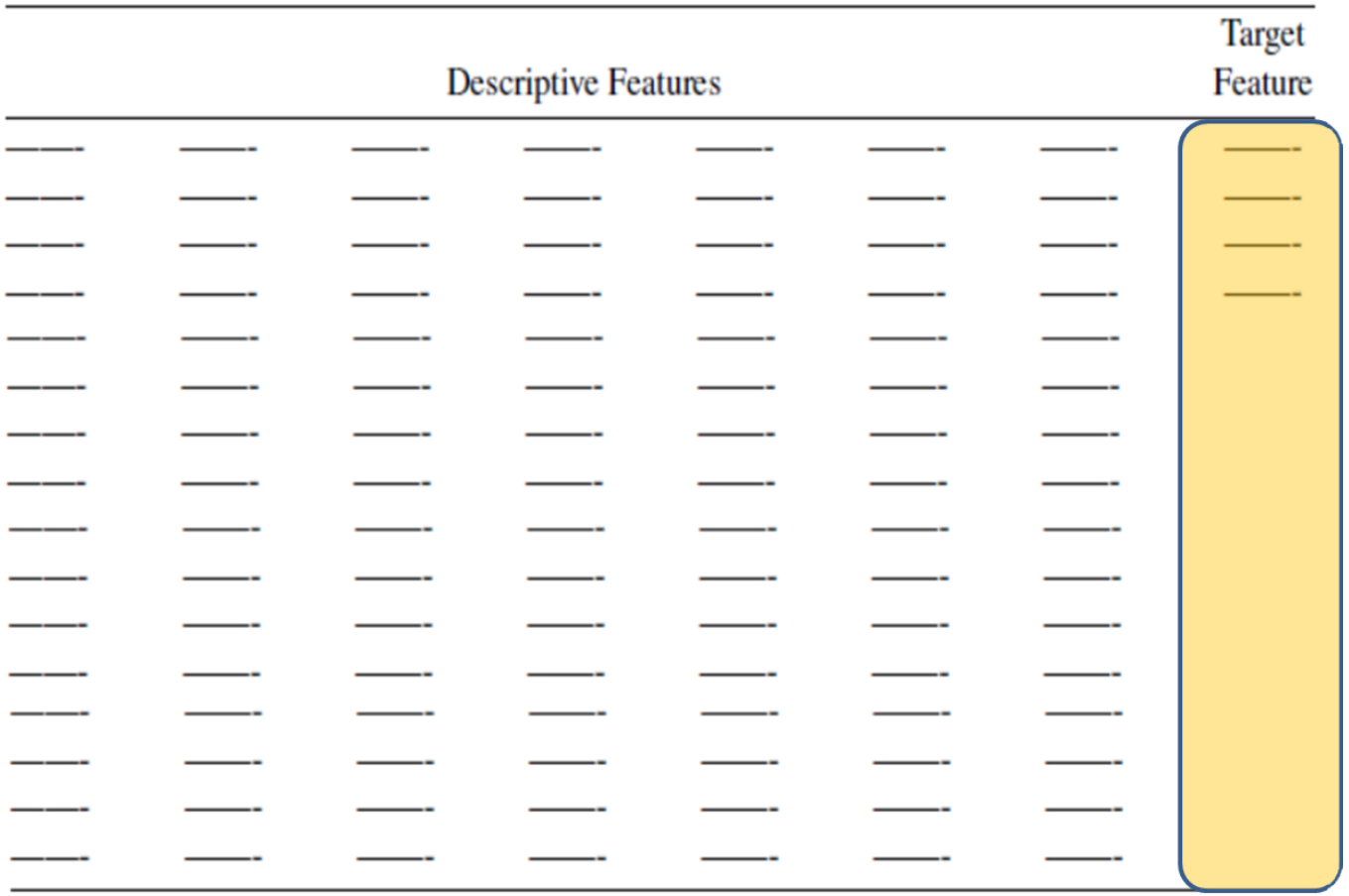
Supervised learning

Descriptive Features							Target Feature

Unsupervised learning



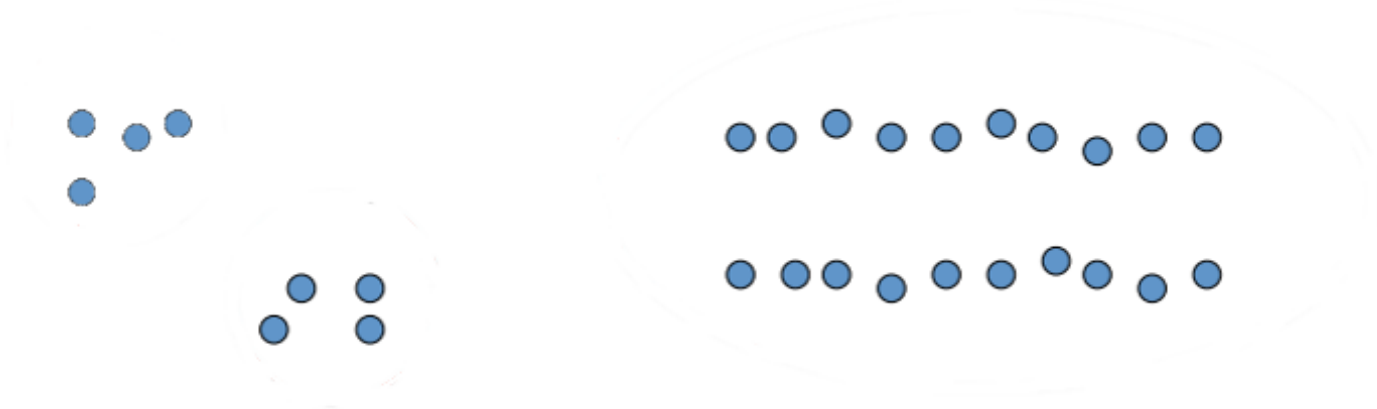
Semi-supervised learning



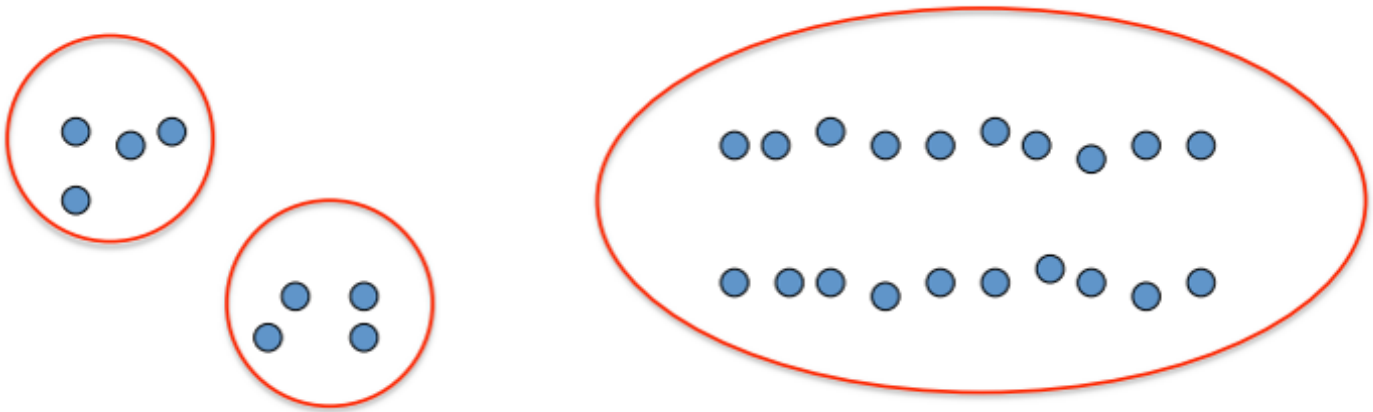
Clustering

- *Cluster analysis* or *clustering* is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters).
- Clustering is a typical unsupervised learning task.

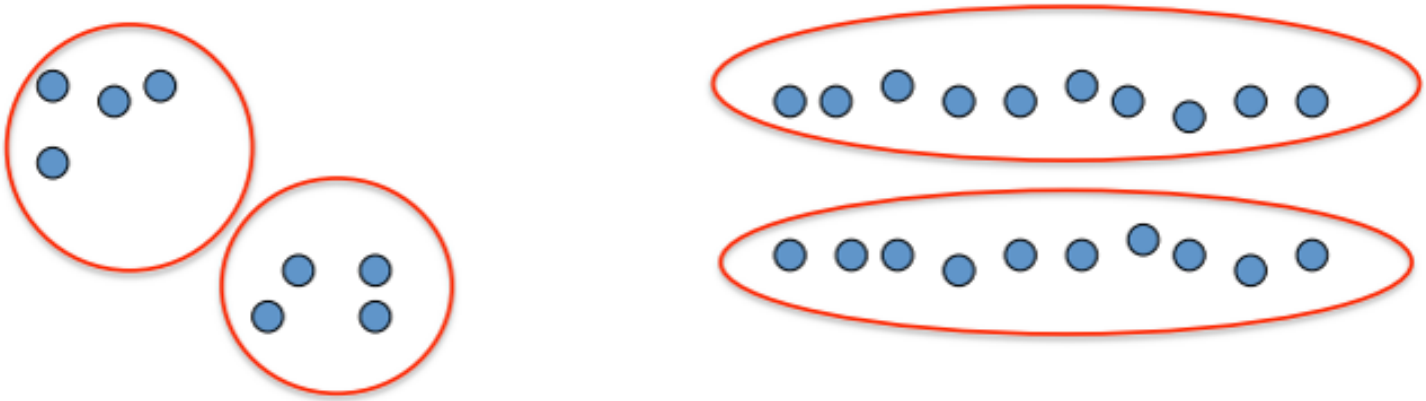
Clustering – An example



Clustering – An example



Clustering – An example



Clustering – An example