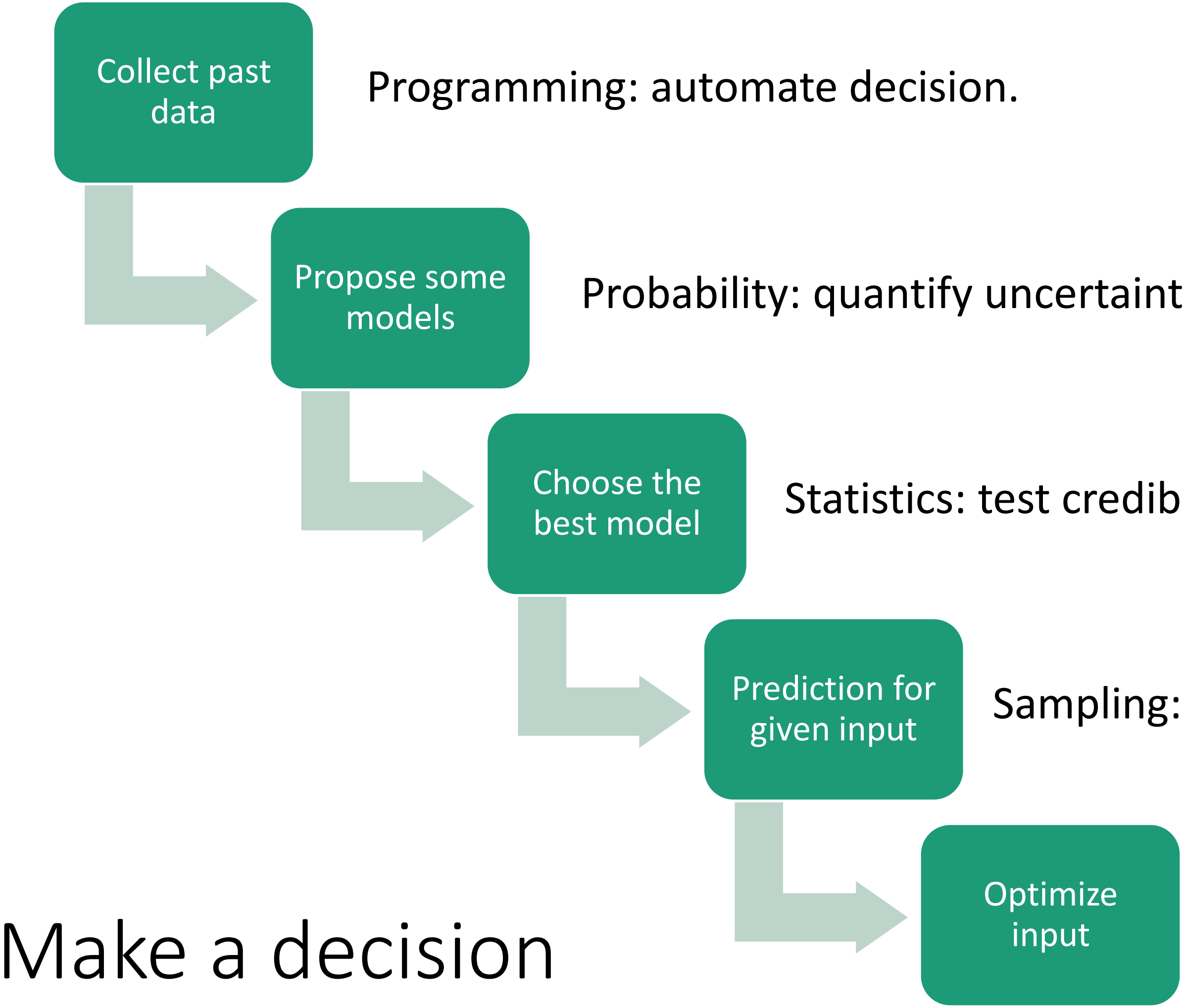
香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

Introduction to Data Science

Lecture 20 Machine Learning: Introduction & “Supervised” Learning

Zicheng Wang

Recap



Collect past data

Programming: automate decision.

Propose some models

Probability: quantify uncertainty. <- A probabilistic model

Choose the best model

Statistics: test credibility. <- e.g. MLE

Prediction for given input

Sampling: calculate complicated objective.

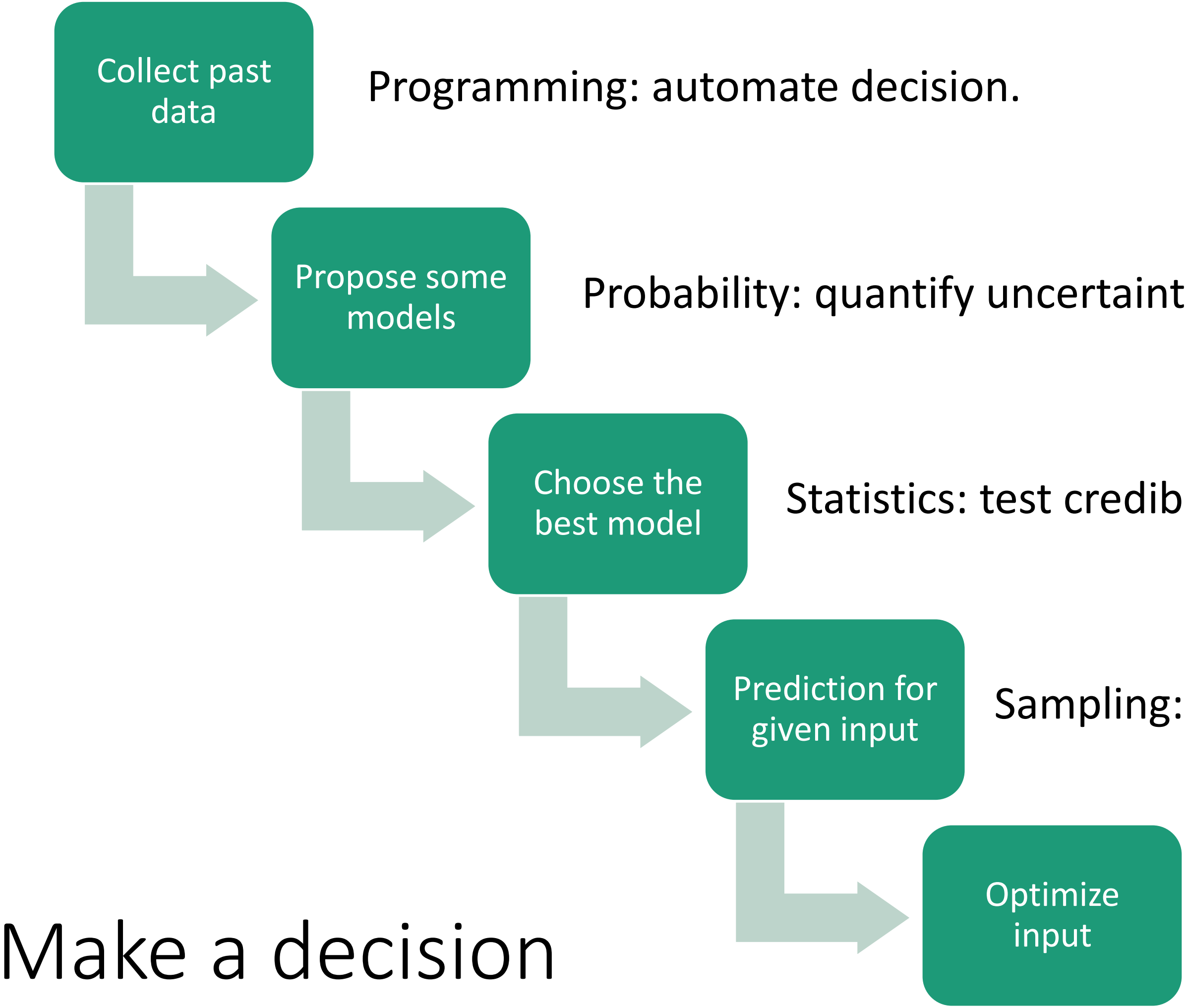
Optimize input

Convex Optimization

Optimization: optimize objectives.

Make a decision

Recap



Collect past data

Programming: automate decision.

No models to propose?

Propose some models

Probability: quantify uncertainty. <- A probabilistic model

Choose the best model

Statistics: test credibility. <- e.g. MLE

Prediction for given input

Sampling: calculate complicated objective.

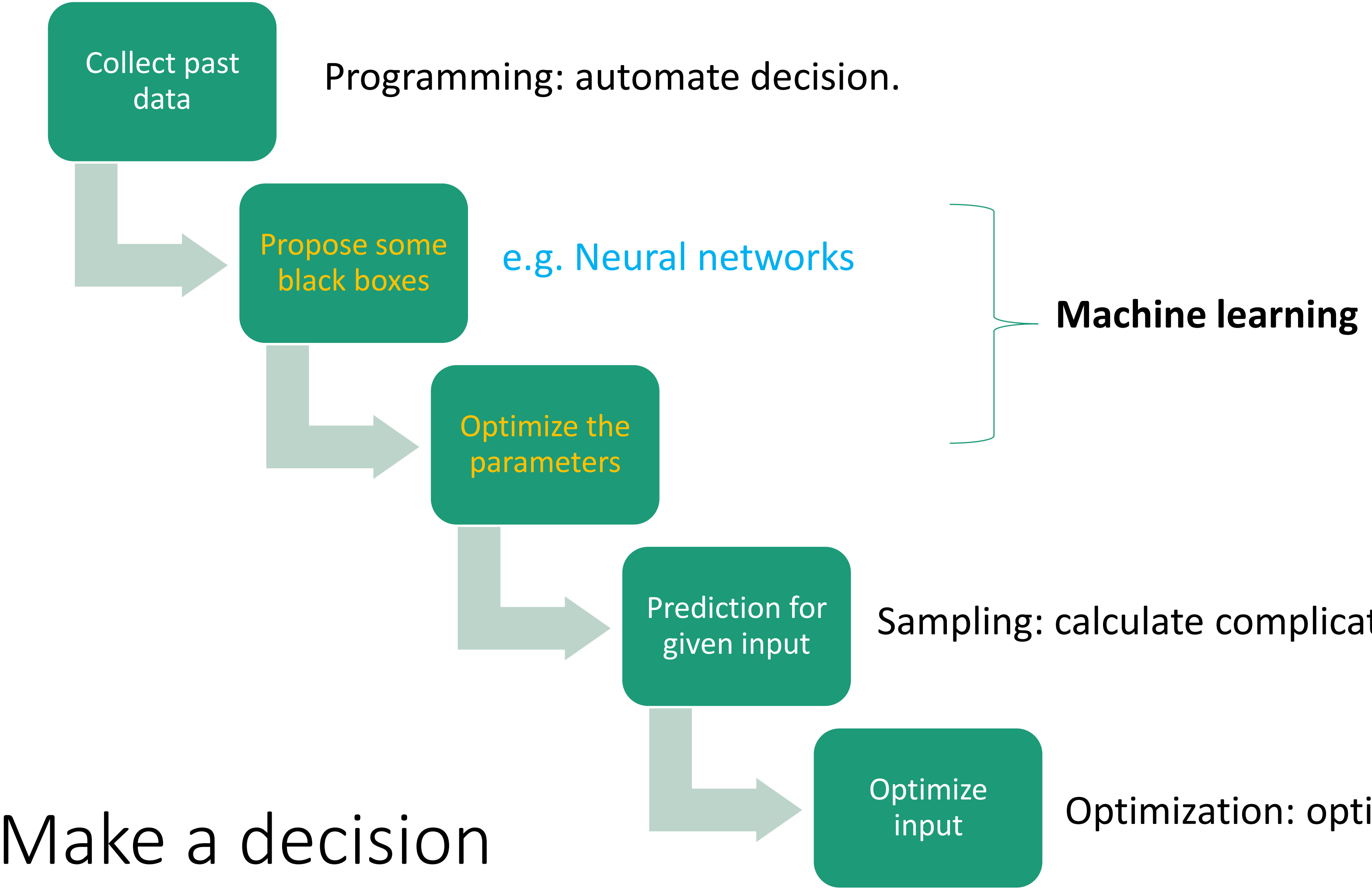
Optimize input

Convex Optimization

Optimization: optimize objectives.

Make a decision

Make a decision



Collect past data

Programming: automate decision.

Propose some black boxes

e.g. Neural networks

Optimize the parameters

Machine learning

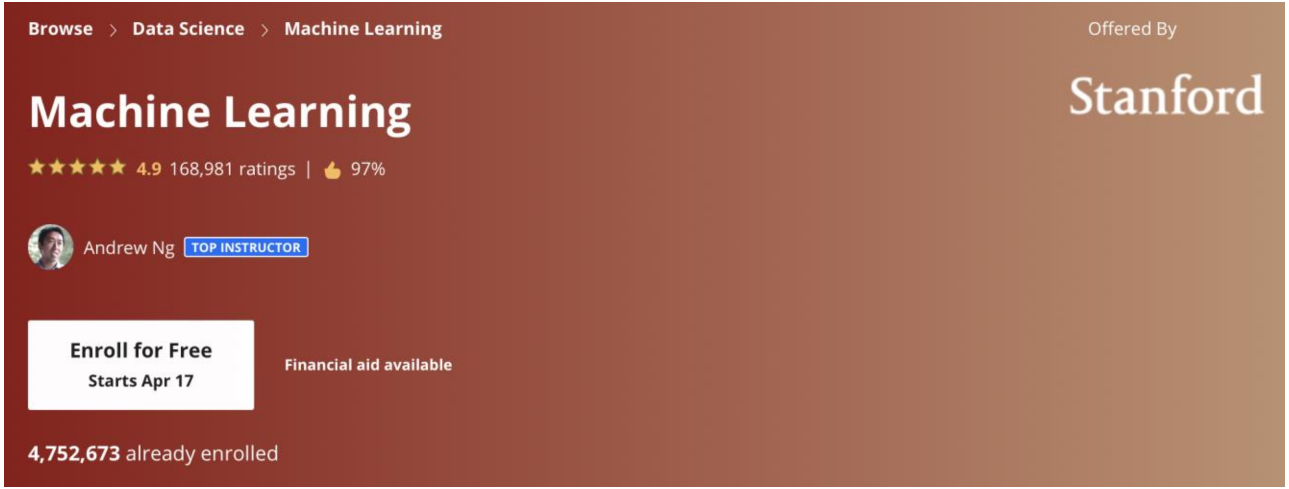
Prediction for given input

Sampling: calculate complicated objective.

Optimize input

Optimization: optimize objectives.

Resources



Browse > Data Science > Machine Learning

Offered By

Stanford

Machine Learning

★★★★★ 4.9 168,981 ratings | 👍 97%

Andrew Ng TOP INSTRUCTOR

Enroll for Free
Starts Apr 17

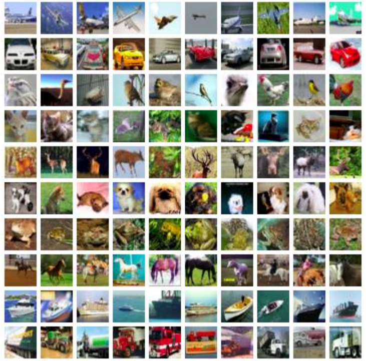
Financial aid available

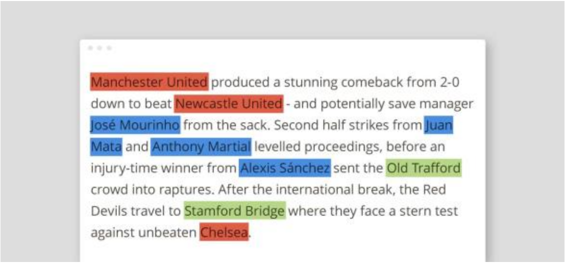
4,752,673 already enrolled

https://github.com/kjaisingh/high-school-guide-to-machine-learning

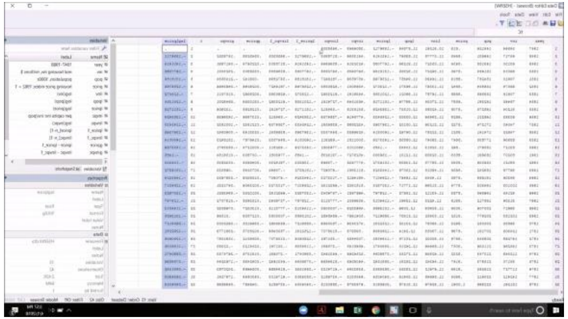
What is Machine Learning?

- Machine learning is to build systems that can **automatically** learn from historical data, identify patterns, and make logical decisions with little to **no human intervention.**
- Diverse forms of input data including numbers, words, clicks and images.





Manchester United produced a stunning comeback from 2-0 down to beat Newcastle United - and potentially save manager José Mourinho from the sack. Second half strikes from Juan Mata and Anthony Martial levelled proceedings, before an injury-time winner from Alexis Sánchez sent the Old Trafford crowd into raptures. After the international break, the Red Devils travel to Stamford Bridge where they face a stern test against unbeaten Chelsea.



What is Machine Learning?

The quality of a machine learning model hinges on two primary factors:

- The quality of the input data.
  - If the input data is of poor quality or disorganized, the model's output will likely be inaccurate.
- The model choice itself.
  - Each algorithm is designed for specific applications. It is vital to choose the appropriate algorithm for the given application.

Why is Machine Learning Important?

- Machine learning is growing in importance due to increasingly enormous volumes and variety of data, the access and affordability of computational power, and the availability of high speed Internet.

- It is possible for one to rapidly and automatically develop models that can quickly and accurately analyze extraordinarily large and complex data sets.

- Many applications: cut costs, mitigate risks, and improve overall quality of life including recommending products/services, detecting cybersecurity breaches, and enabling self-driving cars.

How Does Machine Learning Work?

Step 1: Choose and Prepare a Training Data Set

- Training data consists of representative samples that a machine learning application uses to tune its model parameters.

Step 2: Select and Apply an Algorithm to the Training Data Set

- The type of machine learning algorithm you choose will primarily depend on the nature of the problem the model seeks to solve

Step 3: Model Training and Parameter Tuning

- Training the model involves adjusting the model's variables and parameters to enhance its accuracy in prediction.
- Training model does not require human intervention, showcasing the power of machine learning. The machine learns from the data, needing minimal to no guidance from the user.

Step 4: Deployment and Model Improvement

- Now you can deploy the mode for actual use and improve its effectiveness and accuracy over time with new data.

Dog/Cat Classification

Is it a cat or a Dog?

cat

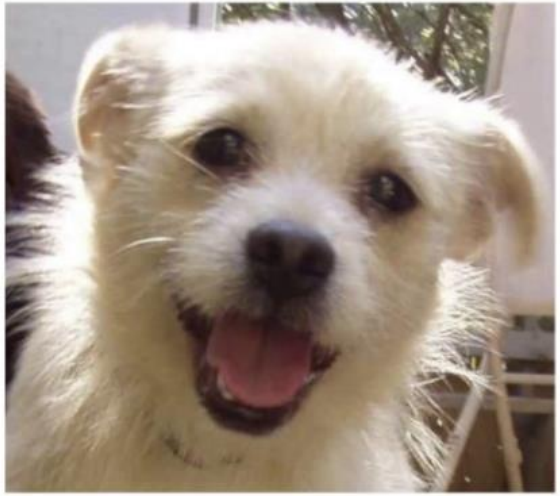
dog

Is it a cat or a Dog?

[✓ cat]

dog

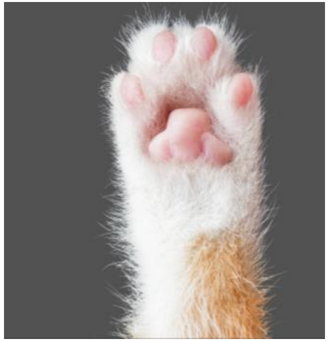
Is it a cat or a Dog?



cat

[✓ dog]

How about more challenging cases?



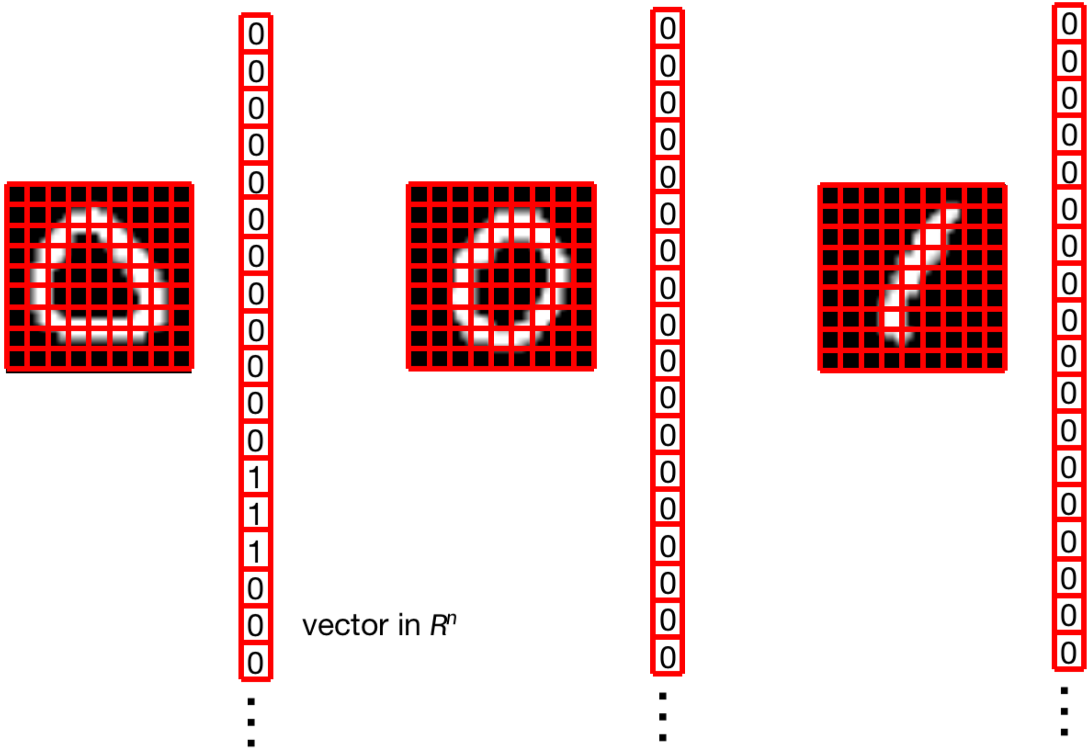
cat

?

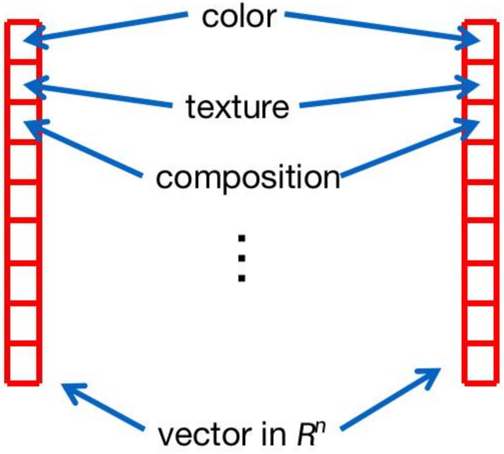
dog

Represent Objects Numerically

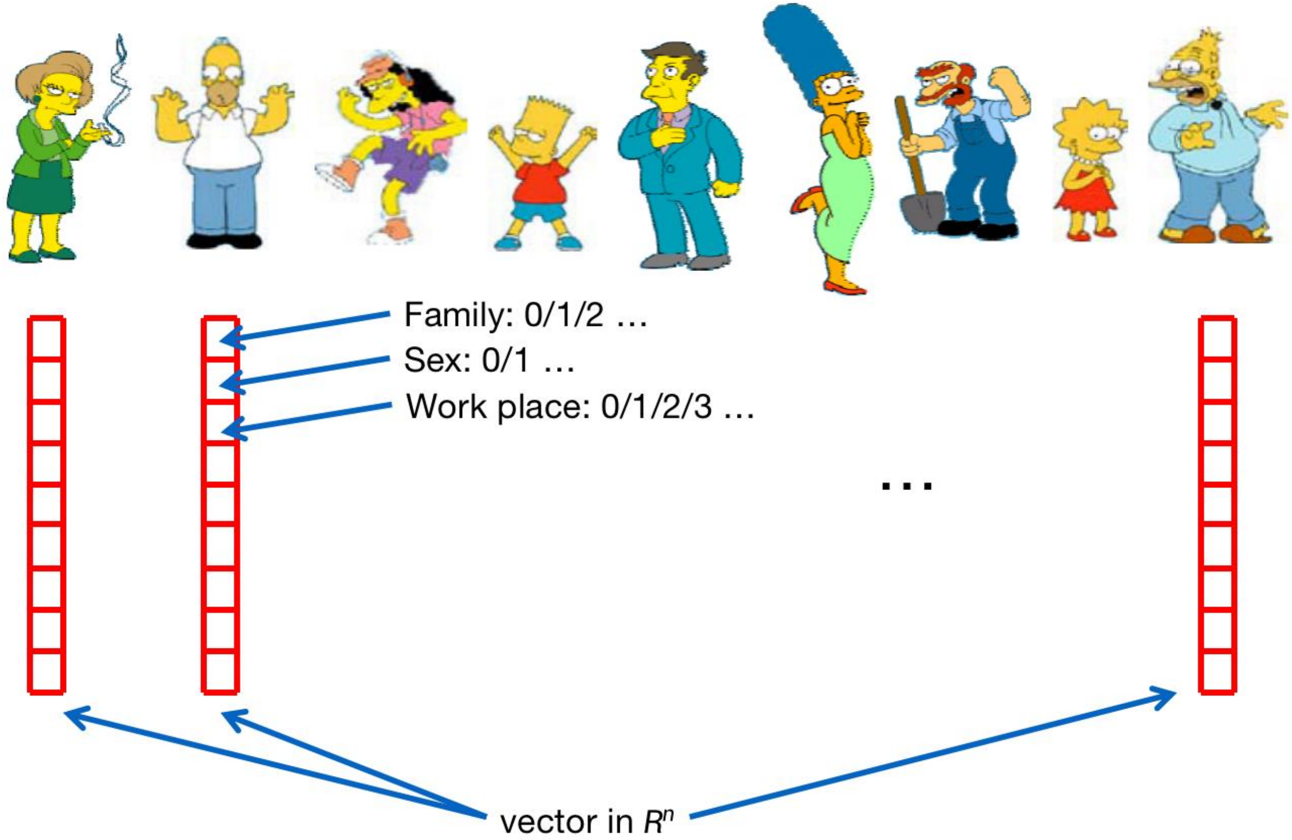
How to represent objects numerically?



Images of different sizes



Objects in real life



Family: 0/1/2 …

Sex: 0/1 …

Work place: 0/1/2/3 …

…

vector in R^n

Supervised Learning

Supervised Learning

- Supervised machine learning algorithms utilize labeled data for training, where the correct outputs corresponding to input data are already known.

- For all samples, $(x^i, y^i), i = 1, \ldots N$, you can observe both the input data x^i and the label y^i

Training data

y=1 (cat)

y=0 (dog)

y=1 (cat)

••• •••

y=0 (dog)

Supervised Learning

Training data

y=1 (cat)

y=0 (dog)

y=1 (cat)

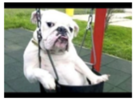
…… …

y=0 (dog)

⇩ Learning algorithm (optimization involved)

Classifier $h: X \rightarrow \{0,1\}$

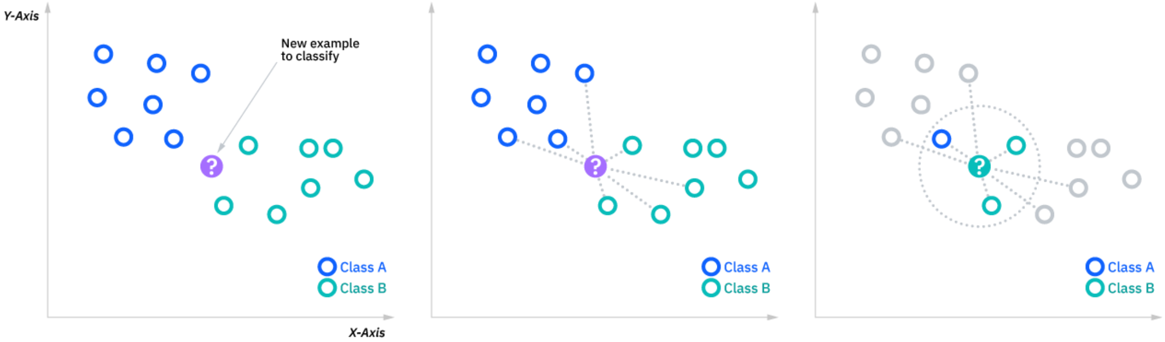
For example:

$h\Bigg($  $\Bigg)$ ⟶ 0

Supervised Learning Algorithm 1: KNN

K-Nearest Neighbor Classifier

- Find K training points x_i closest to x.
- If the majority of K-nearest neighbors of x belong to classifier c, label x as c.



The KNN Algorithm

1. Load the data
2. Set K of your choice to be the number of neighbors
3. For each new data to be classified
    - Calculate the distances between the new data and all the labeled data.
    - Record the entry (d_i, y_i), where d_i is the distance between the new data and the ith labeled data, and y_i is the label of the ith data.
    - Sort the these entries with respect to distance (from smallest to largest).
5. Pick the first K entries from the sorted collection
6. Get the labels of the selected K entries
7. Choose the label with the largest frequency

Example 1

Sweet tofu pudding or salted tofu pudding?

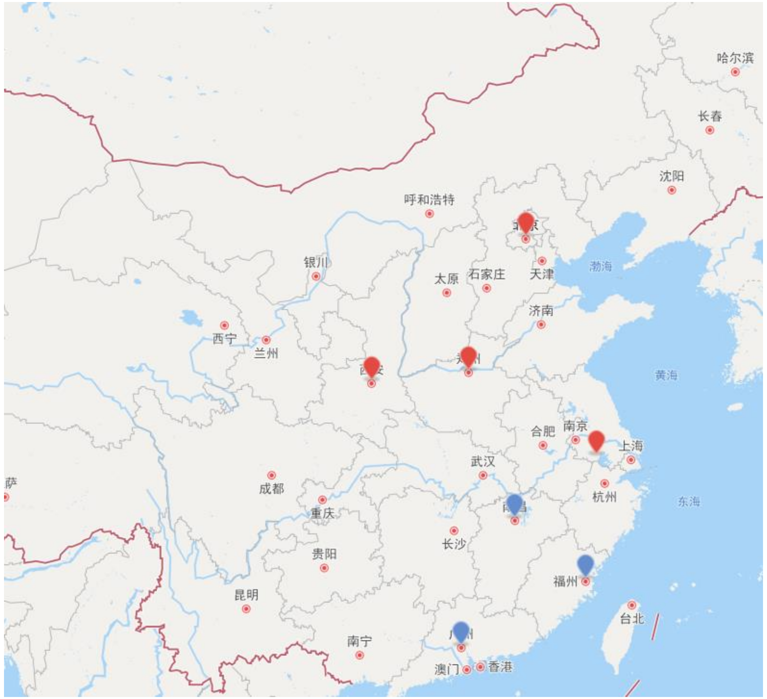
Example 1

Data:

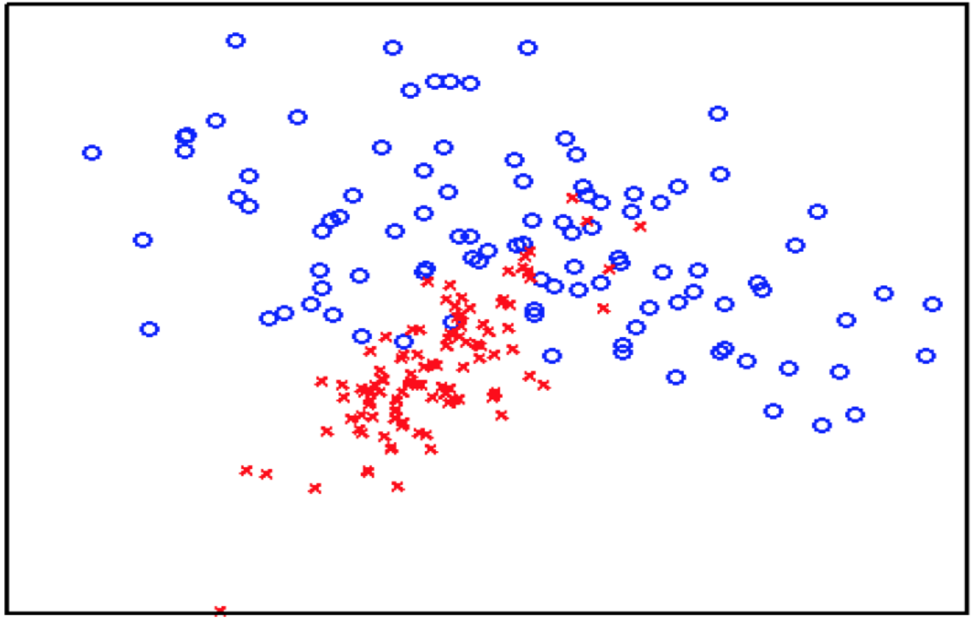
Your hometown location

Label:

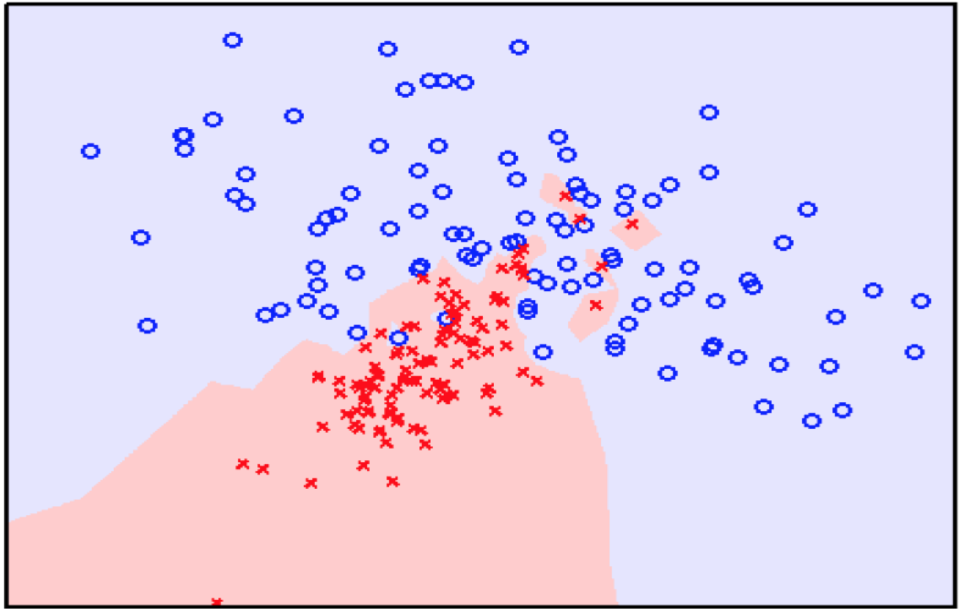
Red (Salted)/Blue (Sweet)



Example 2



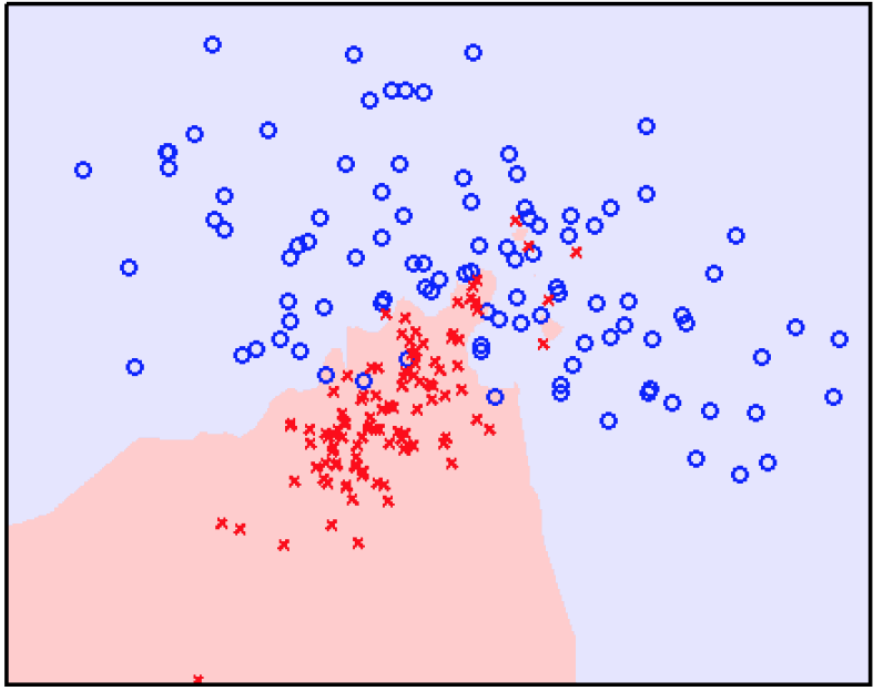
Example 2



K = 1

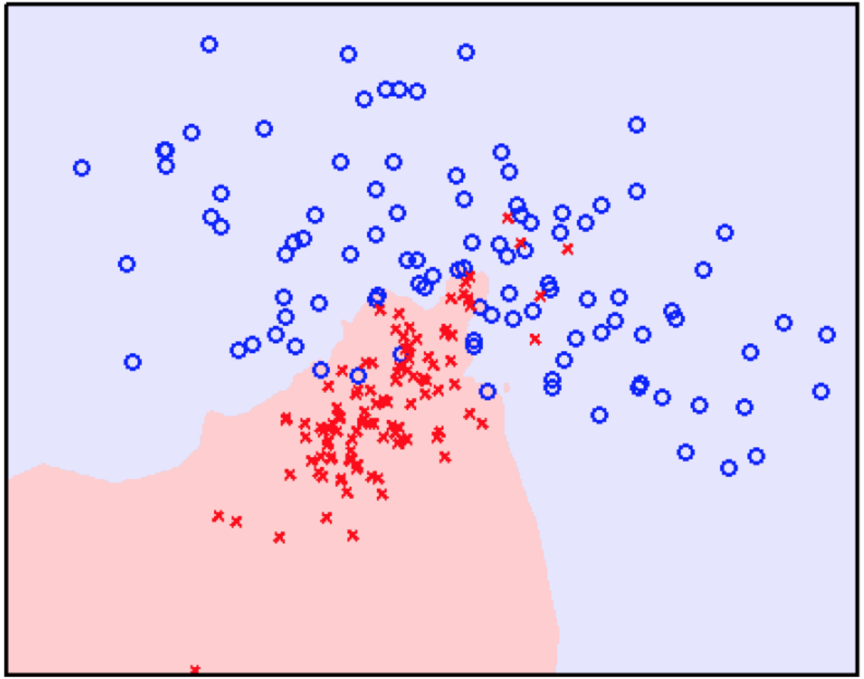
The red/blue region indicates the area where any new data falling within will be classified as ‘red/blue' by the algorithm.

Example 2



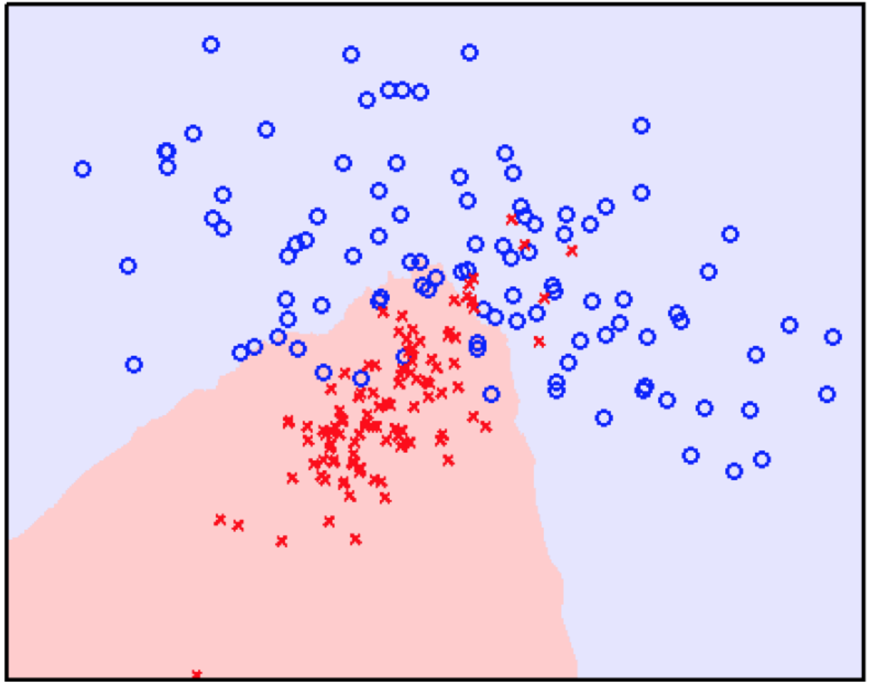
K = 3

Example 2



K = 5

Example 2



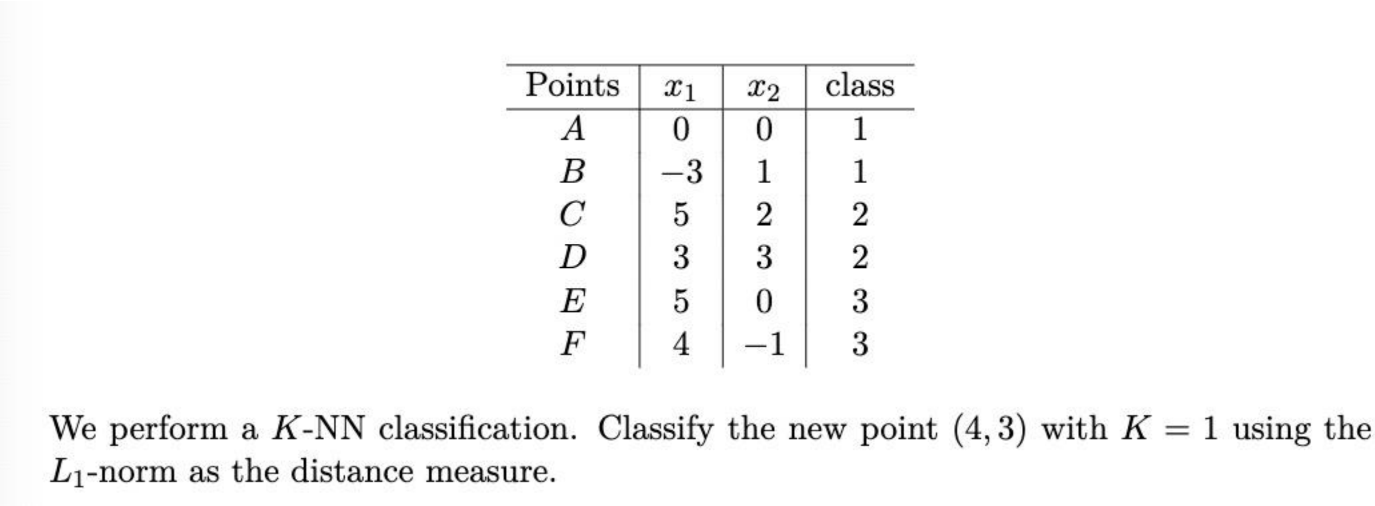
K = *25*

- The selection of K is critical—lower values of K may significantly increase the influence of noise on the outcomes.

- A large value of K may undermine the fundamental principles underlying the K-Nearest Neighbors algorithm.
  - If K>total number of labeled data, the label of any new input will be the labels that appears the most in the samples.

Exercise

- We are given the following data set with points of three different classes:



Points	x_1	x_2	class
A	0	0	1
B	-3	1	1
C	5	2	2
D	3	3	2
E	5	0	3
F	4	-1	3

We perform a K-NN classification. Classify the new point $(4, 3)$ with $K = 1$ using the L_1-norm as the distance measure.

Manhattan distance between x and y: $|x_1 - y_1| + |x_2 - y_2|$