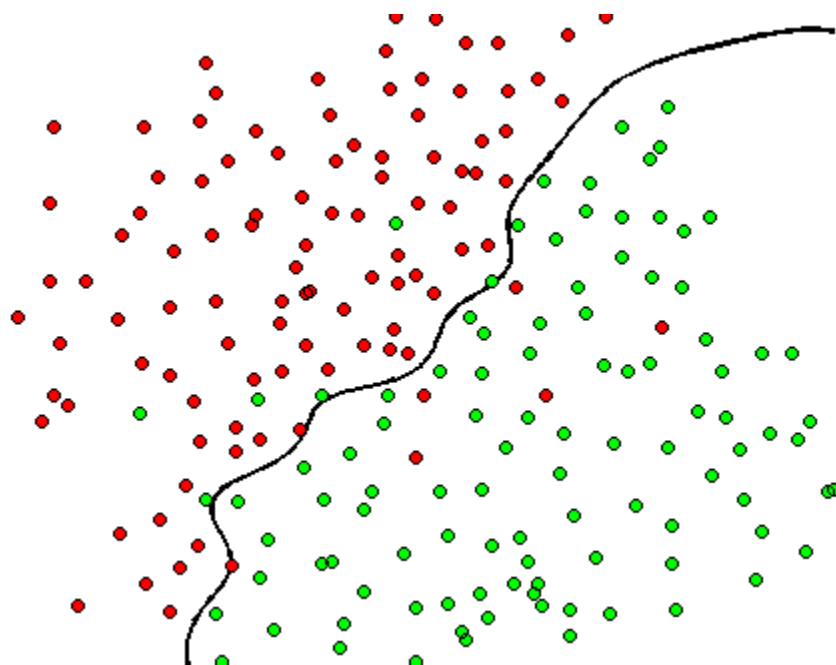


法律声明

- 本课件包括：演示文稿，示例，代码，题库，视频和声音等，小象学院拥有完全知识产权的权利；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或机构不得盗版、复制、仿造其中的创意，我们将保留一切通过法律手段追究违反者的权利。



关注 小象学院



分类模型(1)

--Robin

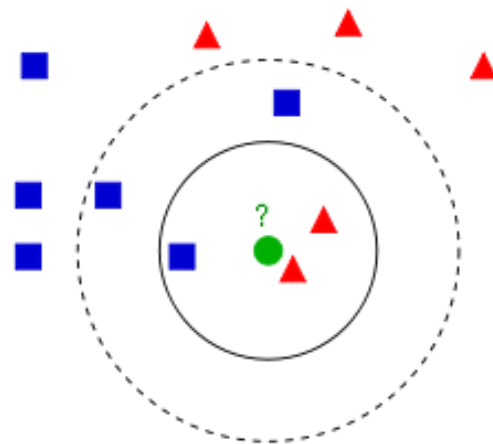
目录

- kNN
- 决策树
- 朴素贝叶斯

kNN

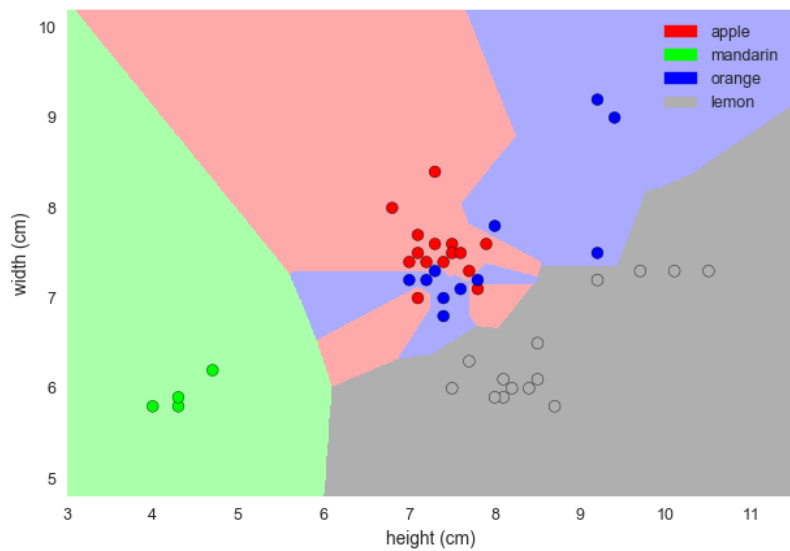
kNN (k-NearestNeighbor), k-近邻算法

- 是一种基于样本/实例的算法
- 无参 (nonparametric) 模型
- 可用于分类和回归
- 步骤:
 1. 计算出测试样本和所有训练样本的距离;
 2. 为测试样本选择k个与其距离最小的训练样本;
 3. 统计出k个训练样本中大多数样本所属分类;
 4. 这个分类就是待分类数据所属分类
- 问题:
 - k必须是奇数吗?

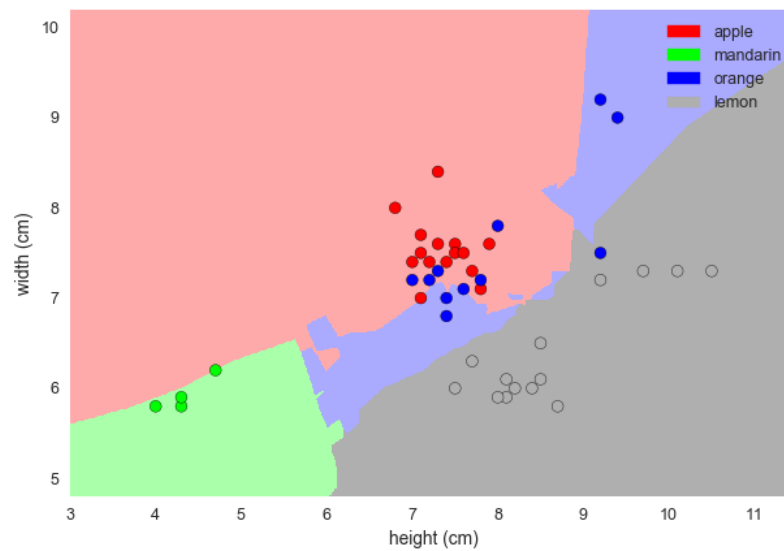


kNN

- 应用在“水果识别”数据集中



k=1

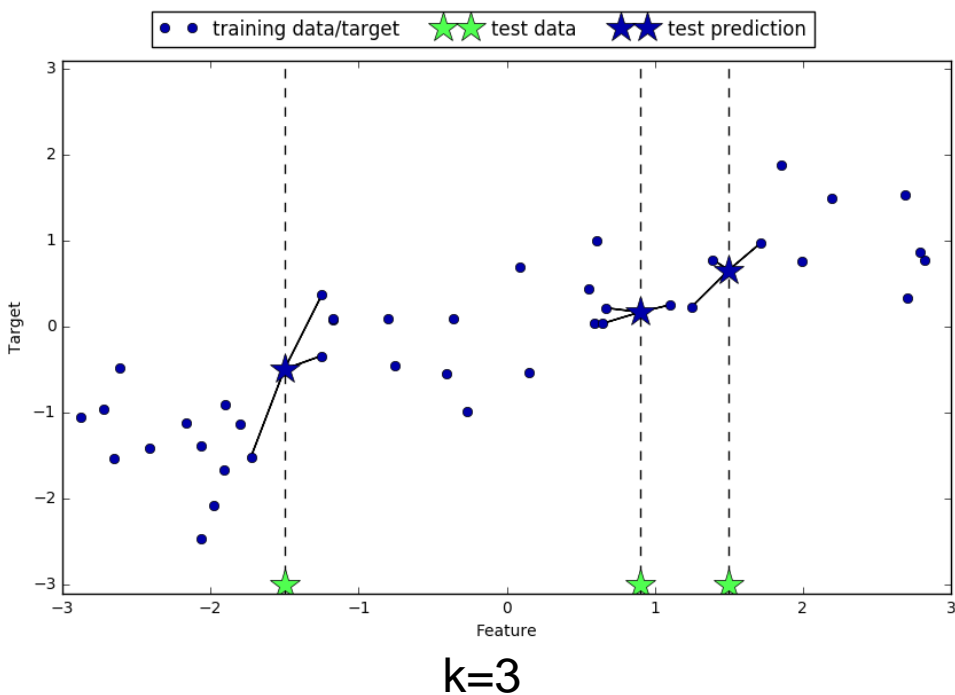
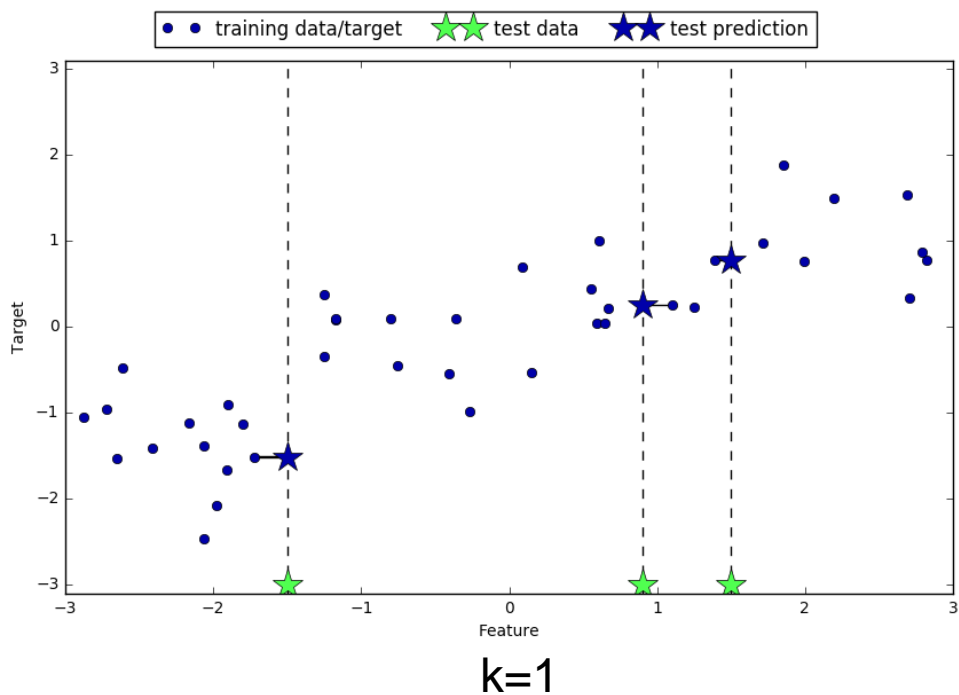


k=10

- 问题:
 - 如果k是无穷大, 结果是怎样的?

kNN

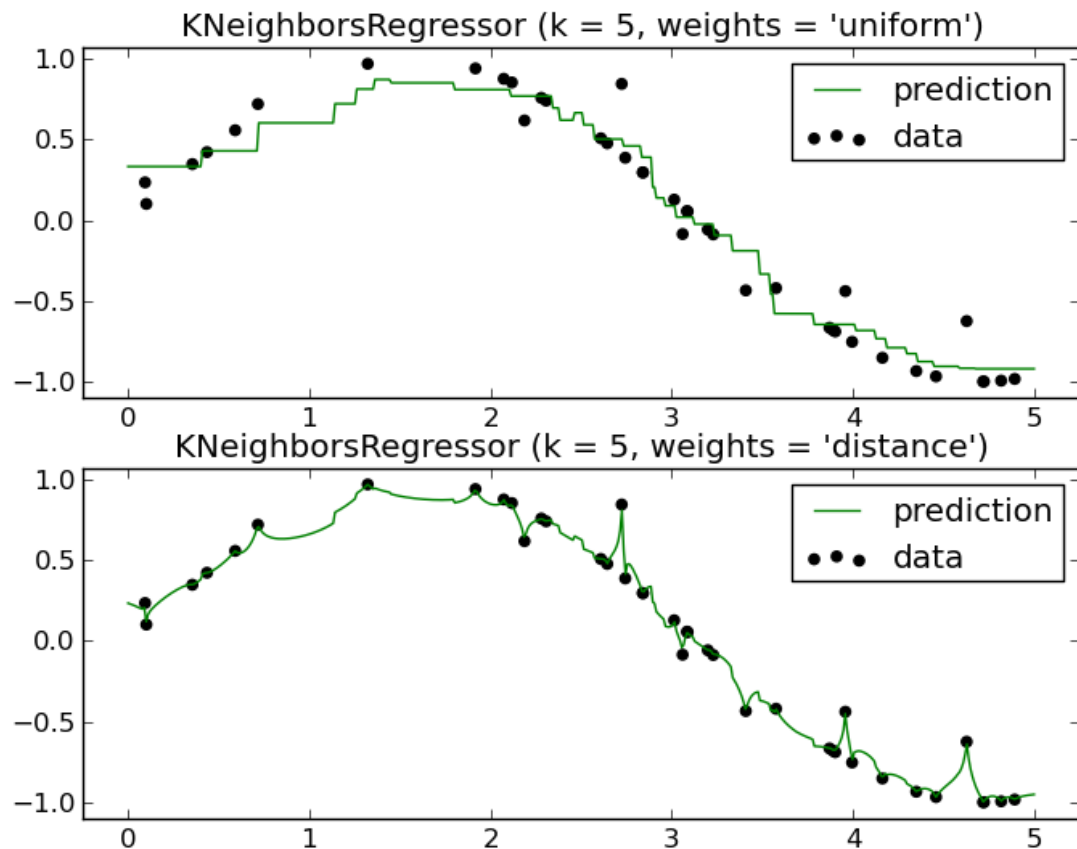
- kNN用于回归



ref: <https://elvinouyang.github.io/study%20notes/python-datasets-and-knn/>

kNN

- kNN用于回归

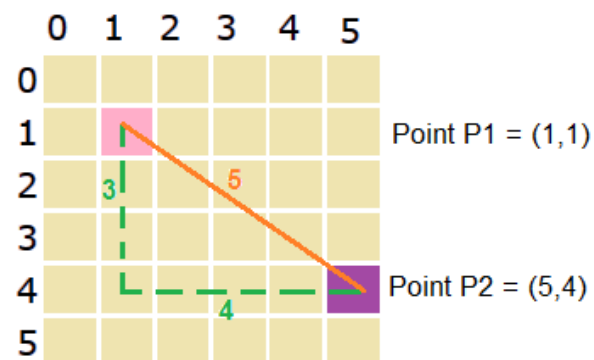


ref: <https://stats.stackexchange.com/questions/104255/why-would-anyone-use-knn-for-regression>

kNN

- kNN中的距离
 - [相似性度量](#)
 - 闵式距离
 - 欧式距离（默认）
 - $p=1$ ，曼哈顿距离
 - $p=2$ ，欧式距离
 - ...

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=0}^{n-1} |x_i - y_i|^p \right)^{1/p}$$



$$\text{Euclidean distance} = \sqrt{(5-1)^2 + (4-1)^2} = 5$$

$$\text{Manhattan distance} = |5-1| + |4-1| = 7$$

kNN

kNN (k-NearestNeighbor), k-近邻算法

- 需要注意的问题
 1. 相似性度量
 2. 紧邻点个数, 通过交叉验证得到最优紧邻点个数
- kNN优缺点

优点	缺点
<ul style="list-style-type: none">• 算法简单直观, 易于实现• 不需要额外的数据, 只依靠数据(样本)本身	<ul style="list-style-type: none">• 计算量较大, 分类速度慢• 需要预先指定k值

```
sklearn.neighbors.KNeighborsClassifier()  
sklearn.neighbors.KNeighborsRegressor()
```

联系我们

小象学院：互联网新技术在线教育领航者

— 微信公众号：**小象学院**

