

第3课时 KNN

K最近邻算法/K近邻算法

主讲教师：欧新宇

January 16, 2020

K近邻算法

- K近邻算法原理
- 二分类任务
- 多分类任务
- 回归分析

K近邻算法案例——酒的分类

K近邻算法案例——糖尿病预测

K近邻算法简介

K近邻算法的原理

KNN算法是一个典型的**监督学习算法**，它的核心思想是：
未标记样本的类别由**距离其最近的 K 个邻居通过投票**来决定。

具体而言，假设存在一个**已经标记好**的数据集。给定一个**未标记**的数据样本，我们的任务是：**预测出该数据样本所属的类别**。

KNN的原理是：

- 计算待标记样本和数据集中每个样本的距离
- 取距离最近的 K 个样本
- 待标记的样本所属类别由这 K 个距离最近的样本投票产生

K近邻算法简介

• KNN算法原理伪代码

- 假设 X_{test} 为待测样本， X_{train} 为已标记的数据集：
- 遍历 X_{train} 中所有的样本，计算每个样本与 X_{test} 的距离，并把距离保存在Distance数组中
- 对Distance数组进行排序，取距离最近的k个点，并保存到 X_{knn} 数组中
- 在 X_{knn} 中统计每个类别的个数，例如： X_{knn} 中有多少给样本属于类别0，多少个样本属于类别1.
- 在 X_{knn} 中样本数最多的类别即待测样本 X_{test} 的预测分类

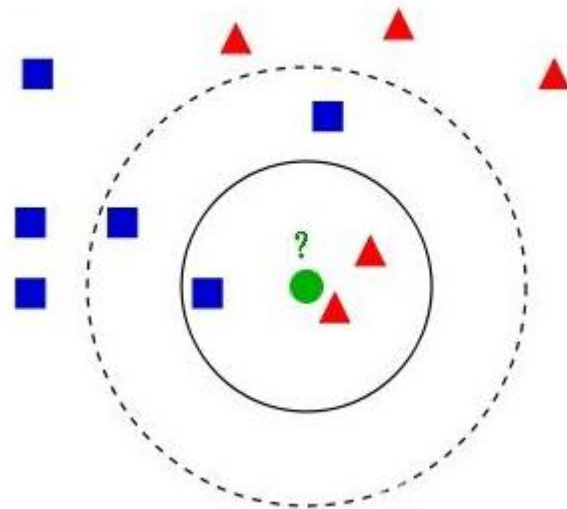
K近邻算法简介

KNN算法简单示例

下图中所显示的数据集是良好的数据集，即都有对应的标签，一类是蓝色正方形，一类是红色三角形，绿色圆形是待分类数据。

* **K = 3时**，范围内红色三角形较多，待分类样本属于红色三角形类

* **K = 5时**，范围内蓝色正方形较多，待分类样本属于蓝色正方形类



如何选择一个最佳的K值？

取决于数据

- 一般情况下，较大的K值能减少噪声的影响，但会使类别之间的界限变得模糊。
- 因此，K的取值通常较小（通常 $K < 20$ ）。
- 在scikit-learn中，K近邻算法的K值通过参数n_neighbors来调节，默认值为5。

K近邻算法简介

算法优缺点

- 优点:

简单，易于理解，无需建模与训练，且易于实现。适合对稀有事件进行分类，适合于多分类问题

- 缺点:

惰性算法，内存开销大，性能较差，可解释性差

K近邻算法的用法

各种需要载入的库文件

- #导入计算库

```
import numpy as np
```

- # 导入绘图工具箱 matplotlib

```
import matplotlib.pyplot as plt
```

- # 导入样数据集生成器

```
from sklearn.datasets import make_blobs
```

- # 从近邻算法子库中导入K近邻分类器KNeighborsClassifier

```
from sklearn.neighbors import KNeighborsClassifier
```

- # 从模型选择子库中导入数据集拆分工具

```
from sklearn.model_selection import train_test_split
```


K近邻算法的用法

二分类任务 (0301Bi-Classification.py)

多分类任务 (0302Multi-Classification.py)

- 生成数据集
- 划分训练集和测试集
- 基于训练集训练KNN模型
- 预测及评分

回归分析 (0303Regression.py)

- 生成数据集
- 划分训练集和测试集
- 基于训练集训练KNN模型
- 预测及评分
- 模型优化

K近邻算法案例分析

酒分类 (0304Case-Wine.py)

- 生成数据集及数据集分析
- 数据集拆分
- KNN建模
- 预测及评分
- 参数分析

糖尿病预测 (0305Case-Diabetes.py)

- 生成数据集及数据集分析
- 数据集拆分
- KNN建模
- 预测及评分
- 结果可视化

欧老师的联系方式

读万卷书 行万里路 只为最好的修炼

QQ: 14777591 (宇宙骑士)

Email: ouxinyu@alumni.hust.edu.cn

Tel: 18687840023