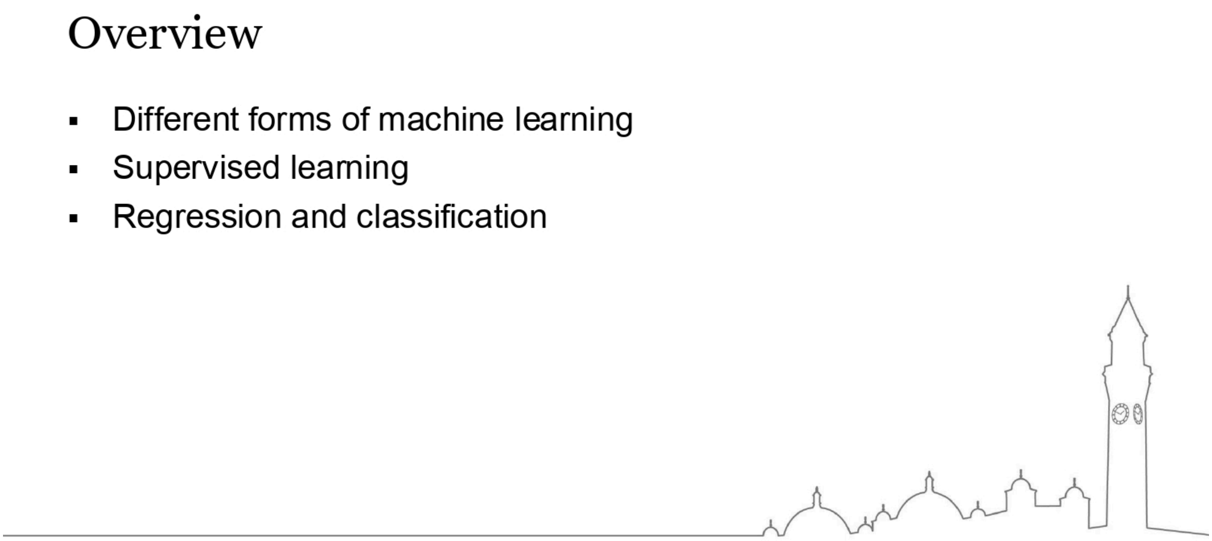
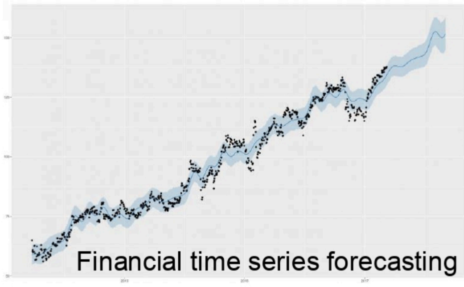
W2.1 Intro to Supervised Learning



Machine Learning Problems

Machine learning problems are those that require a model to be built automatically from data, e.g. to make classifications, estimations or predictions.



UNIVERSITY OF BIRMINGHAM

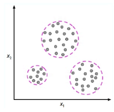
Forms of Machine Learning

- Three forms (if we look at the *input* of the model): supervised learning, unsupervised learning, reinforcement learning
- Supervised learning
  - ➢ The most prevalent form
  - ➢ Learning with a teacher
  - ➢ Teacher: expected output, label, class, etc.
- Solve 2 types of problems (if we look at the *output* of the model): classification, regression problems

UNIVERSITY OF BIRMINGHAM

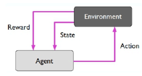
Forms of Machine Learning

- Unsupervised learning
  - ➢ Learning without a teacher
  - ➢ To find hidden structure/insights in data
  - ➢ Clustering, e.g. product recommendation, sport strategy discovery



Clustering

- Reinforcement learning
  - ➢ Learning with (delayed) feedback/reward
  - ➢ Learn series of actions, e.g. chess, robots, …

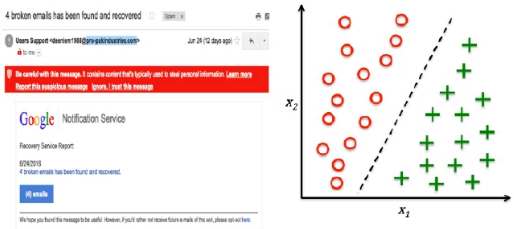


RL

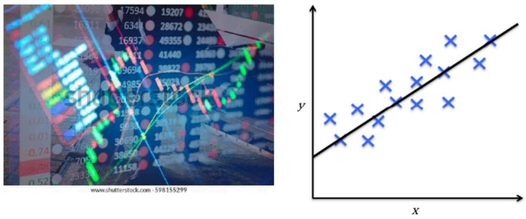
UNIVERSITY OF BIRMINGHAM

Types of supervised learning

Spam detection



Stock price prediction



UNIVERSITY OF BIRMINGHAM

Formulate supervised learning

- Task:
  - Given some input x,
  - Predict an appropriate output y
- Goal: a function f such that $f(x) = y$

The learning process:

1) Have: examples of input-output pairs → training data
   $\left(x^{(1)}, y^{(1)}\right), \left(x^{(2)}, y^{(2)}\right), \ldots, \left(x^{(n)}, y^{(n)}\right)$
2) Supervised learning helps find a good f → training/modelling
3) Given a new input $x^{(n+1)}$, predict its output $y^{(n+1)}$ → prediction

UNIVERSITY OF BIRMINGHAM

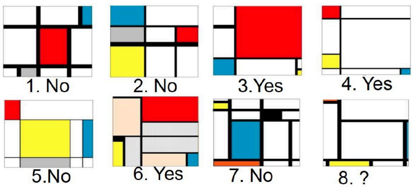
Training Data

- Supervised learning needs annotated data for training:
  In the form of examples of (input, output) pairs.

- After training completed,
  You present it with new input that it hasn't seen before
  It needs to predict the appropriate output.

UNIVERSITY OF BIRMINGHAM

Is painting 8 a genuine Mondrian?



UNIVERSITY OF BIRMINGHAM

Annotated training data

Attributes | Labels

Examples

Number	Lines	Line types	Rectangles	Colours	Mondrian?
1	6	1	10	4	No
2	4	2	8	5	No
3	5	2	7	4	Yes
4	5	1	8	4	Yes
5	5	1	10	5	No
6	6	1	8	6	Yes
7	7	1	14	5	No

Painting 8

Number	Lines	Line types	Rectangles	Colours	Mondrian?
8	7	2	9	4	???

General notations we often use

Lines	Line types	Rectangles	…	Mondrian?
		$\boldsymbol{x}^{(1)}$		$y^{(1)}$
		$\boldsymbol{x}^{(2)}$		$y^{(2)}$
		$\boldsymbol{x}^{(3)}$		$y^{(3)}$
		…		…

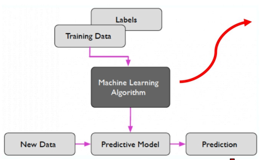
Vector notation:

$$\boldsymbol{x}^{(i)} = \underbrace{\left(x_1^{(i)}, x_2^{(i)}, x_3^{(i)}, \ldots, x_d^{(i)}\right)}_{\text{Attributes, d-dimensional}}$$

The input of the i-th example

UNIVERSITY OF BIRMINGHAM

Supervised learning workflow



Training phase:
$(\boldsymbol{x}^{(1)}, y^{(1)}), (\boldsymbol{x}^{(2)}, y^{(2)}), \ldots, (\boldsymbol{x}^{(n)}, y^{(n)})$

$f(\boldsymbol{x}) = y$

Test/Use phase:
$\boldsymbol{x} \longrightarrow f(\boldsymbol{x}) \longrightarrow y$

New input

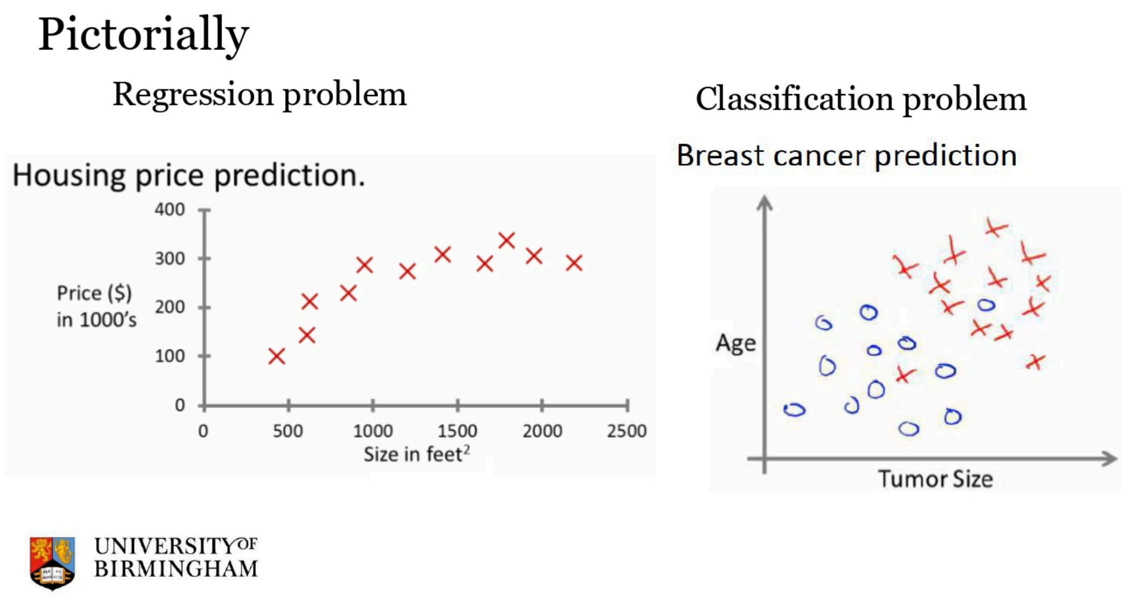
UNIVERSITY OF BIRMINGHAM

Source: Python machine learning, 2nd ed, Sebastian Raschka

Terminology in supervised learning

- Input = attribute(s) = feature(s) = independent variable(s)
- Output = target = response = dependent variable
- Function = hypothesis = predictor

UNIVERSITY OF BIRMINGHAM



Regression

is about predicting a continuous output. In a regression problem, the goal is to predict a numerical value based on input features. For example:

- Predicting the price of a house given its size, number of bedrooms, location, etc.
- Estimating the temperature for the next day based on historical weather data.

Classification

is about predicting a discrete output. In a classification problem, the goal is to assign an input to one of a set of predefined categories. For example:

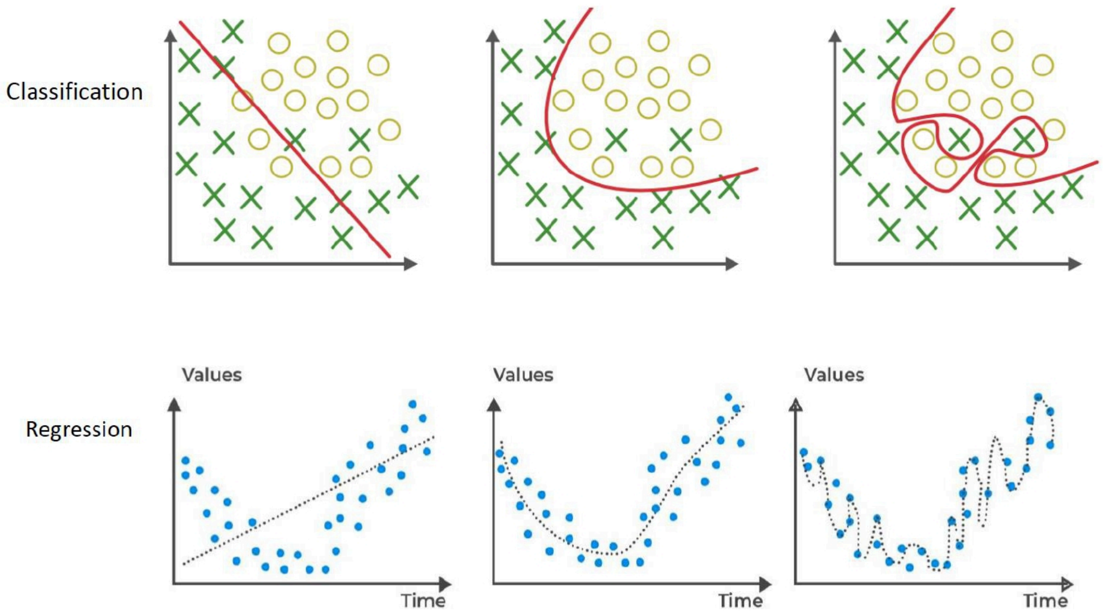
- Identifying whether an email is "spam" or "not spam" based on its content.
- Classifying an image as either a "cat" or "dog."

Pause. Is this some magic?

So…

- There is an unknown function we are after.
- We are given the function values at n specific points only (training set)
- Is it really possible to find out the function values at other points?

UNIVERSITY OF BIRMINGHAM



Classification:

- In a classification problem, we draw decision boundaries to separate different categories.
- Everything on one side of the boundary belongs to one category, and everything on the other side belongs to another.

Regression:

- In regression, instead of drawing a boundary, we fit a line (or curve) through the data points.
- This line represents the relationship between the input features and the continuous output variable.
- Here's what the regression line is saying:
    - **Prediction:**
        - For any given input value, the regression line helps us predict the output value.
        - For example, if you're predicting house prices based on size, you can use the line to estimate the price for a house of a specific size.
    - **Trend:**
        - The slope and shape of the regression line indicate the trend in the data.
        - For instance, if the line slopes upward, it suggests that as the input value increases, the output value also increases.
    - **Extrapolation:**
        - You can use the regression line to estimate values outside the range of your data.
        - However, extrapolation should be done cautiously, as the farther you go from your data points, the less reliable the predictions become.

Overfitting and Underfitting

- Fitting the training data too well is BAD! Why?
- Remember the data you actually want to classify, or predict for, is not the same as the training data –so learning every irrelevant detail (noise) in a training data set will not help.
- Overfitting happens when the model is more complex than required.
- Underfitting happens when the model is simpler than required.

UNIVERSITY OF BIRMINGHAM

Applications of Supervised Learning

- Handwriting recognition
  - When you write an envelope, algorithms can automatically route envelopes through the post.
- Computer vision & graphics
  - When you go out during lockdown, object detection & vision tracking algorithms can automatically detect compliance with the rules.
- Bioinformatics
  - Algorithms can predict protein function from sequence.
- Human-computer interaction
  - Algorithms can recognize speech, gestures, intention.

UNIVERSITY OF BIRMINGHAM