

Machine Learning

Prepared by Dr. Yasser Abdelhamid

DS 620
Machine Learning

Lecture 2





OUTLINE

- ◉ Machine learning preliminaries.
 - ◉ Knowledge
 - ◉ Inference
 - ◉ Inductive inference
 - ◉ Deductive inference
 - ◉ Domain
 - ◉ Instance
 - ◉ Attributes
 - ◉ Dataset
 - ◉ Training set
 - ◉ Model
- ◉ Example

KNOWLEDGE

- ❖ Whenever discussing any form of learning, the term “**Knowledge**” is frequently used to refer to the expected result of the learning process.
- ❖ It is difficult to provide a satisfactory definition of knowledge, however we can describe how it is being used.
- ❖ The purpose of knowledge is **inference**.

HIERARCHY OF KNOWLEDGE

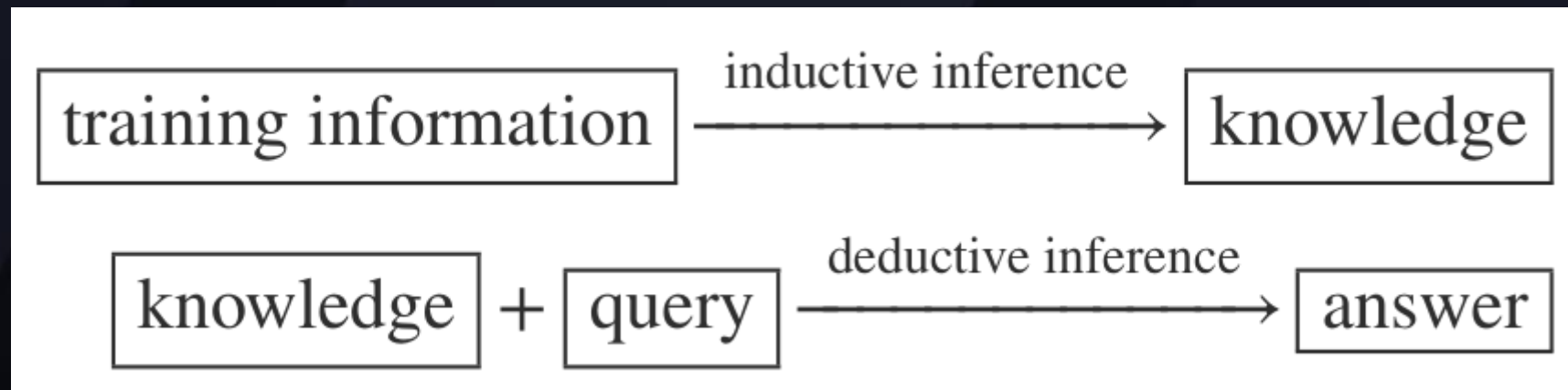
- ❖ **Data**: Is the **raw material** from which information is created.
- ❖ **Information**: is the first processing step towards extracting something useful from data (aggregation, validation, ...) it is **the answer to all questions like who?, What?, When, Where? but not How.**
- ❖ **Knowledge**: Is the process of applying extracted information to fulfill our goals, and it is usually the answer for **How?**
- ❖ **Wisdom**: Is using the knowledge and insights gained from the information to take **proactive decisions or simply the right decision.**



INFERENCE

- ❖ Inference can be considered as the process of using some available knowledge and information to **derive some new knowledge**.
- ❖ Different types of inference are employed when using training information to derive knowledge and when using this derived knowledge.
- ❖ These are :
 - Inductive inference, and
 - Deductive inference.

HOW WE USE INFERENCE?



INDUCTIVE INFERENCE

- ❖ inductive inference is a “specific-to-general” reasoning process.
- ❖ Inductive inference takes a number of individual observations (instances) to generate rules that match these known observations.
- ❖ These rules can be used to predict unknown future observations.
- ❖ Inductive inference is **fallible** and there is no guarantee that it would yield true conclusions when applied to true premises.

DEDUCTIVE INFERENCE

- ❖ Performing deductive inference based on knowledge derived via inductive learning conform to the popular “general-to-specific” pattern.
- ❖ Deductive inference is infallible and its conclusions are guaranteed to be satisfied whenever its premises are true.
- ❖ The infallibility of deductive inference guarantees receiving correct answers to all queries, only if the knowledge on which the deduction is based were perfectly correct, which cannot be expected in practice.

DOMAIN

❖ The domain, designated by X , is the set of all entities that are considered in a given inductive learning task.

❖ Examples:

- Customers,
- Transactions,
- Devices
- ...

INSTANCES

- ❖ Any **single element** of the domain, $x \in X$, is an instance.
- ❖ Instances constitute both **training** information for model creation and **queries** for model application.

ATTRIBUTES

- ❖ Instances are observed by their attributes.
- ❖ An attribute is a function
$$a : X \rightarrow A$$
- ❖ It assigns an attribute value to each instance from the domain.
- ❖ Generally, we assume that there are n attributes defined on the domain X ,
- ❖ $a_1 : X \rightarrow A_1, a_2 : X \rightarrow A_2, \dots, a_n : X \rightarrow A_n$.
- ❖ When speaking of an instance x , we will usually mean the vector of values $a_1(x), a_2(x), \dots, a_n(x)$.

TYPES OF ATTRIBUTES

❖ Nominal.

- Having a finite number of discrete values with no order relation.

❖ Ordinal.

- Having a finite number of discrete values with an order relation.

❖ Continuous

- Having numerical values.

TARGET ATTRIBUTE

- ❖ A distinguished attribute that represents the property of instances that the created model should be able to predict, based on the values of other attributes.

INPUT ATTRIBUTES (FEATURES)

- ❖ Some or all of **non-target** attributes are considered input attributes.
- ❖ The values of which are assumed to be available for the whole domain.
- ❖ There is a possibility of having missing values, which is one of the **common practical data quality issues**.

TRAINING SET

- ❖ A training set is a subset of the domain.
- ❖ the training set $T \subseteq D$ is the set of instances actually used for model creation, where $D \subset X$ denotes the set of all available instances.
- ❖ Why we use only a subset of the whole available set of instances D for model creation?
 - Computational savings,
 - Model evaluation.

MODEL

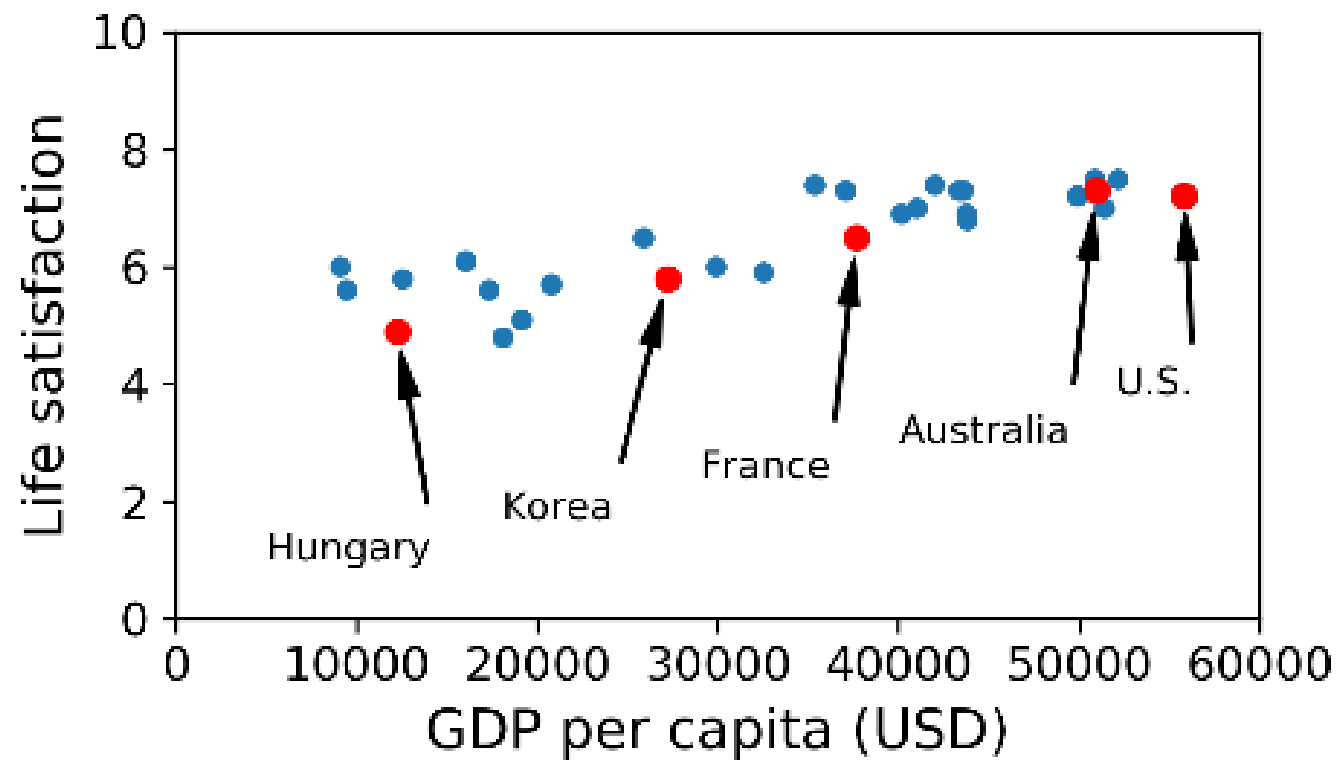
- ❖ Inductive learning task generates a model h based on the provided training set.
- ❖ The generated model h represents knowledge that can be applied to all instances $x \in X$ in an automated way.
- ❖ This is a function that takes an instance as its argument (query) and returns a prediction as its value (answer).
- ❖ It can be therefore considered a new attribute (column), inductively derived from the training data, but defined (computed) for the whole domain.
- ❖ The predicted value is an **approximation** of an existing attribute that is only provided for a limited dataset, but remains unknown for the rest of the domain.

EXAMPLE

❖ Does money make people happier?

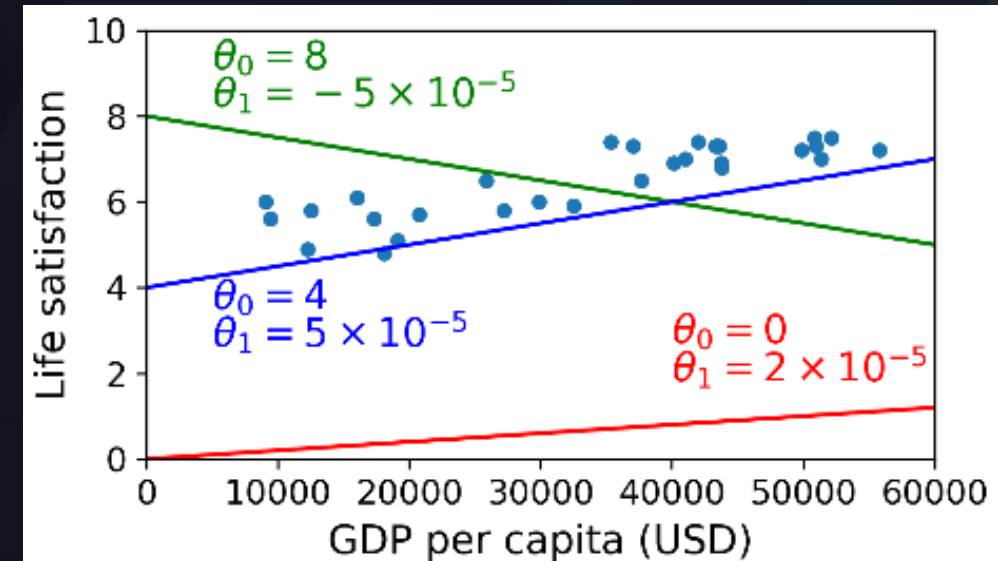
Country	GDP per capita (USD)	Life satisfaction
Hungary	12,240	4.9
Korea	27,195	5.8
France	37,675	6.5
Australia	50,962	7.3
United States	55,805	7.2

EXPLORING DATA



SELECTING THE MODEL

- ❖ You decided to model life satisfaction as a linear function of GDP per capita.
- ❖ you selected a linear model of life satisfaction with just one attribute, GDP per capita.
- ❖ $\text{life_satisfaction} =$
 - $\theta_0 + \theta_1 \times \text{GDP_per_capita}$



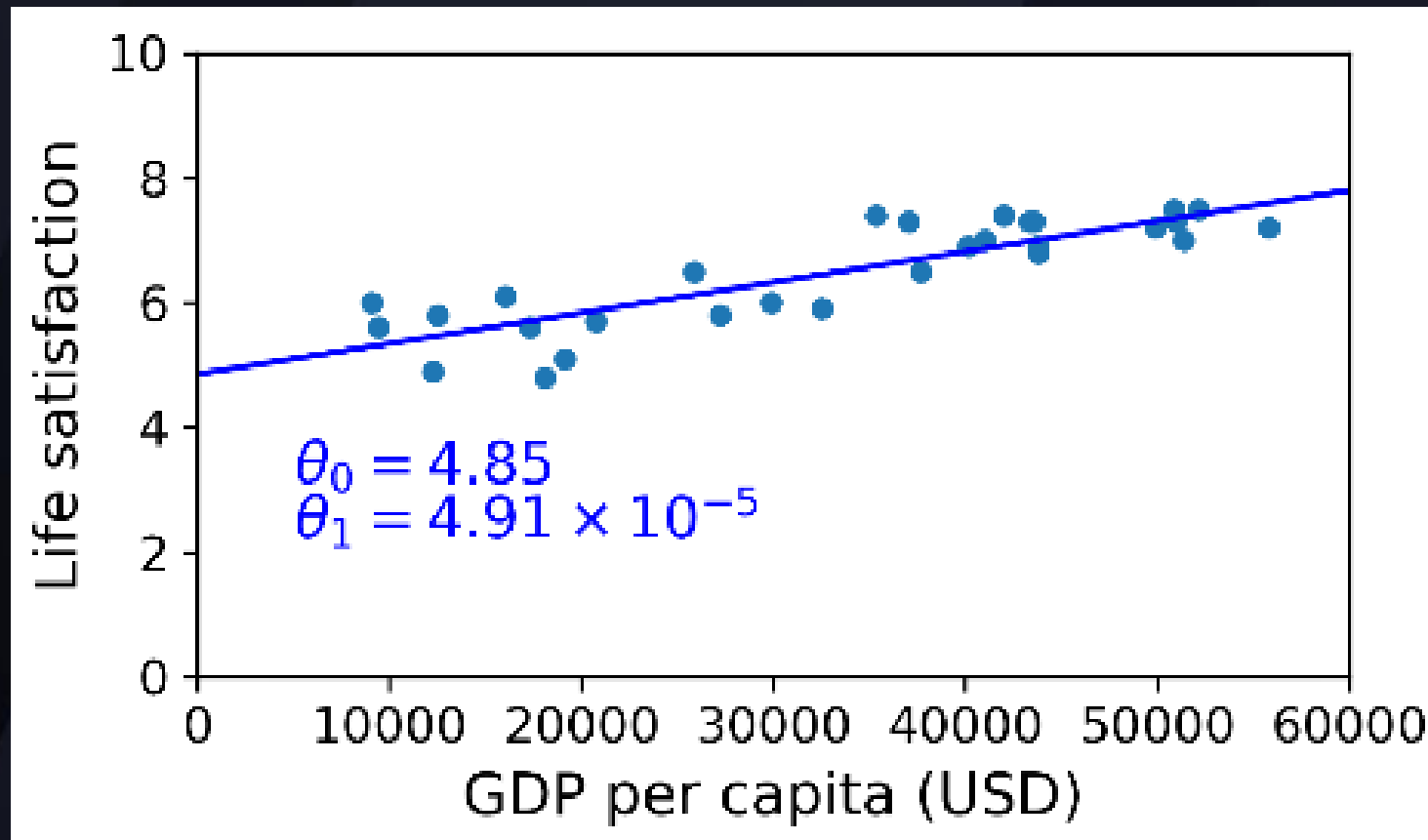
SPECIFY A PERFORMANCE MEASURE

- ❖ A fitness function that measures how good your model is.
- ❖ A cost function that measures how bad it is.

LINEAR REGRESSION

- ❖ Using Linear Regression algorithm, you feed it your training examples, and it finds the parameters that make the linear model fit best to your data.(Training)
- ❖ In our case, the algorithm finds that the optimal parameter values are $\theta_0 = 4.85$ and $\theta_1 = 4.91 \times 10^{-5}$

SOLUTION MODEL



MAKING PREDICTIONS

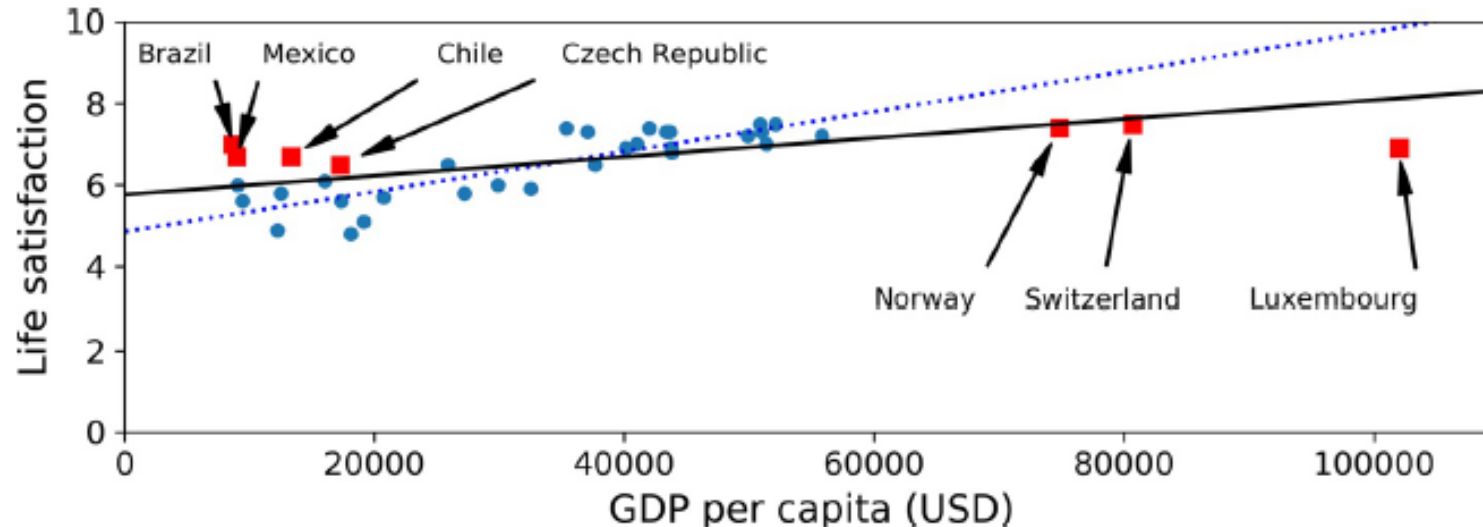
- ❖ Cyprus's GDP per capita is \$22,587
- ❖ Life satisfaction for Cypriots.
- ❖ $4.85 + 22,587 \times 4.91 \times 10^{-5} = \underline{5.96}$.
- ❖ What about KSA?

MAIN CHALLENGES OF MACHINE LEARNING

❖ Insufficient Quantity of Training Data

- Data matters more than algorithms in complex problems.

❖ Non-representative Training Data (Sampling Bias)



MAIN CHALLENGES OF MACHINE LEARNING

❖ Poor-Quality Data

- Mistakes
- Outliers
- Missing
- ...

MAIN CHALLENGES OF MACHINE LEARNING

❖ Irrelevant Features (*feature engineering*)

- *Feature selection*
- *Feature extraction*
- Creating new features (from new data)

MAIN CHALLENGES OF MACHINE LEARNING

❖ Overfitting the Training Data

- The model performs well on the training data, but it does not generalize well.
- Regularization.(reduce the risk of overfitting)
- *Hyper-parameters*: a parameter of a learning algorithm.

❖ Underfitting the Training Data

- The model is too simple to learn the underlying structure of the data.

TESTING AND VALIDATION

❖ *generalization error*: error rate on new cases.



Thank You

midocean.edu.km

