Introduction to AI

2- Supervised learning: Regression and Classification

Jaya Nilakantan

Dawson College

jnilakantan@dawsoncollege.qc.ca

Last week

- Artifical Intelligence
- Machine learning
- Deep learning

Within Machine learning:

- supervised learning
- unsupervised learning
- reinforcement learning

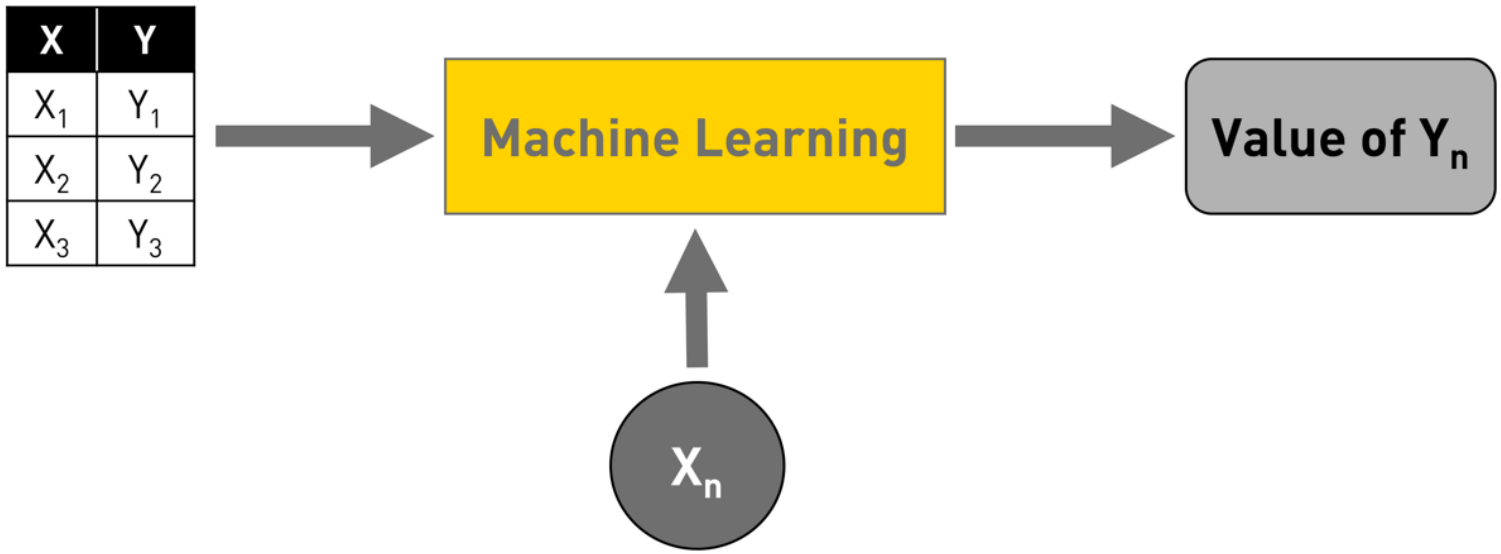
Supervised learning

Supervised learning starts with a set of labeled data (called training data)

- data has features (think as parameters) and labels (think as the answer)

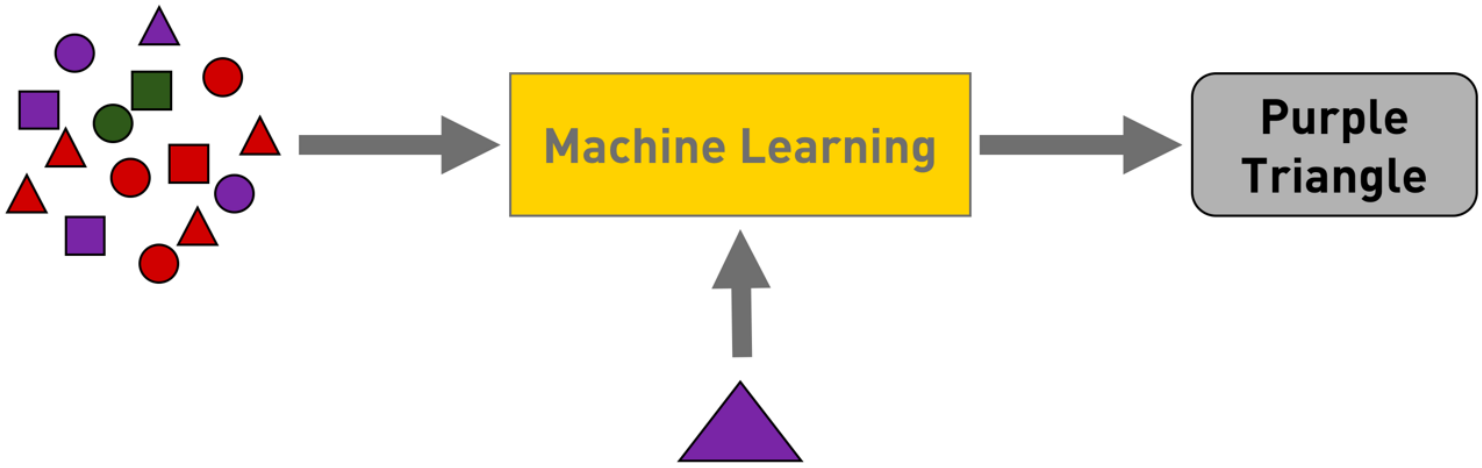
The training data is used to make a mathematical or statistical model.
The model is used for either prediction or classification purposes.
The model is tested with test data: test data is like the training data, it is labelled

Predictions



Train and test model with feature vector $x^{(i)}$ and labels $y^{(i)}$
Use to predict value y given feature vector x
Examples: stock price in 6 months, weather tomorrow at noon, position of car in 10 seconds

Classification



Train and test model with feature vector $x^{(i)}$ and labels $y^{(i)}$
Use to predict value y given feature vector x
Examples: facial recognition, natural language processing, sentiment analysis

Regression problems

The input (features) and output (label) are continuous (e.g. numerical values, such as someone's height or salary)
Since they are continuous, they can be expressed as a function (straight line or more complex)
Family of algorithms:

- Linear single variable regression
- Linear multi-variable regression
- Polynomial regression (not linear)
- regression trees/random forest (build a decision tree based on training data)

Within the algorithm, there are more choices related to optimizing the cost (or objective) function

- Ordinary least squares
- gradient descent

Classification problems

The input (features) may be continuous or discrete, but the output (label) is discrete (e.g. text value, such as emotion)
Family of algorithms:

- Logistic regression: like linear regression, but only 2 output (logistic = logic)
- Naive Bayes: find outcome with highest probability
- K-Nearest-Neighbours: clustering to find likely class

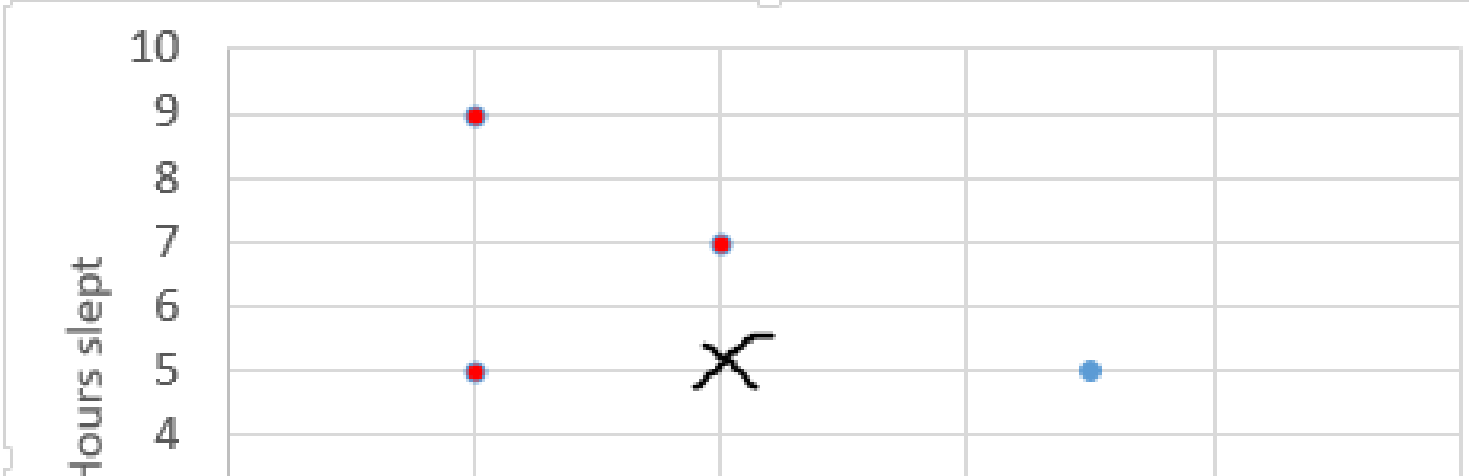
Aside: How to choose the best algorithm?

Each algo has pros and cons depending on your dataset
time efficiency: How long does the algorithm take to complete? (Big O Notation)
space efficiency: how much memory is required?

- at training
- with every test

K-Nearest Neighbours

Hours studied	Hours slept	Result
2	9	Fail
2	5	Fail
4	7	Fail
7	5	Pass
6	2	Pass
8	2	Pass



K-Nearest Neighbours

- Does new data point belong to the Pass cluster or the Fail cluster?
- To which other training data points it is close?
- Guess its classification based on the neighbours.

K-Nearest Neighbours

Which are the three closest neighbours?
The easiest distance measure is *Euclidean*.

$$d(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + ... + (p_n - q_n)^2} \quad (1)$$

For 2 dimensions (e.g., 2 features):

$$d(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2} \quad (2)$$

- use Euclidean distance to find the 3 closest example students to the student

Does the student pass or fail?

The new student is at point (4,5) i.e.: studied 4 hours and slept 5 hours. The three closest points are (4, 7) (distance of 2), (2,5) (distance of 2), and (7,5) (distance of 3).
Two out of 3 neighbours fails -> we predict that this students fails also

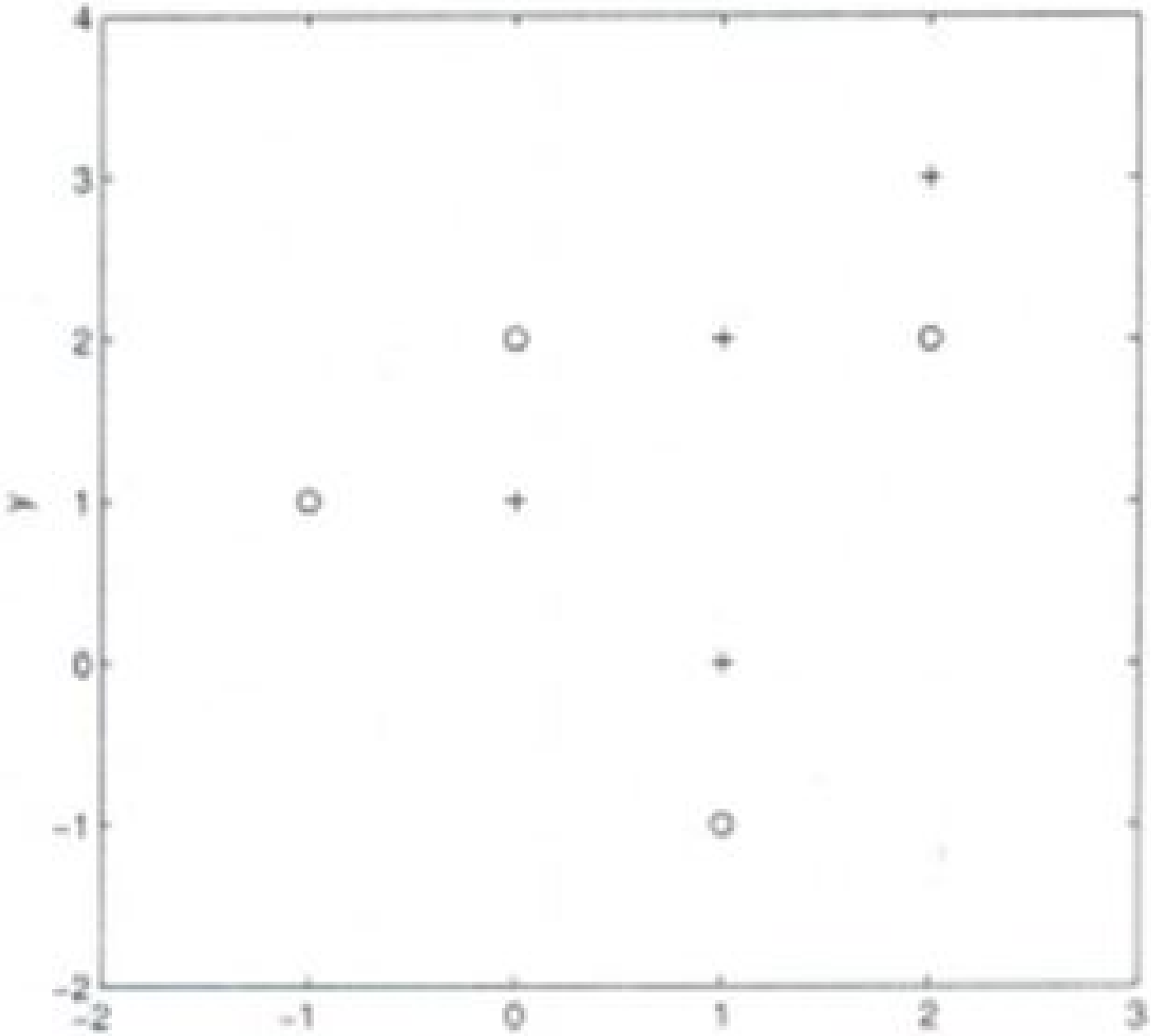
KNN pseudocode

```
Load the data
Initialise the value of k
Iterate from 1 to m (total number of training data points)
   Calculate the Euclidean distance between test datapoint
Sort the calculated distances in ascending order to get the
Count the most frequent classification of these rows
```

Quiz

1- When are the majority of computations done? When we train the model? Or when we test a datapoint?
2- True or false: The algorithm performs with the same efficiency, regardless how many features you have.
3- True or False: the computational complexity for classifying a new datapoint grows linearly with the size of the training set (i.e., it has O(N) complexity, where N is the sixe of the training set)
4- True or False: Euclidean distance is the only distance algorithm that you should use

Quiz



References

https://towardsdatascience.com/selecting-the-best-machine-learning-algorithm-for-your-regression-problem-20c330bad4ef
http://caisplusplus.usc.edu/blog/curriculum-supplement/knn
https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/
https://www.analyticsvidhya.com/blog/2017/09/30-questions-test-k-nearest-neighbors-algorithm/