- Module 3_1 -
Statistical Learning

Outline

- Machine Learning Basics
- Supervised Learning
  - Regression
  - Classification
- Unsupervised Learning
  - Clustering
- Reinforcement Learning

Machine Learning Basics (1)

- **Machine learning** is closely related to the fields of artificial intelligence, data mining and statistics and plays a key role in many areas of science, finance and industry. Examples of learning problems in these areas:
  - **Predict** whether a patient, hospitalized due to a heart attack, will (and when) have another heart attack
    (based on demographic, diet and clinical measurements)
  - **Predict** the price of a stock
    (based on company performance and economic data)
  - **Estimate** the amount of blood glucose of a diabetic person
    (based on the infrared absorption spectrum of that person's blood)
  - **Identify** the risk factors for prostate cancer
    (based on clinical and demographic variables)
  - **Identify** the numbers in a handwritten ZIP code
    (from a digitized image)
  - …

Machine Learning Basics (2)

- In general, machine learning enables the tackling of tasks that
  - are too complex to solve with fixed programs designed and written by humans
  - require adaptation after deployment
- A **machine learning algorithm** is an algorithm that is able to learn from data
- T.M. Mitchell (1997) provided a succinct definition of **learning** as follows:

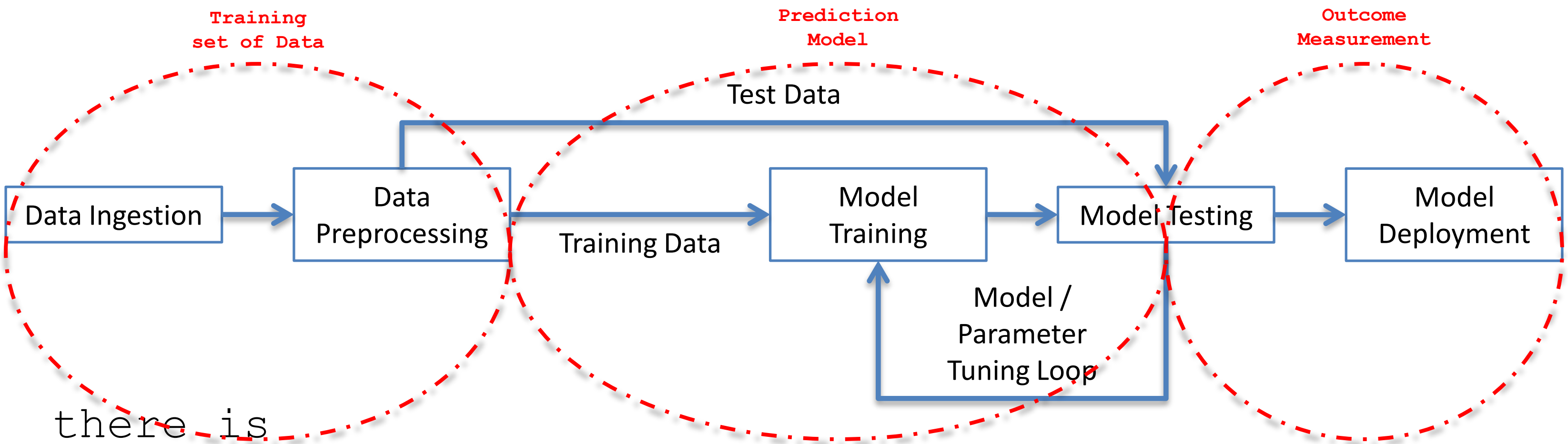
"A computer program is said to learn
from **experience** E
with respect to some class of **tasks** T
and **performance measure** P,
if its performance at tasks in T,
as measured by P,
improves with experience E."

Machine Learning Basics (3)

- Machine learning algorithms can be broadly categorized as **supervised learning, unsupervised learning** or **reinforcement learning** by the kind of experience they are allowed to have during the learning process
  - **Supervised learning** algorithms experience a **dataset** containing **features**, but each **sample** (or **example)** is also associated with a **label** (or **target** or **outcome**)
  - **Unsupervised learning** algorithms experience a dataset containing features, then learn useful properties of the structure of this dataset
  - **Reinforcement learning** algorithms interact with an environment, so there is a feedback loop between the learning system and its experiences, i.e., learns from a series of reinforcements (rewards or punishments)

Machine Learning Basics (4)

- In a typical (supervised) **machine learning workflow,**



there is

- a **training set** of **data,** in which the <u>outcome</u> and <u>feature</u> measurements is <u>observed</u> for a set of <u>objects</u> (e.g., cars)
- a **prediction model** is <u>built</u> using this <u>training data</u>, which will enable the <u>outcome prediction</u> for <u>new unseen data</u>
- an <u>outcome</u> measurement, usually **quantitative** (e.g., stock price) or **qualitative** (or **categorical**) (e.g., heart attack/no heart attack), that is <u>predicted</u> based on a set of <u>features</u> (e.g., diet and clinical measurements)

- A <u>good</u> <u>prediction model</u> is one that <u>accurately</u> <u>predicts</u> such an <u>outcome</u>

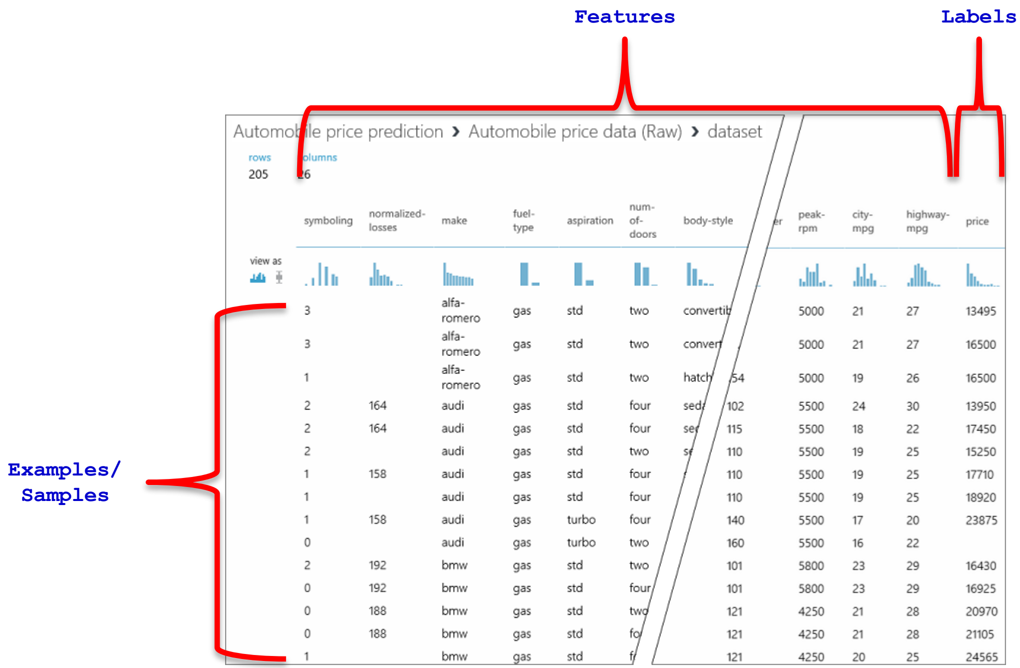
Machine Learning Basics (5)

- Usually the performance measure P is specific to the task T being carried out by the system
- To determine whether the machine learning algorithm **generalizes well** to **new unseen data**, the performance measures are evaluated using a **test set** of **data** that is separate from the training set of data used for training and optimizing the machine learning system

- To evaluate the abilities of a machine learning algorithm, quantitative measures of its performance are required:
  - **Accuracy:** proportion of examples for which the model produces the correct output
  - **Error Rate:** proportion of examples for which the model produces an incorrect output

Machine Learning Basics (6)

- One common way of describing a dataset is with a **design matrix.** A design matrix is a matrix containing a different sample in each row
- Each column of the matrix corresponds to a different feature



symboling	normalized-losses	make	fuel-type	aspiration	num-of-doors	body-style	er	peak-rpm	city-mpg	highway-mpg	price
3		alfa-romero	gas	std	two	convertib		5000	21	27	13495
3		alfa-romero	gas	std	two	convert		5000	21	27	16500
1		alfa-romero	gas	std	two	hatch	54	5000	19	26	16500
2	164	audi	gas	std	four	seda	102	5500	24	30	13950
2	164	audi	gas	std	four	se	115	5500	18	22	17450
2		audi	gas	std	two	s	110	5500	19	25	15250
1	158	audi	gas	std	four		110	5500	19	25	17710
1		audi	gas	std	four		110	5500	19	25	18920
1	158	audi	gas	turbo	four		140	5500	17	20	23875
0		audi	gas	turbo	two		160	5500	16	22	
2	192	bmw	gas	std	two		101	5800	23	29	16430
0	192	bmw	gas	std	four		101	5800	23	29	16925
0	188	bmw	gas	std	two		121	4250	21	28	20970
0	188	bmw	gas	std	fo		121	4250	21	28	21105
1		bmw	gas	std	f		121	4250	20	25	24565

- In the case of supervised learning, the sample contains a label as well as a collection of features

Machine Learning Basics (7)

- Note that raw data rarely comes in the form that is necessary for the optimal performance of a machine learning algorithm
- **Data preprocessing** of the raw data is a crucial step - some of the selected features may be highly correlated and therefore redundant to a certain degree, in these cases, **dimensionality reduction techniques** are useful for compressing the features onto a lower dimensional subspace
  - Reducing the dimensionality of the feature space have the advantages that less storage space is required and that the learning algorithm can run much faster

Machine Learning Basics (8)

- Variable types and terminology:
  - **Input variable** denoted by symbol X
    (if X is a vector, its components can be accessed by subscripts X_j)
  - **Quantitative output** denoted by Y
  - **Categorical output** denoted by G
  - **Generic aspects** of a variable are written in uppercase (e.g., X, Y, G)
  - **Observed values** are written in lowercase (e.g., the ith observed variable of X is written as x_i, where x_i can be a scalar or vector)
  - **Matrices** are represented by bold uppercase letters (e.g., $\boldsymbol{X}$)
  - **Vectors** are represented by bold (when it has N elements) lowercase and assumed to be column vectors (e.g., the ith row of $\boldsymbol{X}$ is x_i^T)
  - Set of **measurements** (observed data): (x_i, y_i) or (x_i, g_i), $i = 1, \dots, N$

Supervised Learning (1)

- The main goal of supervised learning is to learn a model from labeled training data that allows for predictions on new unseen data. The term supervised refers to a set of samples where the desired labels are known

- Given a **training set** of N example (labeled) input-output pairs

$$(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$$

where x and y can be any value and each y_i was generated by an unknown function $y_i = f(x_i)$

➔ the goal is to discover a hypothesis h that approximates the true function f

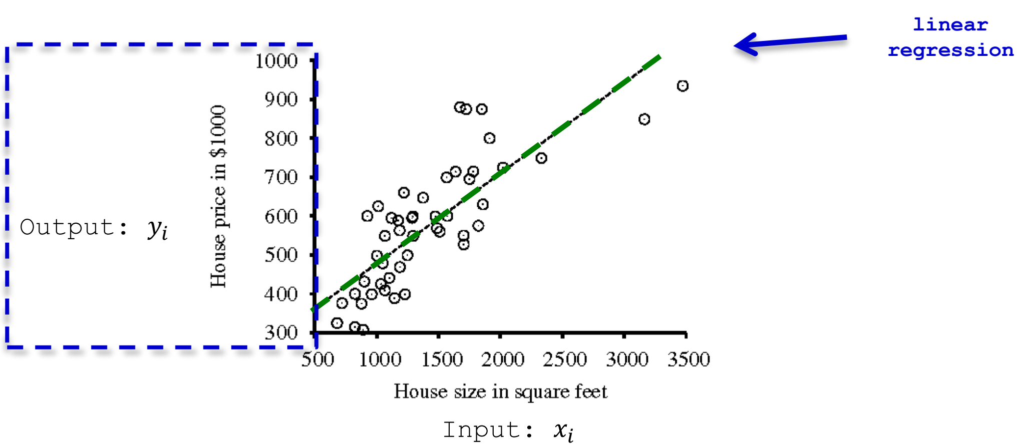
Supervised Learning (2)

- Each training input x_i is typically a vector of feature values
  (E.g., weight and height of a person; humidity and atmospheric pressure of a location on a given day)

- Each training output y_i can be
  - a number (**continuous** or **quantitative** output value)
    (E.g., temperature in Celsius of a location on a given day)
    ➔ **regression**
  - a non-numerical value (**categorical** or **qualitative** output)
    (E.g., {*sunny*, *cloudy*, *rainy*, *snowy*})
    ➔ **classification**

Supervised Learning: Regression (3)

- **Regression** is a subcategory of supervised learning where the goal is the prediction of continuous outcomes. In regression, given a number of features, p, and a continuous outcome, the objective is to find a relationship (a function $f\colon \mathbb{R}^p \to \mathbb{R}$) between those features to predict an outcome

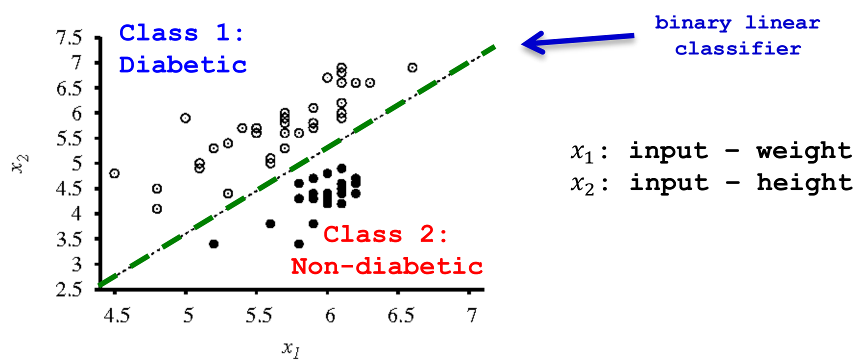


Supervised Learning: Regression Applications (4)

- Regression Applications:
  - House price prediction
    - Input: house square footage, location, size of land, employment rate, etc.
    - Output: house price
  - Cancer survival prediction
    - Input: tumor size, age at diagnosis, family history, treatments received, etc.
    - Output: likelihood of 5-year survival
  - Stock price prediction
    - Input: oil prices, exchange rates, interest rates, stock price indices in other countries, other side information (twitter tweets), etc.
    - Output: tomorrow's stock prices

Supervised Learning: Classification (5)

- **Classification** is another subcategory of supervised learning where the goal is the prediction of categorical labels. In classification, given a number of features, p, and a categorical outcome, the objective is to find a relationship (a function $f\colon \mathbb{R}^p \rightarrow \{1, \ldots, C\}$) to predict which of C **classes** (or categories) a sample belongs to

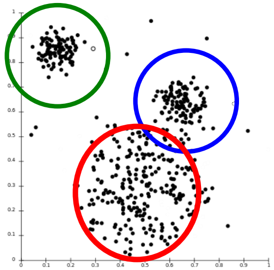


Supervised Learning: Classification Applications (6)

- Classification Applications:
  - Spam filtering
    - a binary classification problem
    - Input: email subject and message content
      (extracting keywords to form “bag of words”,
      i.e., number of occurrences of ‘buy’, ‘viagra’, etc.)
    - Output: spam or non-spam
  - Movie genre classification
    - a multiclass classification problem (E.g., a genre is a class)
    - Input: plot summary
      (extracting keywords from the summary to form “bag of words”,
      i.e., number of occurrences of ‘love’, ‘laugh’, etc.)
    - Output: movie genre (romance, comedy, thriller, etc.)
  - Object recognition
    - a multiclass classification problem (can be lots of classes)
    - Input: image (pixel RGB values)
    - Output: object class (car, traffic light, motorcycle, pedestrian, bicycle, etc.)

Unsupervised Learning (1)

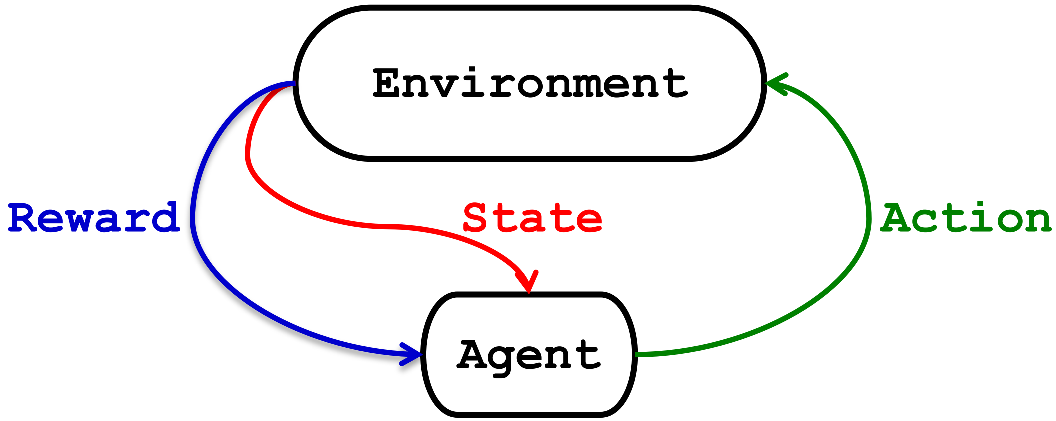
- In unsupervised learning, only features are observed with no measurements of the outcome. The task is to explore the structure of the data (how the data are organized or clustered) in order to extract meaningful information without the guidance of a known outcome variable
- **Clustering** is an exploratory data analysis technique that allows the organization of information into meaningful subgroups (or **clusters**) without having any prior knowledge of their group memberships
  - Each cluster that arises defines a group of objects that share a certain degree of similarity but are more dissimilar to objects in other clusters



Unsupervised Learning: Applications (2)

- Unsupervised learning Applications:
  - E-commerce
    - Clustering
    - Online retailers cluster users into groups based on their previous purchases or web-surfing behaviour ➔ send targeted advertising to each group of potential customers
  - Fraud detection
    - Anomaly detection
    - Credit card companies track the spending behaviour of a user and detects transactions that deviate from prior transactions
  - Data visualization
    - Dimensionality reduction
    - Projections of high-dimensional data (e.g., gene expression) into a lower-dimensional space (e.g., 2D) for easier data visualization

Reinforcement Learning (1)



- In reinforcement learning, the agent learns from experience in the absence of existing training data
  ➔ the agent collects the training samples (e.g., "this action was good", "that action was bad") through **trial-and-error** as it attempts its task, with the goal of maximizing **long-term reward**

Reinforcement Learning: Applications (2)

- Reinforcement learning Applications:
  - Games: DeepMind's AlphaGo Zero (2017)
    - Trained solely by self-play reinforcement learning: starting from random play, without any supervision or use of data from real human games → after three days of self-play training, AlphaGo Zero defeated AlphaGo by 100 games to 0
  - Biomedicine
    - Lots of data but difficult to create good training datasets ← reinforcement learning may be useful in designing drugs for drug targets, predicting protein folding and predicting drug effects

https://aws.amazon.com/deepracer/