第8章：最近邻学习

蒋良孝 |  中国地质大学（武汉）

机器学习与数据挖掘团队

ljiang@cug.edu.cn

http://www.escience.cn/people/jlx/

本章内容

一、最近邻学习基础知识

二、最近邻学习基本思想

三、最近邻学习常见问题

一、最近邻学习基础知识

- 根据第一章绪论中对分类的定义可知：分类包含两个阶段：训练阶段和工作阶段。
- 到目前为止，我们前面介绍的所有机器学习技术都有显式的训练过程，都是在训练阶段就对训练样本进行学习处理，构建起分类模型，这类机器学习技术统称为“积极学习” (eager learning)。
- 这一章我们将要介绍的最近邻学习，没有显式的训练过程，在训练阶段只是把训练样本保存起来，建模工作延迟到工作阶段才进行处理，这类机器学习技术统称为“消极学习”(lazy learning)

一、最近邻学习基础知识

- 最近邻学习不是在整个样本空间上一次性地估计目标函数，而是针对每个待测样本作出局部的目标函数逼近。当目标函数很复杂，但它可以用不太复杂的局部函数来逼近时，这样做有非常明显的优势。
- 最近邻学习可以为不同的待测样本构建起不同的目标函数逼近，因此相比于那些积极的学习技术，最近邻学习往往具有较高的分类性能。

二、最近邻学习基本思想

- 最近邻学习的本思想非常简单：给定待测样本，首先基于某种近邻索引方法找出训练集中与其最靠近的K个样本，然后基于这K个样本的后验概率来预测待测样本的类标记。具体算法分为两个阶段：
- 训练阶段：

  将每个训练样本保存起来
- 工作阶段：

  给定一个待测样本
  - ✓ 基于某种近邻索引方法找出训练样本集中与其最靠近的K个样本
  - ✓ 基于这K个训练样本的后验概率来预测待测样本的类标记

三、最近邻学习常见问题

- 尽管最近邻学习的基本思想非常简单，但在实际应用中，经常会遇到很多的现实问题：

1）近邻索引问题

2）维度灾害问题

3）邻域大小问题

4）后验概率问题

5）计算效率问题

6）归纳偏置问题

三、最近邻学习常见问题

1）近邻索引问题：

- 最近邻学习的所有计算几乎都花费在索引近邻问题上。所以，如何有效地索引近邻样本，以减少分类时所需计算是一个重要的实践问题。
- 目前，使用最多的近邻索引方法就是通过计算待测样本与每一个训练样本之间的距离，然后基于距离排序，选择距离最短的K个训练样本作为待测样本的最近邻样本。因此如何度量样本点之间的距离就变得非常重要了。

三、最近邻学习常见问题

1）近邻索引问题：

- 为度量样本点之间的距离，学者们提出许多经典的距离度量函数。根据样本点的数据类型分，主要有：

✓ 连续属性：Euclidean距离、Manhattan 距离等

✓ 离散属性：Overlap Metric距离、 Value Difference Metric距离等

✓ 混合属性： Heterogeneous Euclidean-Overlap Metric (HEOM)距离、Heterogeneous Value Difference Metric (HVDM)等

简单说，样本点之间的距离总可以分解成样本点在每一维上的差；然后再根据每一维上的数据类型来选择合适的距离函数。比如HEOM距离就是Euclidean距离和Overlap距离异构而成的，HVDM距离就是Euclidean距离和VDM距离异构而成的[Wilson & Martinez, 1997]。

三、最近邻学习常见问题

1）近邻索引问题：

- 除了上述基于距离排序的索引方法之外，目前还开发了许多对存储的训练样本进行索引的方法，以便更快速地确定最近邻样本。比如KD-Tree方法把训练样本存储在树的叶子结点上，邻近的样本存储在相同或相近的叶子结点上，然后通过测试待测样本在内部结点上的划分属性把待测样本划分到相关的叶子结点上。

这种方法因为树的构建是在训练阶段进行的，所以比基于距离排序的索引方法所需的计算量要小得多。但如何构建有效的树成了另外一个需要解决的问题。

三、最近邻学习常见问题

2）维度灾害问题：

- 前面讲到的许多学习方法，比如决策树学习，只测试部分属性就可作出判断，而最近邻学习中样本间的距离是根据样本的所有属性来计算的。如果目标函数仅依赖于很多属性中的几个时，样本间的距离会被大量不相关的属性所支配，从而导致相关属性的值很接近的样本相距很远。
- 这种由于存在很多不相关属性所导致的难题，被称为维度灾难。解决维度灾害问题的常用方法主要包括：1）属性加权；2）属性选择。

三、最近邻学习常见问题

3）邻域大小问题：

- 最近邻学习有一个很重要的参数，那就是邻域的大小，即最近邻样本的数目K，最近邻学习的预测结果与K的大小密切相关。同样的数据，K值不同可能导致不同的预测结果。
- 目前对于K值的选取主要有两种办法：1）基于经验直接给定；2）基于数据自动学习。

三、最近邻学习常见问题

4）后验概率问题：

- 给定待测样本的K个最近邻样本，估计其后验概率的常用方法包括：投票法、加权投票法、局部概率模型法。当计算得到的后验概率出现相同的情况下，可以采用随机分类或者拒判的方法进行处理。
- **在计算后验概率的过程经常会采用一些常用的概率估计方法：基于频率的极大似然估计、拉普拉斯估计、基于相似度（距离）加权的拉普拉斯估计、m-估计，朴素贝叶斯估计等等。**

三、最近邻学习常见问题

5）计算效率问题：

- 最近邻学习推迟所有的计算处理，直到接收到一个新的待测样本，所以分类每个新的待测样本就需要大量的计算。
- 高效的近邻索引方法可以在一定程度上缓解计算效率问题，比如KD-Tree近邻索引方法。

三、最近邻学习常见问题

6）归纳偏置问题：

- 最近邻学习的归纳偏置是：在输入空间上相近的样本点具有相似的目标函数输出。也就是说，一个待测样本的类标记与它在输入空间中相邻的训练样本的类标记相似。
- 有效的距离度量方法可以在一定程度上缓解归纳偏置问题，比如属性加权的距离度量方法。