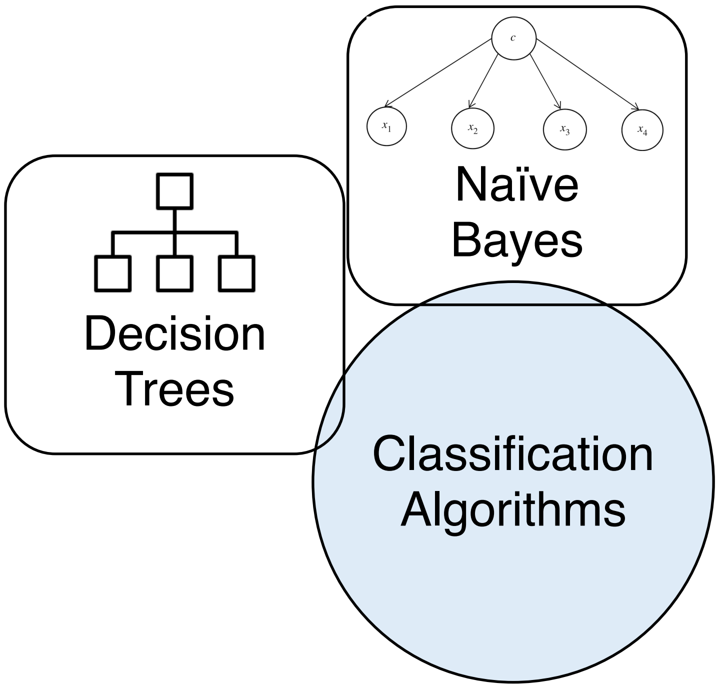
机器学习

第五讲: K-近邻分类

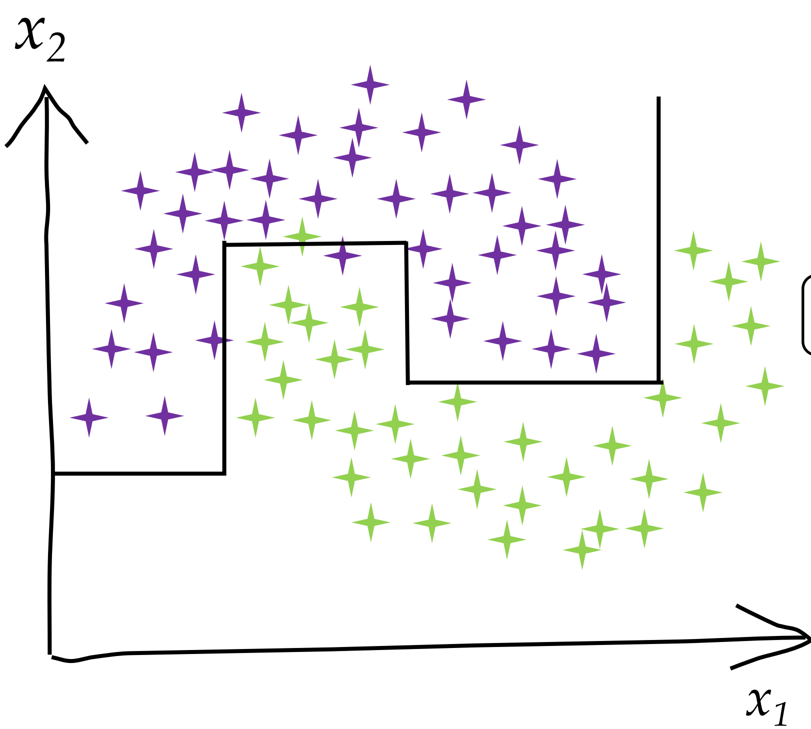
顾小东
上海交通大学计算机学院

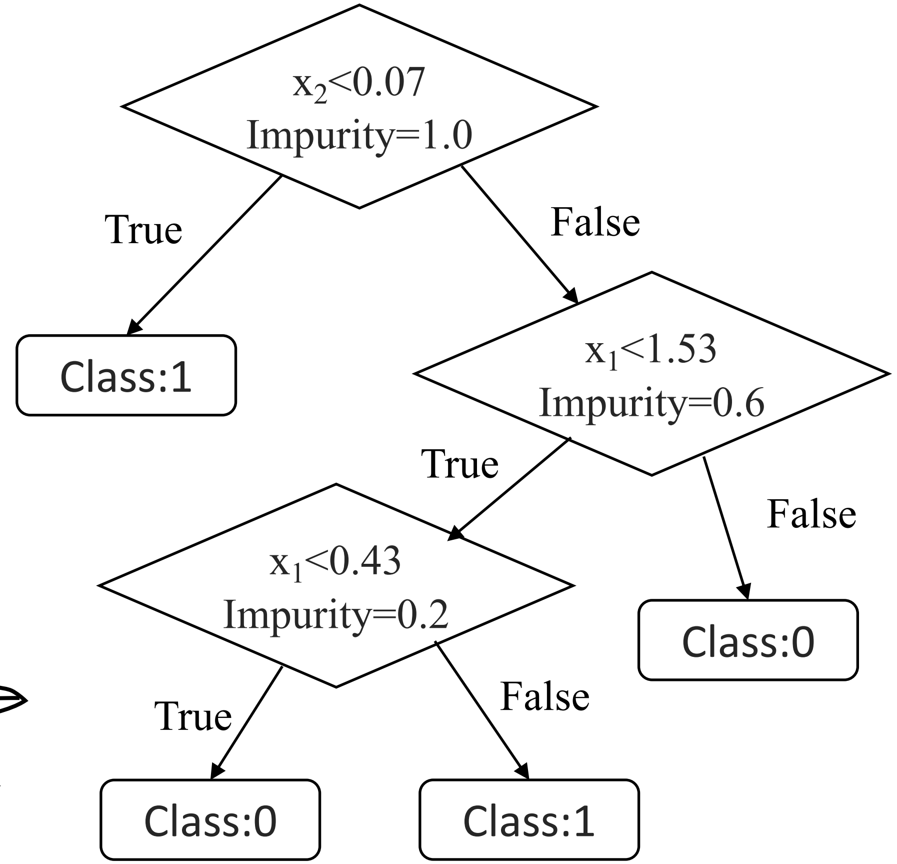
分类器家族



回顾: 决策树

根据数据的不纯净度（Impurity）作分类决策

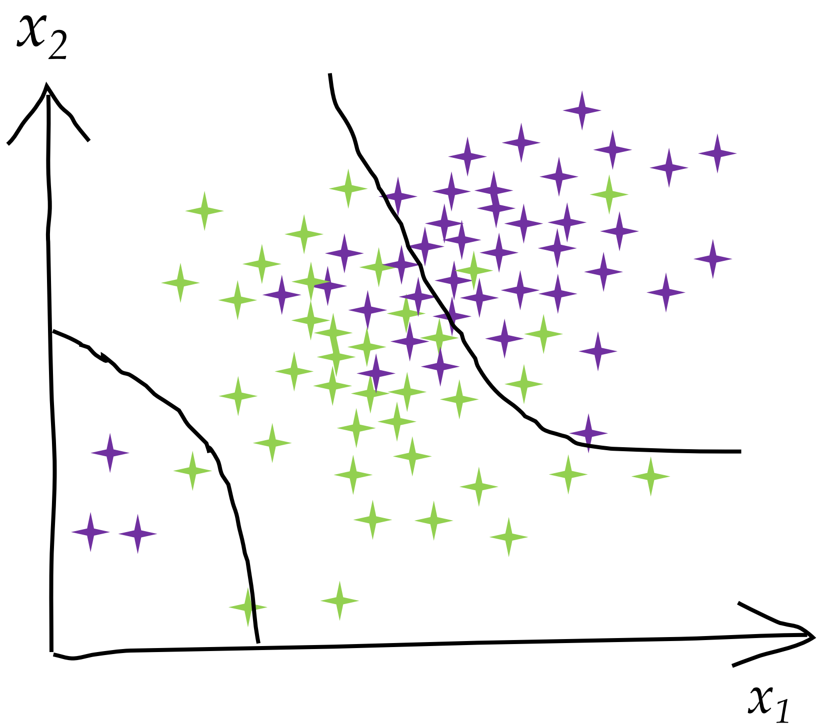


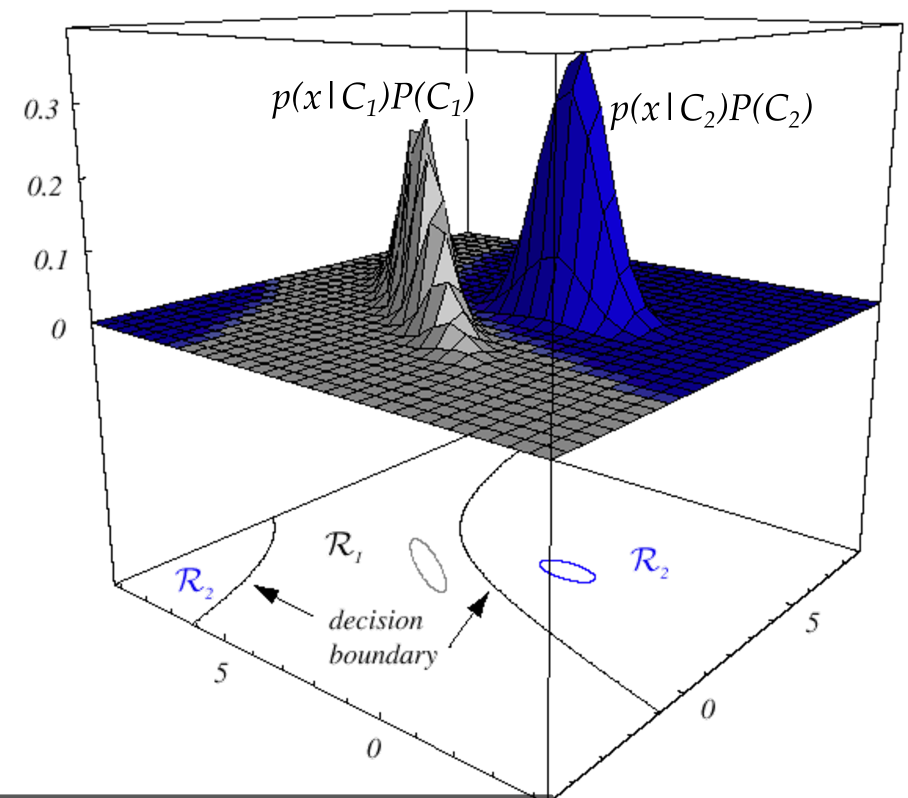


回顾: 贝叶斯分类

根据数据的概率密度作分类决策

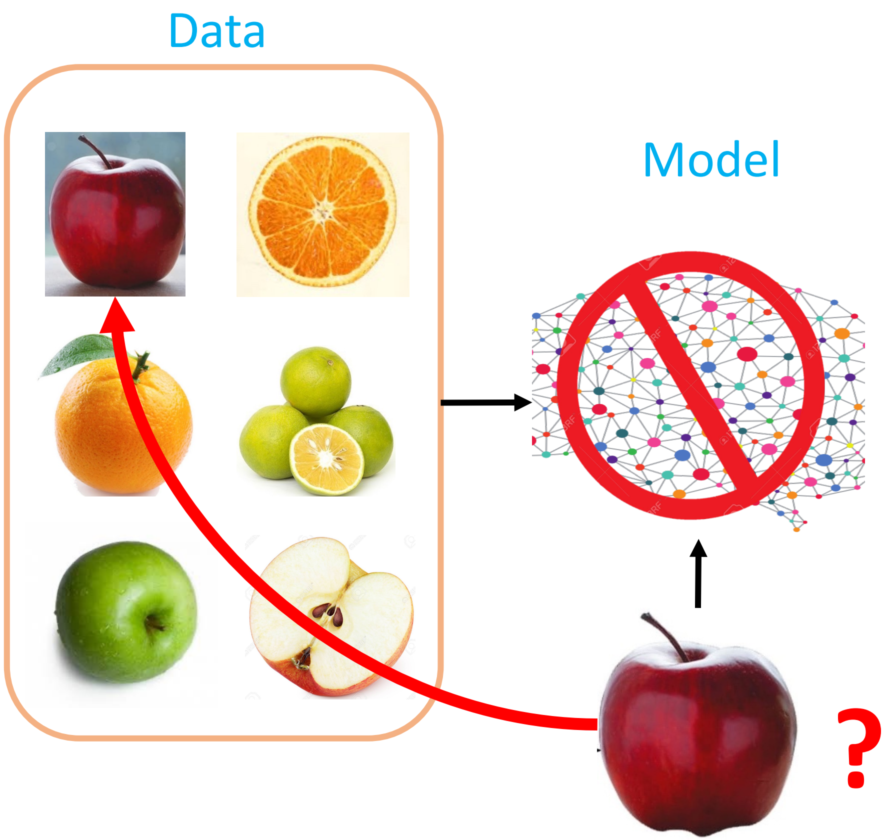
$$P(C_i|x) = \frac{p(x|C_i)P(C_i)}{p(x)}, \quad i=1,2$$

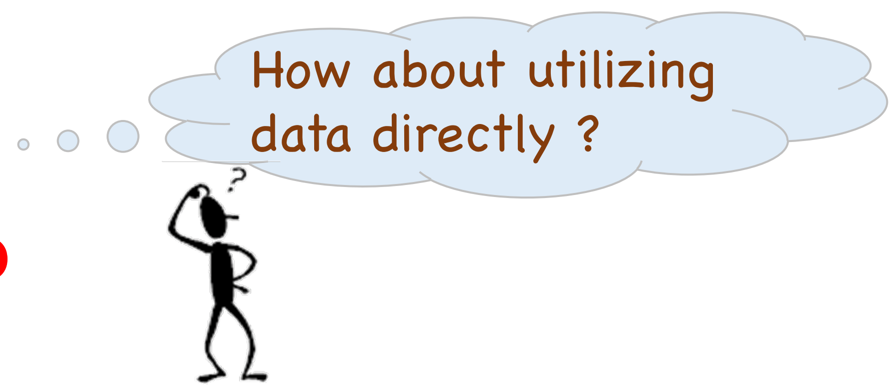




决策规则: Classify x into C_i if $p(C_i|x) > p(C_j|x)$ else C_j.

Do we really need a model ?

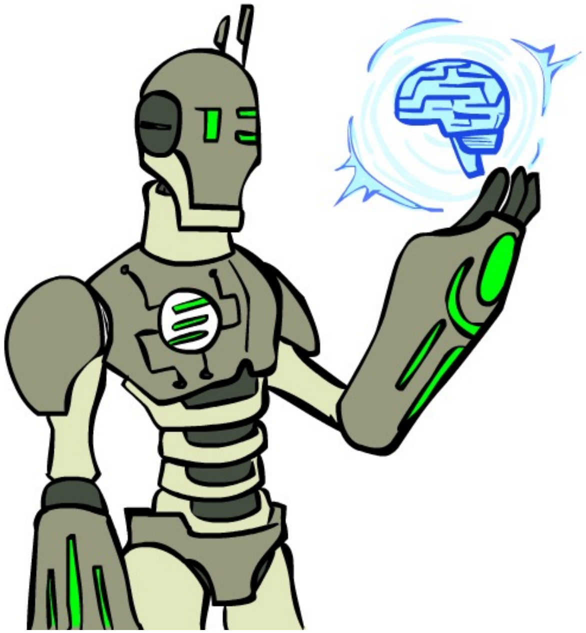




Today

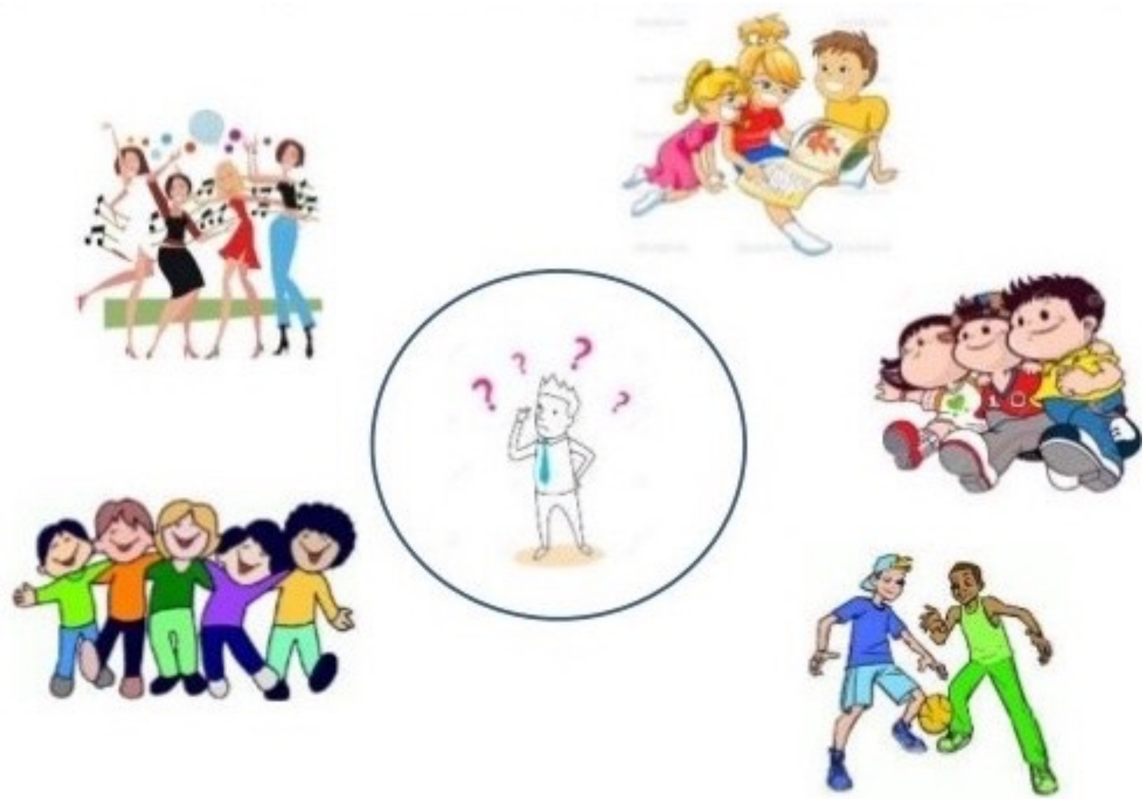
Let's learn a simple (naïve) method for classification.

- K-Nearest Neighbor



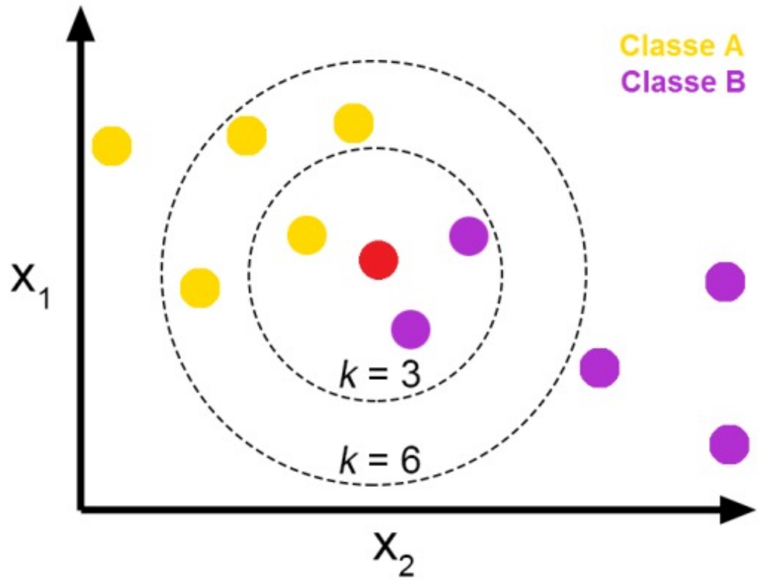
A Simple Analogy… 物以类聚人以群分

- Tell me about your friends (who your neighbors are). Then I will tell you who you are.



K-NN Classification

- Classified by "MAJORITY VOTES" from neighbor classes.
- An object is classified to the most common class amongst its k nearest neighbors **measured by a "distance" function**



How to determine whether an object falls into the k-nearest neighbors?

Distance Measures

- **Euclidean Distance (欧氏距离)**

$$D(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{i=1}^{\mathrm{n}} (x_i - y_i)^2}$$

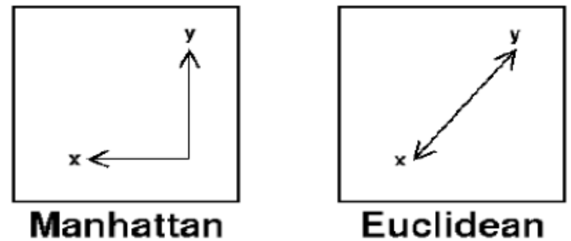
where $\boldsymbol{x} = (x_1, x_2, \ldots, x_\mathrm{n})$ and $\boldsymbol{y} = (y_1, y_2, \ldots, y_\mathrm{n})$ represent the n attribute values of two records.

doesn't work well in high dimensions and categorical variables because it ignores the similarity between attributes.

Distance Measures

- **Manhattan Distance** (a.k.a. city block distance)

$$D(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$



$$|x_1 - x_2| + |y_1 - y_2|$$

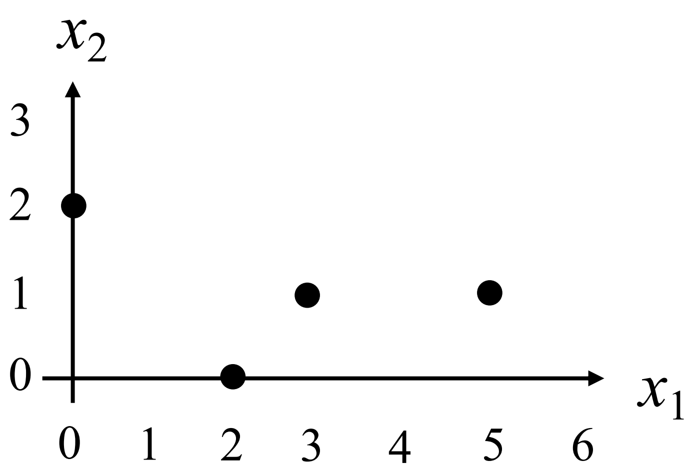
- **Minkowski Distance(闵氏距离)**

$$D(x, y) = \left(\sum_{u=1}^{n} |x_u - y_u|^p \right)^{\frac{1}{p}}$$

p=1 Manhattan distance
P=2 Euclidean distance

Distances

Example



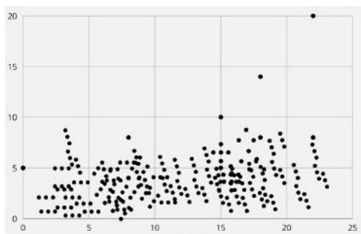
point	x_1	x_2
p1	0	2
p2	2	0
p3	3	1
p4	5	1

Euclidean Distance Matrix

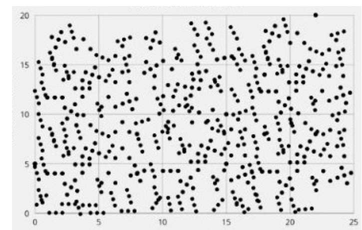
	p1	**p2**	**p3**	**p4**
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Normalization (归一化)

- Standardize the range of attributes (features of data)



ID	Grade	Age	Score
1	3	22	92
2	4	23	87
3	2	21	82



ID	Grade	Age	Score
1	0.5	0.5	1
2	1	1	0.5
3	0	0	0

Z-score normalization: rescale the data so that the mean is 0 and the standard deviation is 1.

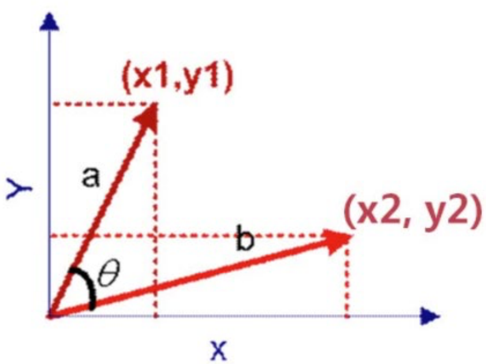
$$x_{norm} = \frac{x - \mu}{\sigma}$$

Min-max normalization: scale the data to a fixed range of [0, 1].

$$x_{morm} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Similarity vs. Distance

- **Similarity:** numerical measure of how alike two data objects are.
  - Is higher when objects are more alike.
  - Often falls in the range [0, 1].
- **Cosine Similarity**



$$\cos(\theta) = \frac{a \bullet b}{||a|| \times ||b||}$$

$$= \frac{(x_1, y_1) \bullet (x_2, y_2)}{\sqrt{x_1^2 + y_1^2} \times \sqrt{x_2^2 + y_2^2}}$$

$$= \frac{x_1 x_2 + y_1 y_2}{\sqrt{x_1^2 + y_1^2} \times \sqrt{x_2^2 + y_2^2}}$$

$$\cos(d_1, d_2) = \begin{cases} 1: \text{exactly the same} \\ 0: \text{orthogonal} \\ -1: \text{exactly opposite} \end{cases}$$

Example

$d_1 = [3\ 2\ 0\ 5\ 0\ 0\ 0\ 2\ 0\ 0]$

$d_2 = [1\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 2]$

$d_1 \cdot d_2 = 3\times1+2\times0+0\times0+...+0\times2 = 5$

$||d_1|| = (3^2+2^2+...+0^2)^{0.5} = 6.481$

$||d_2|| = (1^2+0^2+...+2^2)^{0.5} = 2.245$

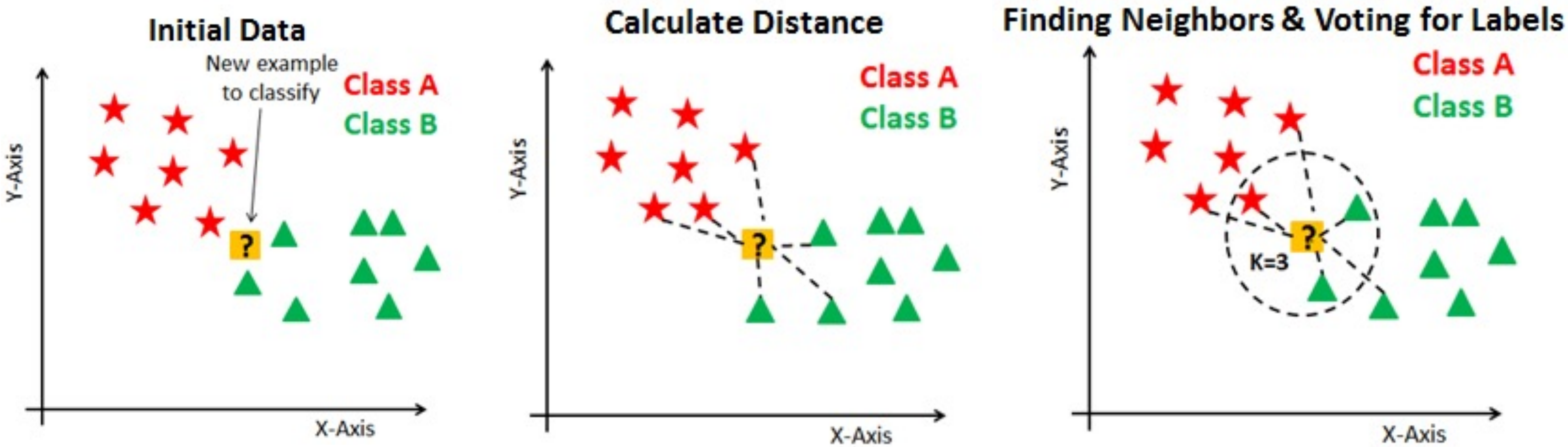
$\cos(d_1,d_2) = (d_1 \cdot d_2)/||d_1||\ ||d_2|| = 0.3150$

The Algorithm

Algorithm

1. Determine parameter K
2. Choose a sample from the test data that needs to be classified and compute its distance to all the training examples.
3. Sort the distances obtained and take the k-nearest data samples.
4. Assign the test class to the class based on the majority vote of its k neighbors.

The Algorithm



KNN – Example

Training set

ID	width	height	Type
1	5.2	8.0	lemons
2	6.7	9.8	lemons
3	7.2	7.5	orang
4	6.1	5.3	orang
5	4.1	6.5	lemons

Test instance

6	6.2	9.7	?

=lemon

Step1 – calculate distances

	6
1	5.099
2	2.828
3	5.623
4	3.162
5	7.892

sort

Step2 – rank the neighbors.

	6	
2	2.828	lemon
4	3.162	orange
1	5.099	lemon
3	5.623	
5	7.892	

Step4 – voting.

Step3 – select the nearest neighbors.

What is the best value of K to use?

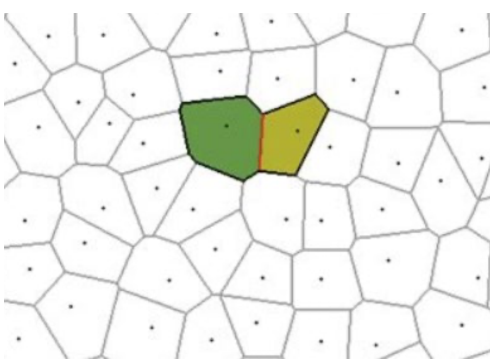
Decision Boundary

Voronoi Tessellation (维诺图，沃罗诺伊分割)

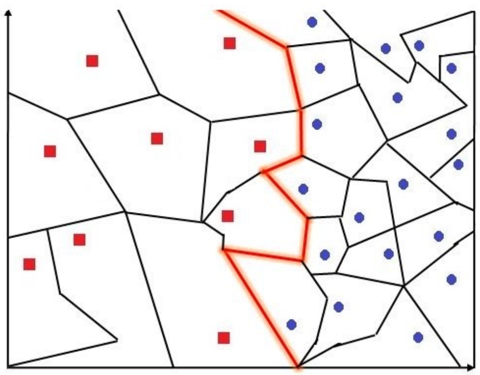
- Partition the space into areas that are nearest to any given point
- Boundary: points at the same distance from two different training examples.

Decision Boundary: boundaries that separates two different classes.

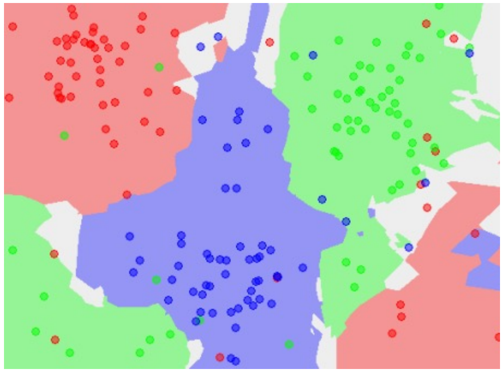
K=1



K=1



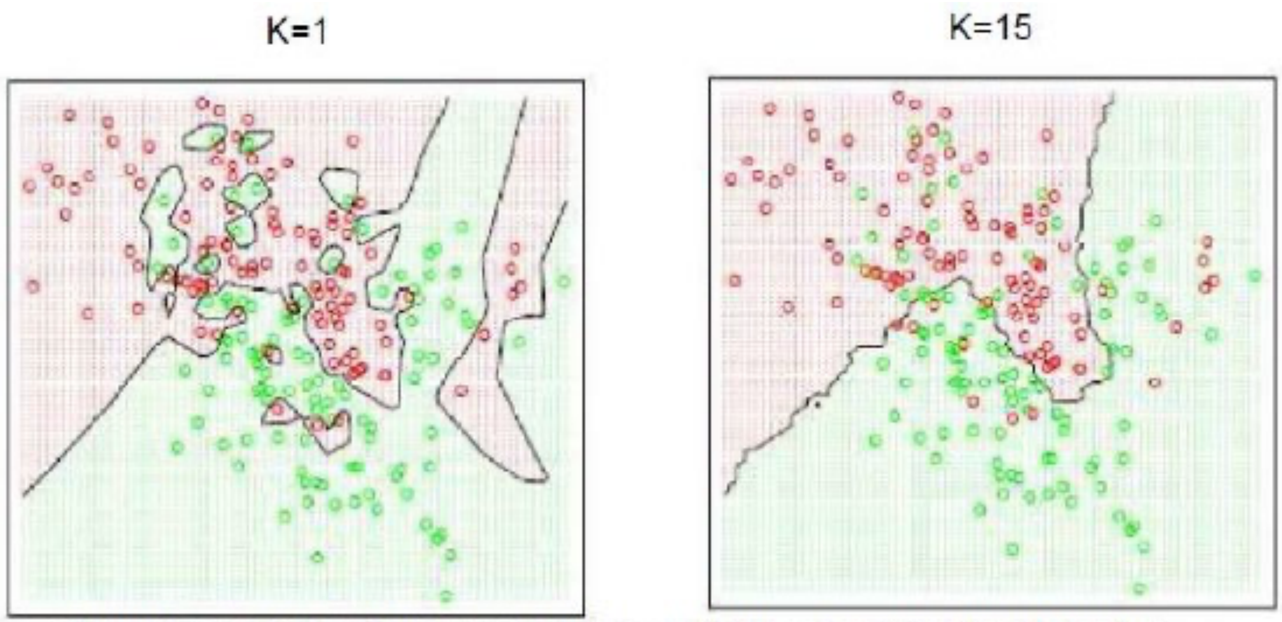
K>1



With large number of examples and possible noise in the labels, the decision boundary can become nasty!

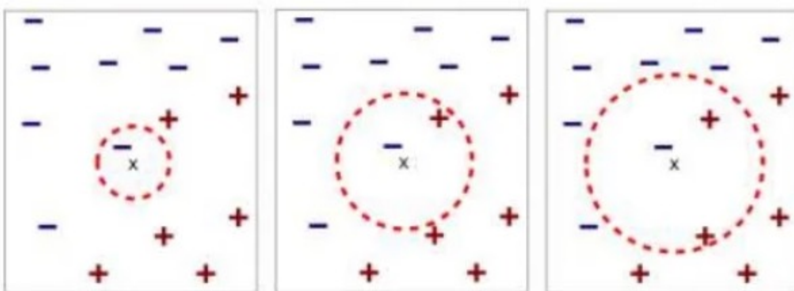
Effect of K

- Larger K produces smoother boundary effect
- When K==N, always predict the majority class



Figures from Hastie, Tibshirani and Friedman (Elements of Statistical Learning)

Discussion: which model is better between K=1 and K=15? Why?

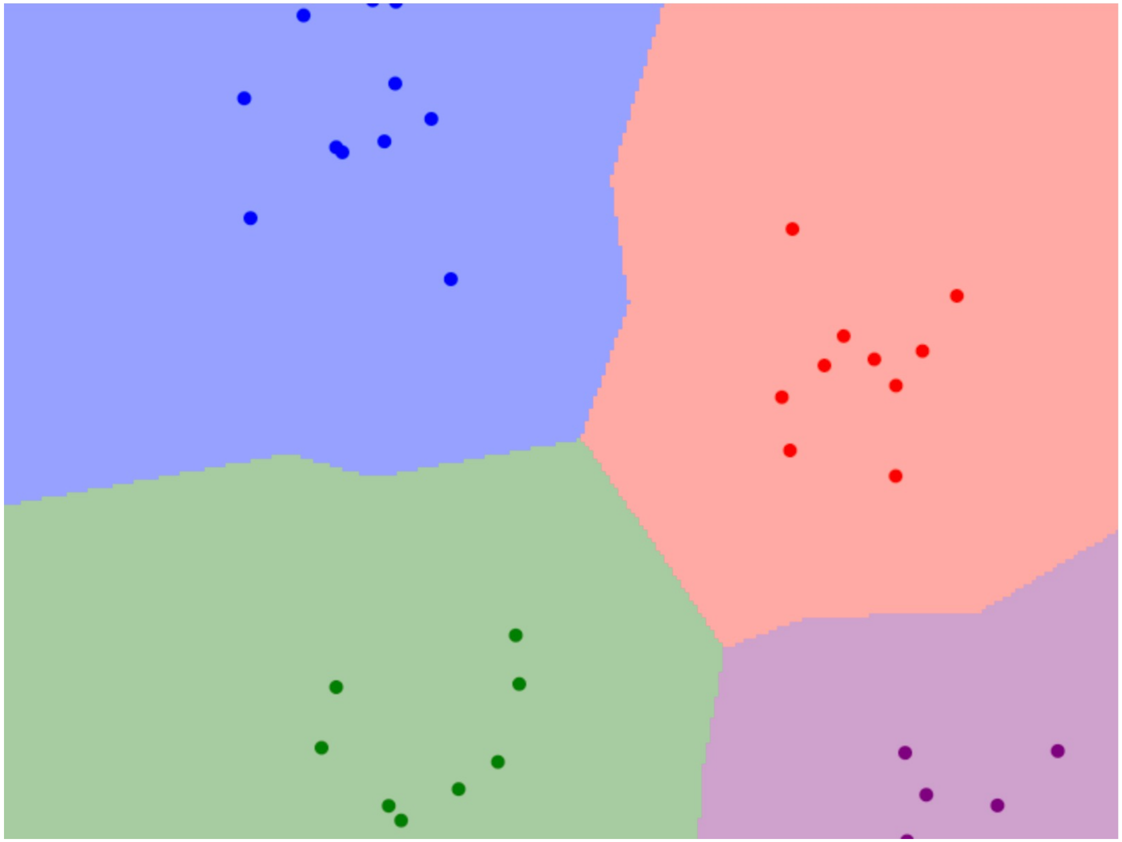


(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

How to choose K?

- If K is too small, efficiency is increased but becomes susceptible to noise.
- Larger K works well. But too large K may include majority points from other classes, but risk of over-smoothing classification results

Try it yourself



http://vision.stanford.edu/teaching/cs231n-demos/knn/

Pros and Cons

Advantages:

- Simple to understand, explain, and implement,
- No effort for training,
- New data can be added seamlessly without hampering the model accuracy

Pros and Cons

Disadvantages:

- Does not scale with large data sets (calculating distance is computationally expensive)
- Highly susceptible to the curse of dimensionality
- Large storage requirements
- Data normalization is required

Tutorial: KNN with Python

https://www.kaggle.com/code/lohitha17/knn-without-scikit-learn

IT'S TIME

FOR

coding