ZIXING AI
自兴人工智能

k-近邻算法
(k Nearest Neighbors, kNN)

陈白帆

ZIXING AI
自兴人工智能

参考教材

- Peter Harrington著，李锐、李鹏、曲亚东、王斌译. 机器学习实战，人民邮电出版社，2013

ZIXING AI
自兴人工智能

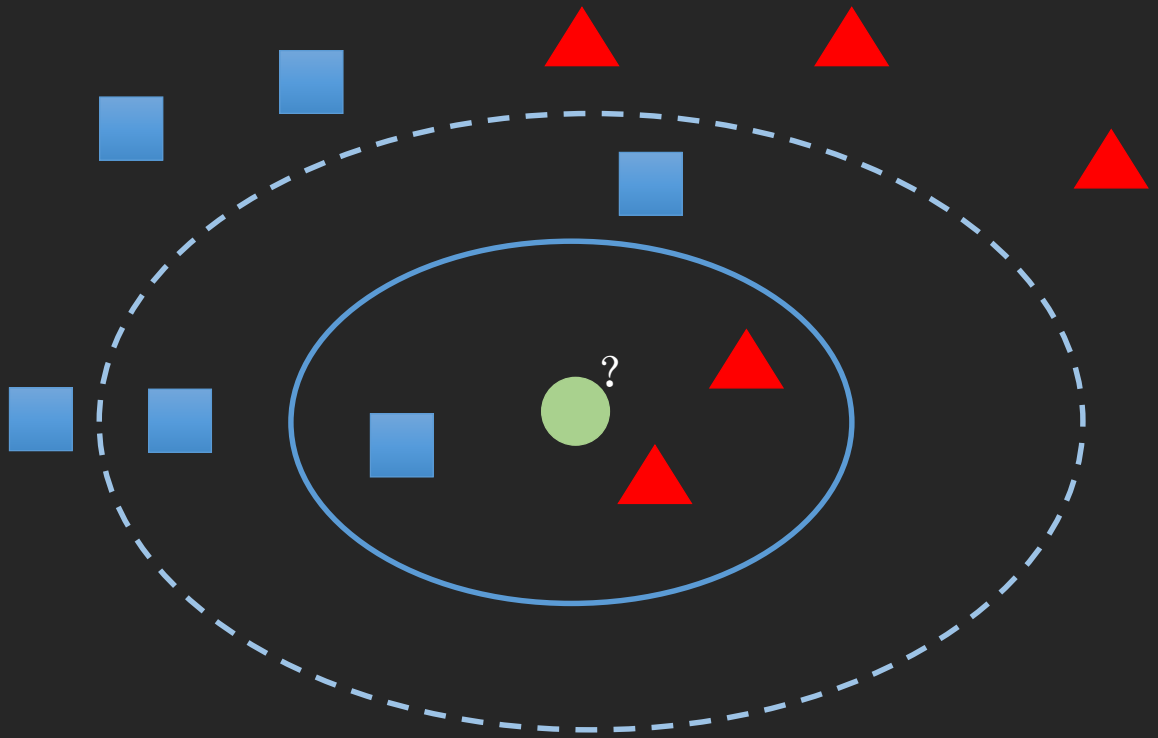
Q1：电影题材分类

- 类型
  - 如：动作片、爱情片
- 特征
  - 打斗场景、接吻镜头

分类的思想

- 给定用于训练的观测样本集 (X, Y)
- 目标：学习出一个输入和输出之间的规律或模型，有时可表示为函数 $f: X \rightarrow Y$
- 当出现新的未知观测 x，$f(x)$ 能预测出输出 y。

kNN的思想



kNN算法概述

- 监督学习
- 核心：采用测量不同特征值之间的距离进行分类
- 原理：存在一个训练样本集，已知样本集中每一数据的所属分类。当输入没有标签的新数据后，选择新数据与样本数据特征最相似（k近邻）的分类作为新数据的分类。
- k<=20

Q2：电影？？属于哪一类？

电影名称	打斗镜头	接吻镜头	电影类型
California Man	3	104	爱情片
He’s Not Really into Dudes	2	100	爱情片
Beautiful Woman	1	81	爱情片
Kevin Longblade	101	10	动作片
Robo Slayer 3000	99	5	动作片
Amped II	98	2	动作片
??	18	90	未知

例：电影分类

电影名称	与未知电影的距离
California Man	20.5
He's Not Really into Dudes	18.7
Beautiful Woman	19.2
Kevin Longblade	115.3
Robo Slayer 3000	117.4
Amped II	118.9

- $k = 3$

爱情片

kNN算法流程

- 计算已知类型样本集中的点与待分类点之间的距离；
- 按距离递增次序排序；
- 选取与当前点距离最小的k个点；
- 确定前k个点所在类别出现频率；
- 返回前k个点出现频率最高的类别作为当前点的预测分类。

ZIXING AI
自兴人工智能

关于“距离”

- 欧氏距离Euclidean distance

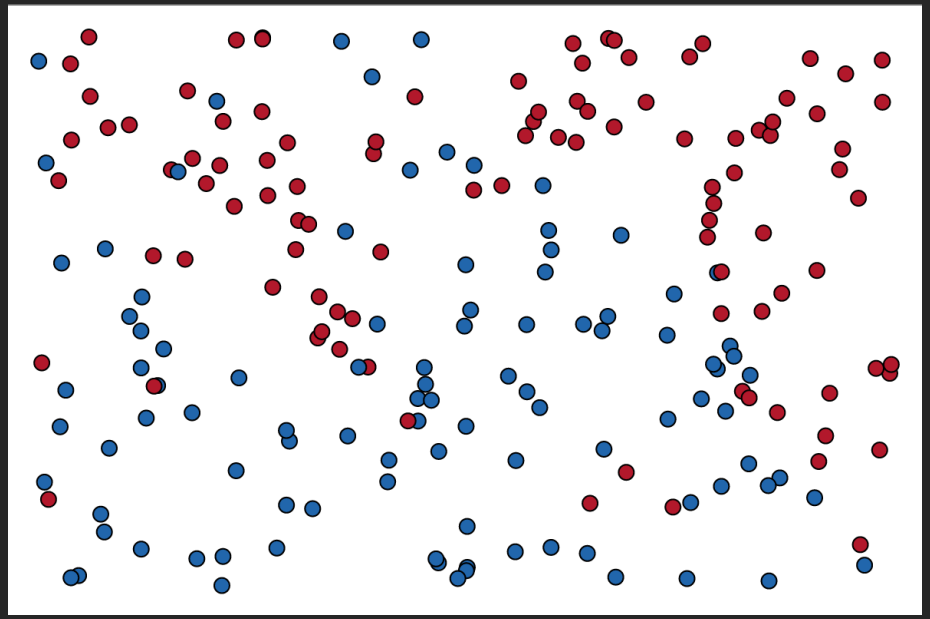
$$d(x, x') = \sqrt{(x_1 - x'_1)^2 + (x_2 - x'_2)^2 + \ldots + (x_n - x'_n)^2}$$

- 曼哈顿距离Manhattan distance

$$d_{12} = \sum_{k=1}^{n} |x_{1k} - x_{2k}|$$

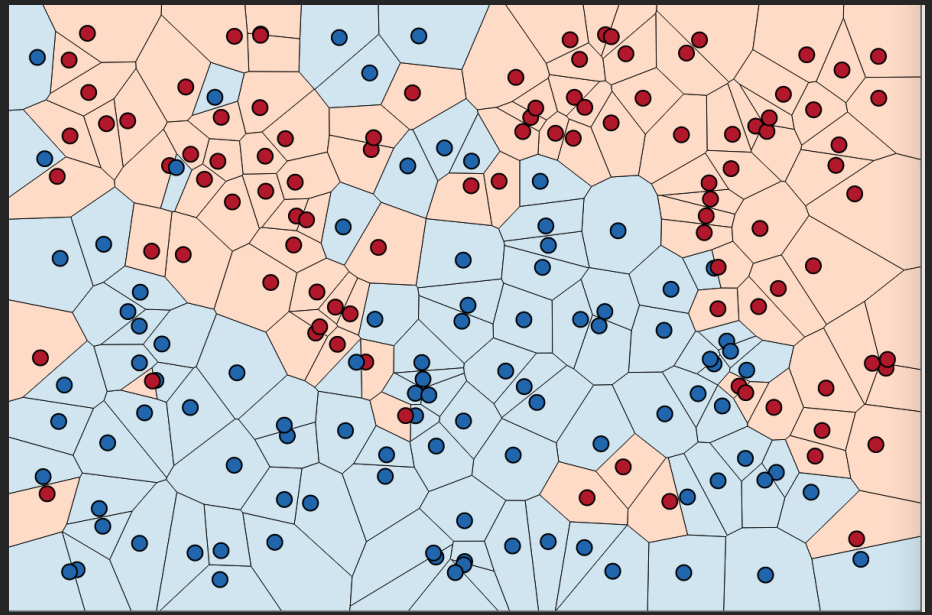
- 切比雪夫距离Chebyshev distance
- 汉明距离Hamming distance
- 马氏距离Mahalanobis distance

例：选民党注册



图片来源：http://scott.fortmann-roe.com/docs/BiasVariance.html

例：选民党注册



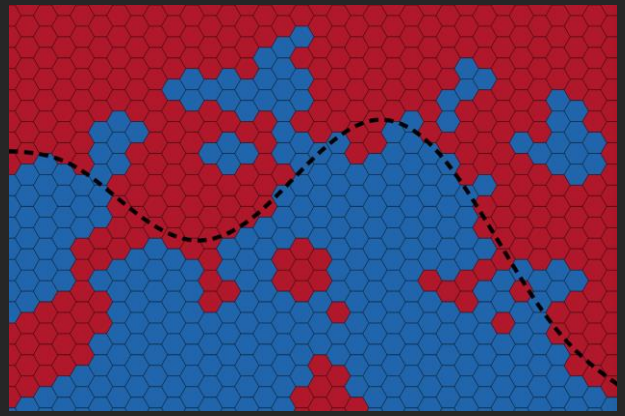
图片来源：http://scott.fortmann-roe.com/docs/BiasVariance.html

ZIXING AI
自兴人工智能

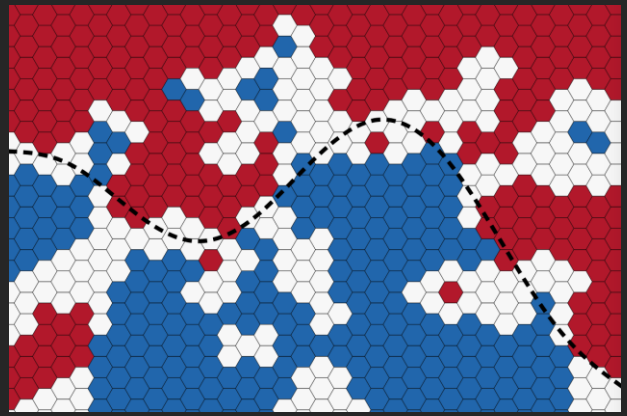
关于“k”

- 靠经验
- 交叉验证
  - 部分样本做训练
  - 部分做测试
- 初始较小值，不断调整到最佳。
- 一般取奇数

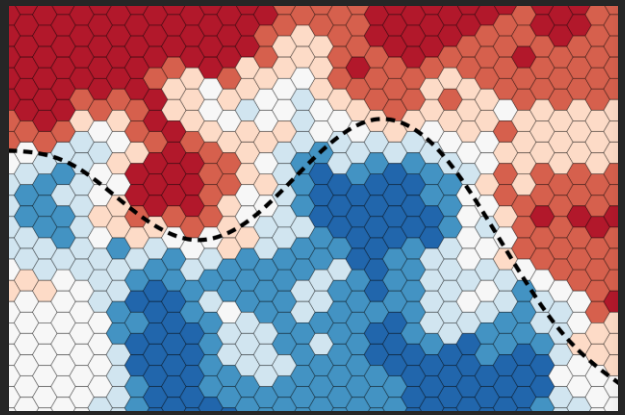
图片来源：http://scott.fortmann-roe.com/docs/BiasVariance.html



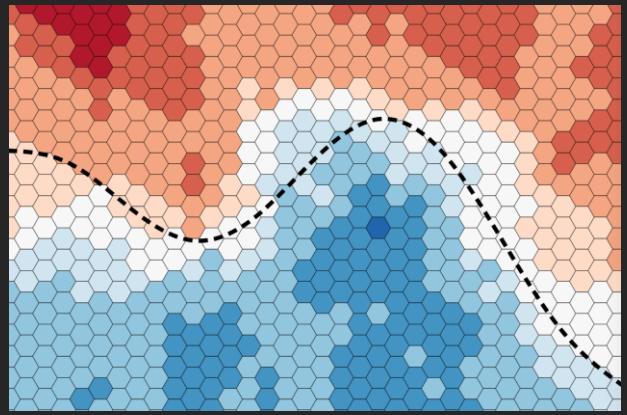
$k=1$



$k=2$



$k=5$



$k=20$

图片来源：http://scott.fortmann-roe.com/docs/BiasVariance.html

ZIXING AI
自兴人工智能

kNN实现

- 收集数据
- 准备数据
- 分析数据
- 训练算法
- 测试算法
- 使用算法

总结

- 适用数据范围
  - 数值型、标记型
- 监督学习
- 优点
  - 简单、精度高、鲁棒
- 缺点
  - 计算复杂度高、空间复杂度高
  - 如何改进?

ZIXING AI
自兴人工智能

练习：鸢尾花识别

- 数据集：Iris Plants Database, IFD
  - https://archive.ics.uci.edu/ml/datasets/Iris
- 示例数：150
- 特征（4）
  - sepal_length, sepal_width
  - petal_length, petal_width
- 种类（3）
  - Iris Setosa，Iris Versicolour，Iris Virginica

- 要求
  - 画出数据集带分类的图示
  - 获得准确率
  - 获得最佳k值，并分析