CPSC 5310 MACHINE LEARNING WINTER 2026

DR. DIALA EZZEDDINE

Week 1 Session 2

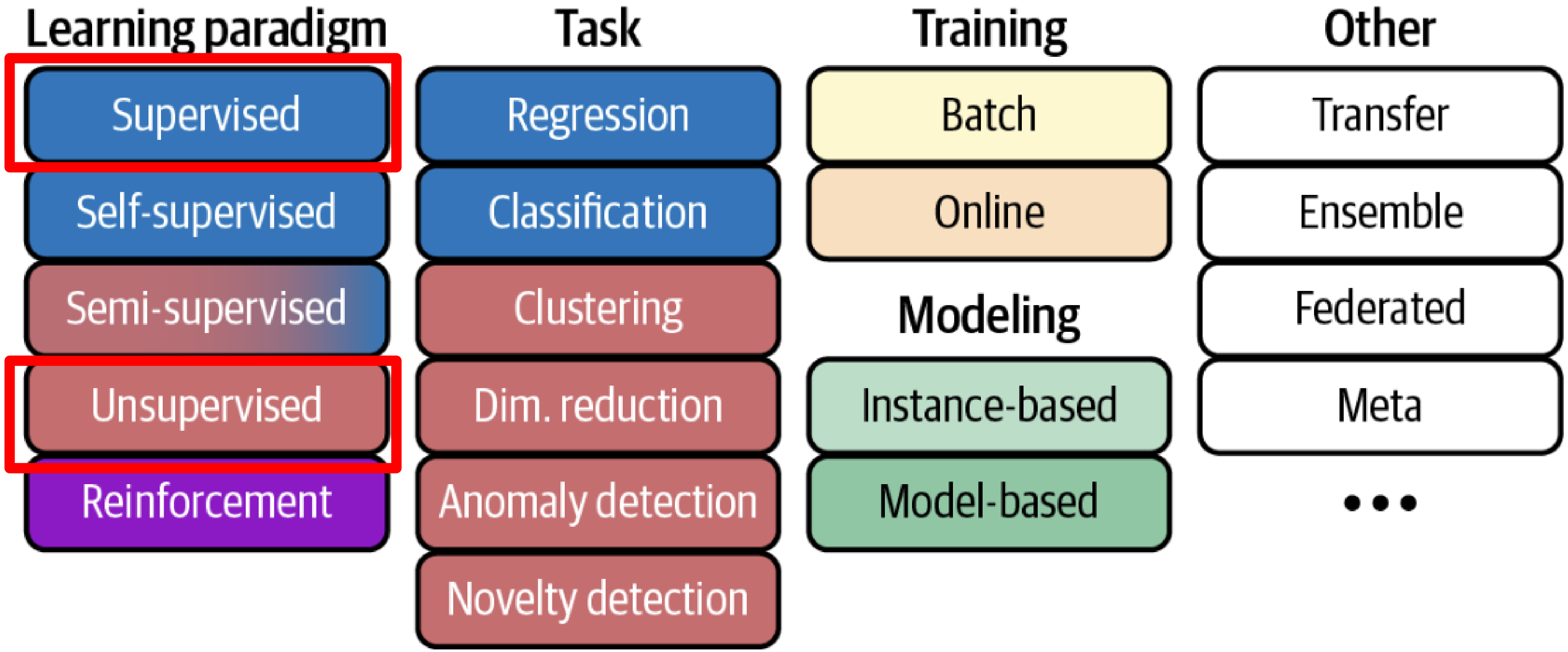
AGENDA

- ML Types
- Challenges of ML
- In-Class activity 2
- Testing and Validating
- Practice: EX2 (Homework)

ML TYPES

Reference:
Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow by Aurélien Géron, 3rd edition, O'Reilly Media.

TYPES OF MACHINE LEARNING SYSTEMS



Learning paradigm

Supervised

Self-supervised

Semi-supervised

Unsupervised

Reinforcement

Task

Regression

Classification

Clustering

Dim. reduction

Anomaly detection

Novelty detection

Training

Batch

Online

Modeling

Instance-based

Model-based

Other

Transfer

Ensemble

Federated

Meta

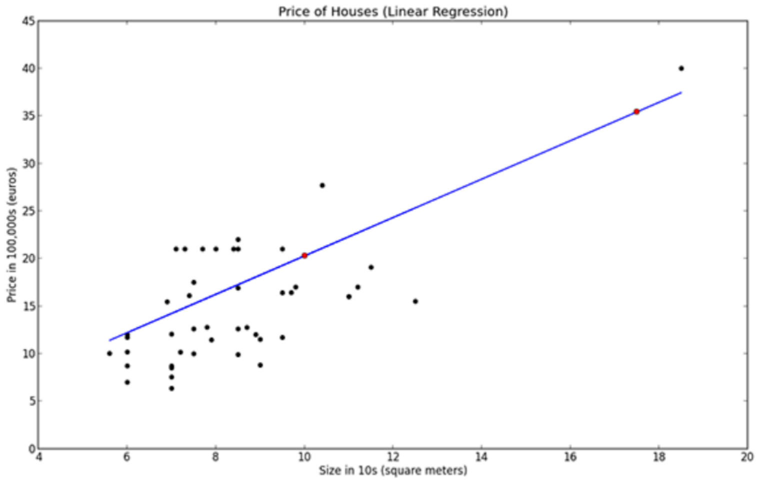
• • •

Figure 1-20. Overview of ML categories

SUPERVISED LEARNING

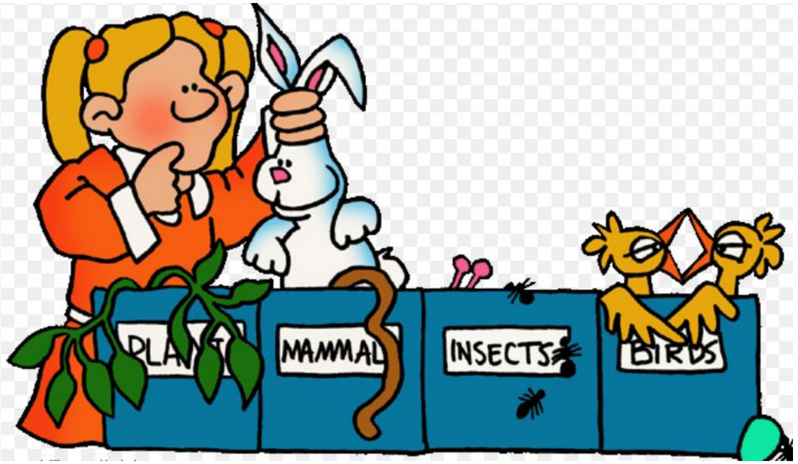
- Supervised learning is defined as when a model gets trained on a Labeled Dataset.
- Labelled datasets have both input and output features.
- In Supervised Learning, algorithms learn to map between inputs and correct outputs.

Supervised Learning



Given a set of data, a supervised machine learning algorithm attempts to optimize a function(model) to find the combination of feature values that result in the target output.

1. The learning task of predicting which category example belongs to is known as **classification**. In classification, the target feature to be predicted is a categorical feature known as the **class** and is divided into categories called **levels**.
2. The learning task of predicting numeric data such as test scores, house values or counts of items is known as **numeric prediction.** The widely used **is linear regression**.



UNSUPERVISED LEARNING

- Unsupervised Learning works with unlabeled data, meaning there are no predefined outputs.
- The algorithm finds hidden patterns, groups or relationships within the data on its own.
- It’s mainly used for clustering, dimensionality reduction, patter discovery and data visualization.

UNSUPERVISED MACHINE LEARNING

- Unsupervised machine learning refers to the solving a problem of finding hidden structure within unlabeled data.
- The common unsupervised ML tasks are:
  - **Clustering** is used to dividing a dataset into homogeneous groups. This method is used for Segmentation analysis and recommendation systems.
  - **Dimensionality Reduction** is used to remove irrelevant features, noise, from the data.
  - **Pattern discovery** is used to identify useful associations within data. This method is used often for Market basket analysis.

sample

Cluster/group

SEMI-SUPERVISED VS SELF-SUPERVISED LEARNING

- Semi-supervised learning combines supervised and unsupervised learning by using both labeled and unlabeled data to train a machine learning algorithm for classification and regression tasks.
- Self-supervised learning combines supervised and unsupervised learning by using unlabeled data to generate implicit labels that is then used to train a machine learning algorithm for classification and regression tasks.

REINFORCEMENT LEARNING

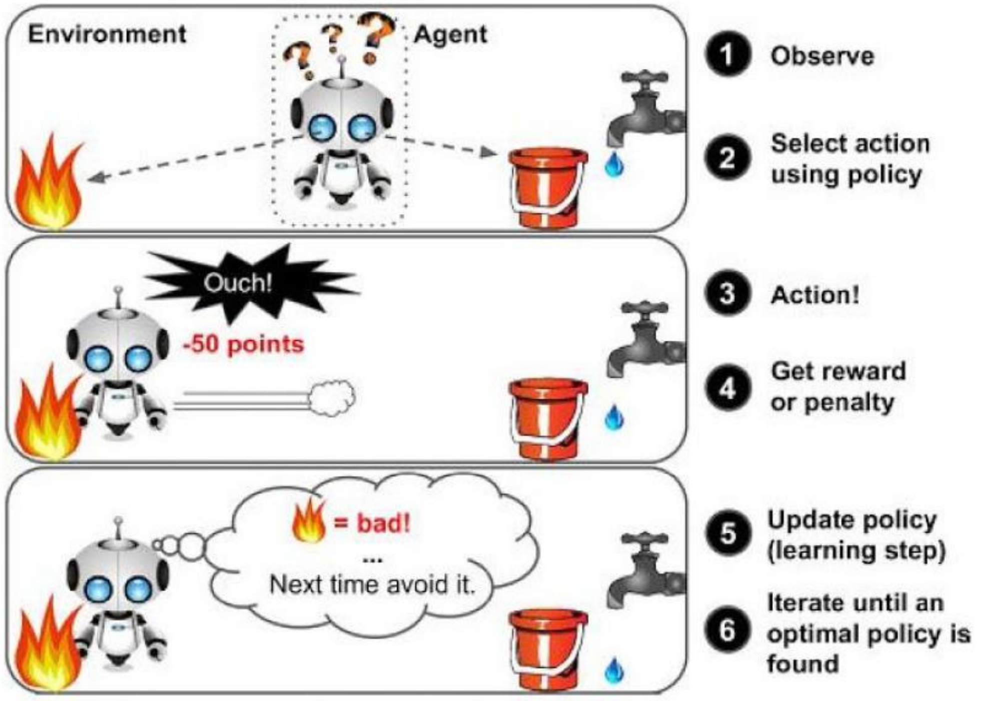
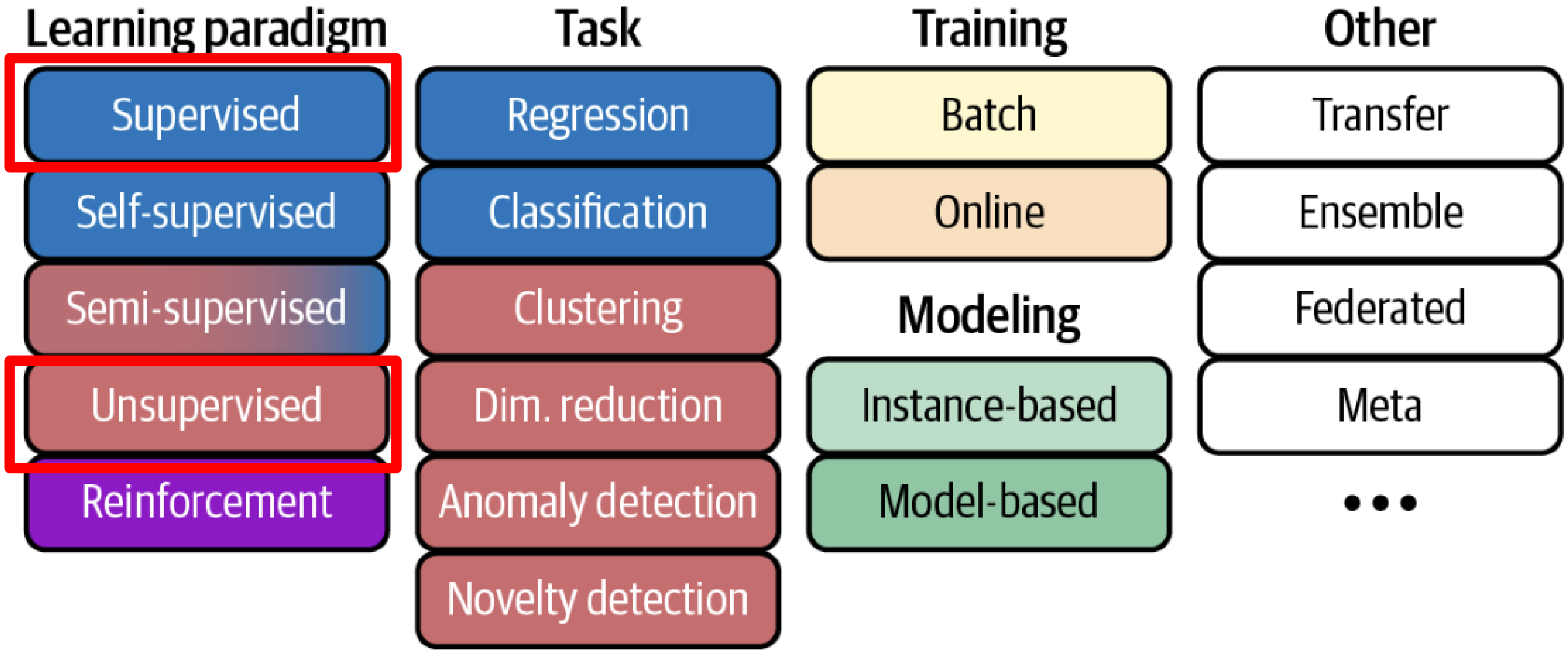

Environment
Agent
1 Observe
2 Select action using policy
Ouch!
-50 points
3 Action!
4 Get reward or penalty
= bad!
...
Next time avoid it.
5 Update policy (learning step)
6 Iterate until an optimal policy is found

Figure 1-12. Reinforcement Learning

- Reinforcement learning is an ML method that enables an agent to learn in an interactive environment by trial and error using feedback from its own actions and experiences.
- It uses a reward system.
- The purpose is to maximize its reward gain.
- It must then learn by itself what is the best strategy, called a ***policy***, to get the most reward over time.
- A policy defines what action the agent should choose when it is in a given situation.

TYPES OF MACHINE LEARNING SYSTEMS



Learning paradigm

- Supervised
- Self-supervised
- Semi-supervised
- Unsupervised
- Reinforcement

Task

- Regression
- Classification
- Clustering
- Dim. reduction
- Anomaly detection
- Novelty detection

Training

- Batch
- Online

Modeling

- Instance-based
- Model-based

Other

- Transfer
- Ensemble
- Federated
- Meta
- •••

Figure 1-20. Overview of ML categories

PRACTICE ML TYPES

What is the appropriate ML type to use in the following examples:

1. Predict a patient's blood pressure level based on age, weight, and exercise habits.
2. Identify if a crop leaf is healthy or diseased from a photo.
3. Netflix is trying learn more about users by their watching habits (action lovers, comedy fans, weekend binge watchers.
4. Learning a language by reading many sentences. You decide to hide some words and learn to guess the missing ones.
5. Reduce 200 survey questions into 2 personality types.
6. An AI tutor learns how to explain topics, so students click "helpful" more often.
7. Studying the chopping cart of your customers.
8. A grocery store wants to group shopper by shopping needs using store collects shopping data and few customers that filled out a survey with their shopping goals.

BATCH VS ONLINE LEARNING

- Batch (Offline) learning: System is trained with all data at once, then used. If you get new data, you have to re-train using old and new data
- Online (Incremental) Learning: System is trained incrementally, online or offline (memory does not fit all the data at once).

INSTANCE-BASED VS MODEL-BASED LEARNING

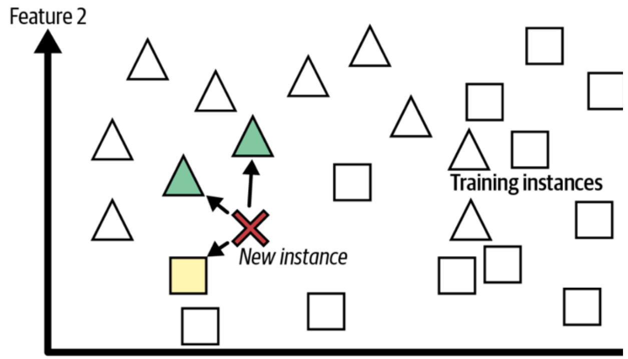


Figure 1-15. Instance-based learning: in this example we consider the class of the thı
the training set

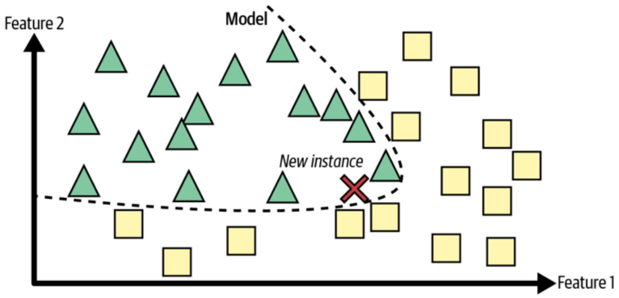


Figure 1-16. Model-based learning

One more way to categorize machine learning systems is by how they generalize.

- Instance-based learning: the system learns the examples by heart, then generalizes to new cases by using a similarity measure.
- Model-based learning: Another way to generalize from a set of examples is to build a model of these examples and then use that model to make predictions.

IN CLASS ACTIVITY

PYTHON LAB!

- Let’s try the Example1-1 from the textbook to compare the instance-based learning to a model-based learning.

Suppose you want to know if money makes people happy, so you download the Better Life Index data from the OECD’s website, and World Bank stats about gross domestic product (GDP) per capita. Then you join the tables and sort by GDP per capita.

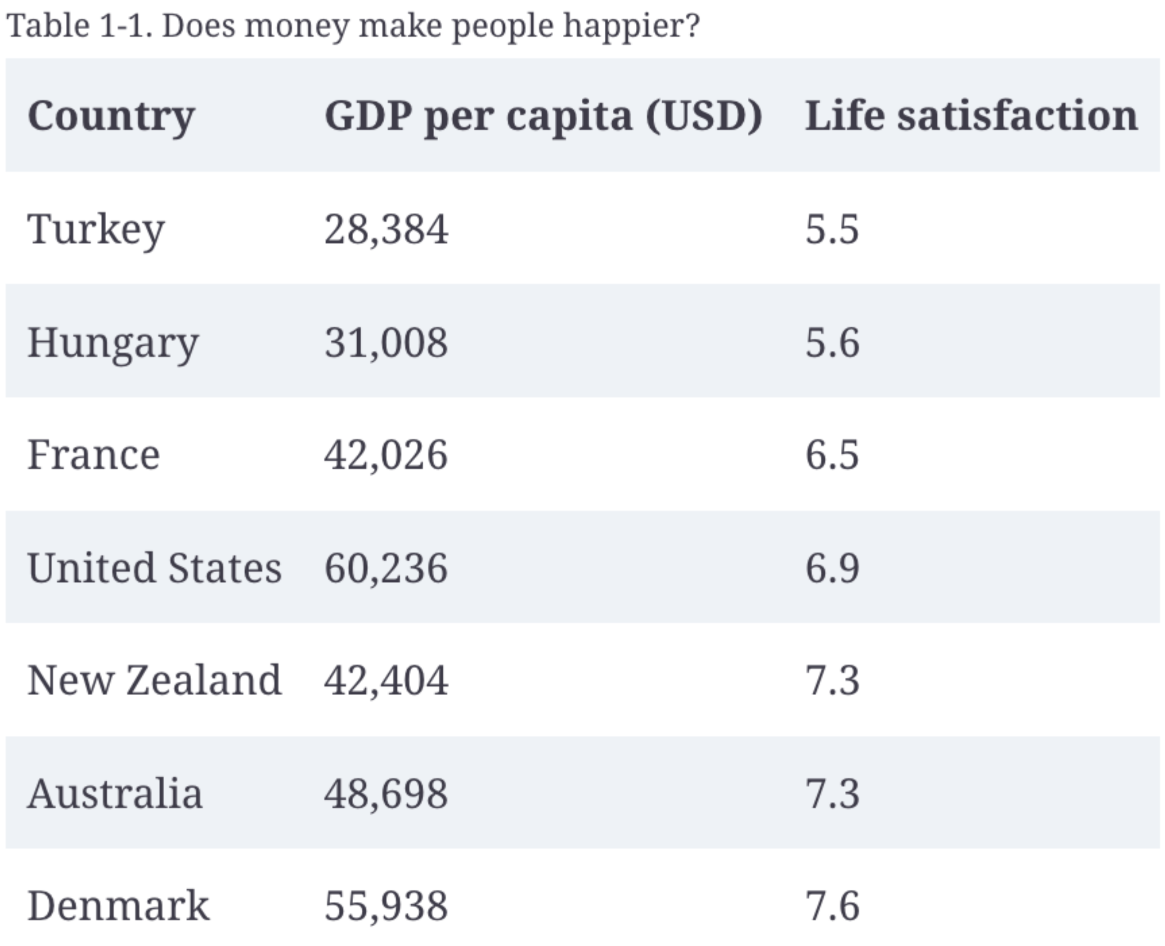


Table 1-1. Does money make people happier?

Country	GDP per capita (USD)	Life satisfaction
Turkey	28,384	5.5
Hungary	31,008	5.6
France	42,026	6.5
United States	60,236	6.9
New Zealand	42,404	7.3
Australia	48,698	7.3
Denmark	55,938	7.6

MAIN CHALLENGES OF MACHINE LEARNING

- **Bad Data:** The system will not perform well if your training set is too small, or if the data is not representative, is noisy, or is polluted with irrelevant features (garbage in, garbage out):
  - Insufficient Quantity of Training Data
  - Nonrepresentative Training Data
  - Poor-Quality Data
  - Irrelevant Features
- **Bad algorithm**: Your model needs to be neither too simple (in which case it will underfit) nor too complex (in which case it will overfit):
  - Overfitting the Training Data
  - Underfitting the Training Data
- **Deployment issues**: Lastly, you must think carefully about deployment constraints.

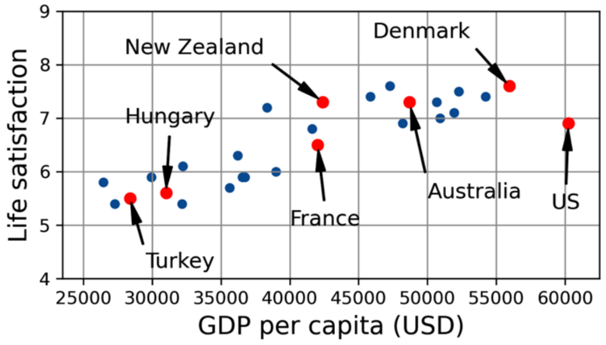


Figure 1-17. Do you see a trend here?

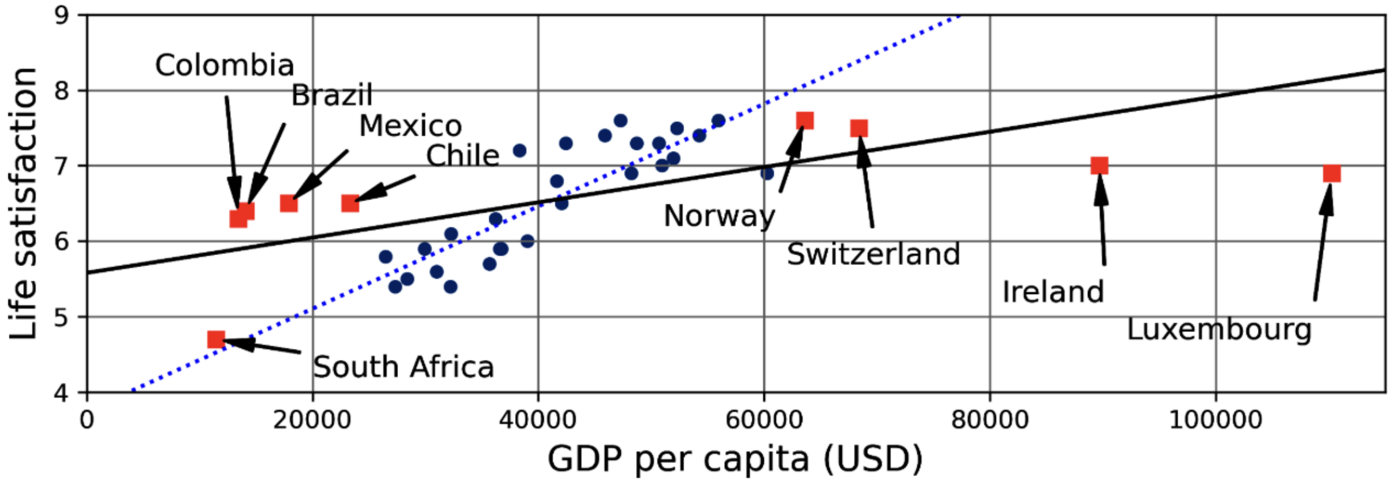


Figure 1-22. A more representative training sample

BAD DATA: NONREPRESENTATIVE TRAINING DATA

BAD DATA: IRRELEVANT FEATURES

- A critical part of the success of a machine learning project is deciding on a good set of features to train on. This process, called ***feature engineering***, involves the following steps:
  - *Feature selection*: selecting the most useful features to train on among existing features.
  - *Feature extraction*: combining existing features to produce a more useful one.
  - Creating new features by gathering new data.

BAD ALGORITHM: OVERFITTING

- Overfitting means that the model performs well on the training data, but it does not generalize well.
- Overfitting happens when the model is too complex, so it starts to learn random patterns in the training data.
- Here are possible solutions:
  - Simplify the model by selecting one with fewer parameters, by reducing the number of attributes in the training data, or by constraining the model.
  - Gather more training data.
  - Reduce the noise in the training data.

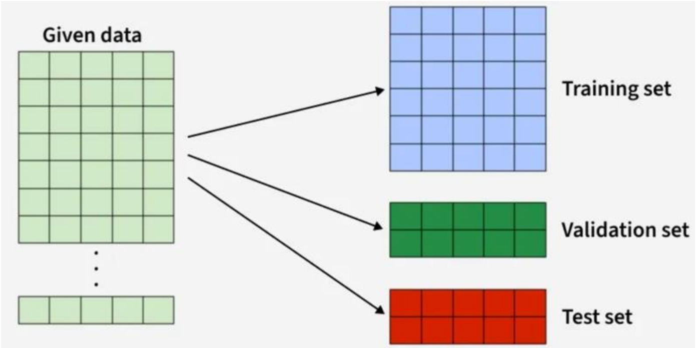
BAD ALGORITHM: UNDERFITTING

- Underfitting occurs when your model is too simple to learn the underlying structure of the data.
- Here are the main options for fixing this problem:
  - Select a more powerful model, with more parameters.
  - Feed better features to the learning algorithm.
  - Reduce the constraints on the model.

PROBLEMS ML *CAN'T* SOLVE WELL

- ***Questions with no data or examples to learn from.***
  ML can't make predictions about something if it has never seen information about it.
- ***Problems that depend on human values, opinions, or personal judgment.***
  Example: decide what is "fair," "beautiful," or "ethical." There is no single correct answer for the model to learn.
- ***One-time events that never repeat.***
  Example: predicting the exact outcome of a surprise family argument that happened only once.
- ***Things that are truly random.***
  Example: guessing the winning lottery numbers.
- ***Questions that need perfect logical certainty, not patterns.***
  Example: proving a new math theorem. ML finds patterns, it doesn't create guaranteed proofs.

TESTING AND VALIDATING



- ***Training Set*** is the portion of the dataset used to fit the machine learning model. During training, the algorithm learns patterns, relationships and parameters directly from this data.
- ***Validation set (***or dev-set, holdout validation): hold out part of the training set to evaluate several candidate models and select the best one. It is used to tune model hyperparameters and make design decisions during training. Unlike the training set, it is not used to update model weights directly. Instead, it provides an unbiased estimate of model performance during development.
- ***Test set*** is a completely independent subset used to evaluate the final model's performance after all training and tuning are complete. It simulates how the model will perform on unseen, real-world data and provides the most reliable estimate of generalization.
- ***Train-dev set***: holdout of 2 separates validation sets: The train-dev set is used when there is a risk of mismatch between the training data and the data used in the validation and test datasets.

https://www.geeksforgeeks.org/machine-learning/training-vs-testing-vs-validation-sets/

EX2-DATA ANALYSIS

You will work on a set of problems of exploratory data analysis.

Part 1:

Practice EDA with the given examples.

Part 2:

Write a summary of your findings from data analysis.

COMING UP…
IN THE NEXT EXCITING EPISODE

NEXT CLASS

- Week 2 Session 1:End to end ML project part 1
- Homework
  - Review the slides from today.
  - DC2: Preprocessing for Machine Learning in Python, due on Thursday Jan 15 at midnight.
  - ER1: Overfitting and Underfitting, due on Monday Jan 12 at midnight.
  - Submit EX2- Data Analysis, due on Monday Jan 12 at midnight.