

K-nearest neighbors classifier (KNN)

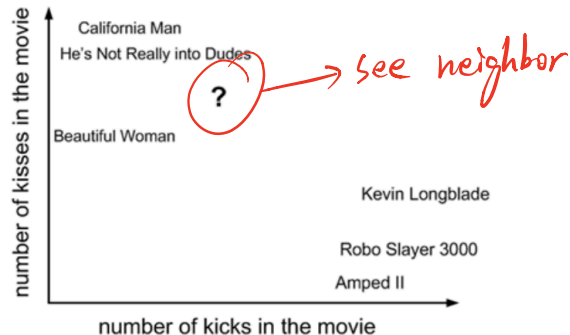
Hung-Hsuan Chen 陳弘軒
Computer Science and Information Engineering
National Central University
hhchen@ncu.edu.tw

Slides adapted from Xiaoli Fern (Oregon State), Rong Jin (MSU)

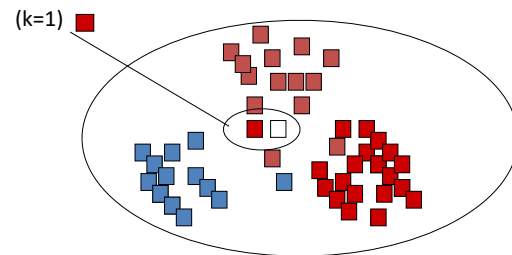
An example of movie type classification

| Movie title | # of kicks | # of kisses | Type of movie |
|-----------------------------------|------------|-------------|---------------|
| <i>California Man</i> | 3 | 104 | Romance |
| <i>He's Not Really into Dudes</i> | 2 | 100 | Romance |
| <i>Beautiful Woman</i> | 1 | 81 | Romance |
| <i>Kevin Longblade</i> | 101 | 10 | Action |
| <i>Robo Slayer 3000</i> | 99 | 5 | Action |
| <i>Amped II</i> | 98 | 2 | Action |
| ? | 18 | 90 | Unknown |

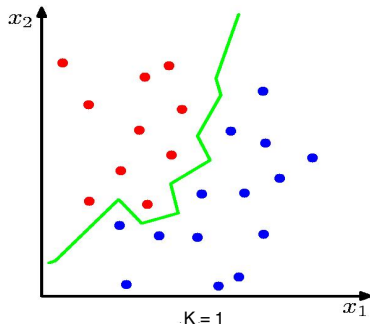
Visualize the data



K-Nearest-Neighbor (k NN) Classifier



K Nearest Neighbour (kNN) Classifier



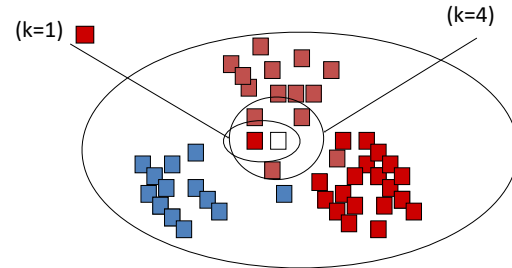
K Nearest Neighbour (kNN) Classifier

K is a **hyperparameter** of the model

need human to decide

v.s. parameter decide by

machine in ML

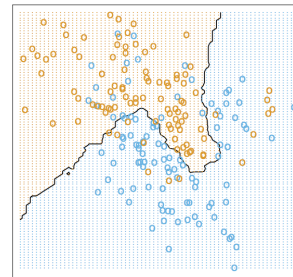


KNN classifier

Given a query point x_0 , find the k nearest training instances to x_0 , and then make prediction based on majority vote among these k neighbors

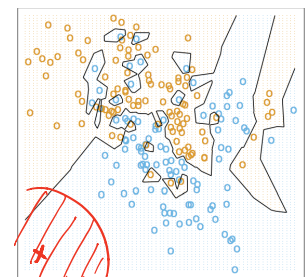
Feature normalization is usually performed as a preprocessing step

How to select k ?



15-nearest neighbors

underfit

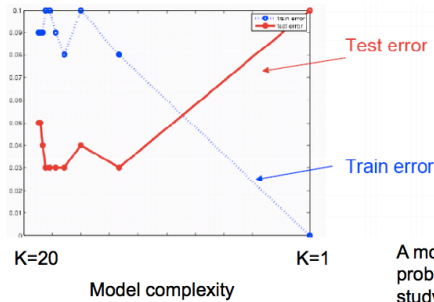


1-nearest neighbors

overfit

*itself → training error 0
k=1*

Overfitting vs underfitting of KNN



A model selection problem that we will study later

Discussion

- How to select k ?
- What if k is an even number?
- What if k equals 1?
- What if k equals the number of the training instances?
- How fast for model training/testing? n

Discussion (cont')

- How to accelerate testing time?
 - E.g., by designing a special data structure and algorithm? By approximation?
 - Although this may increase the training time
 - This could be a research project!
- How to measure "distance"?
 - Euclidean distance, Cosine distance, etc.
- Can we apply knn when some features are categorical?

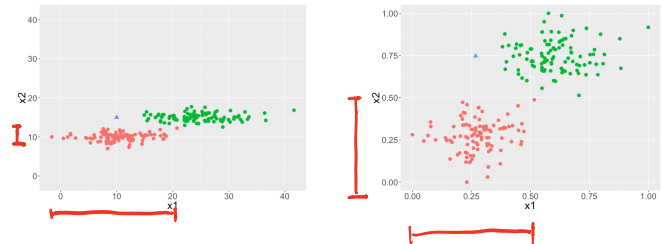
$A = \{a, b\}$
 $B = \{a, c\}$
 $C = \{d, e\}$

yes,

↓
 depend on if you can define
 distance between
 category

Discussion (cont')

- Why feature normalization?




Popular feature normalization methods

- Normalized to $N(0, 1)$

- Scale to $[0, 1]$

- Scale to $[-1, 1]$


$$\frac{x_i - \mu}{\sigma}$$

Quiz

- Training data: (X =height, y =gender)

- Given a new student's height, explain how to predict the student's gender by kNN ($k=3$)?

- Explain overfitting and underfitting

- When using KNN with a very small k , will the model tend to overfit or underfit?

→ find whose height nearest students