Naive Bayes、KNN、Softmax分类器的比较研究（附代码地址）

本文主要针对naïve bayes、knn和softmax三种分类模型进行比较研究，通过分析不同模型的特点找到相应场景应使用的模型。

1. 算法介绍
   1. Naïve Bayes

朴素贝叶斯分类器主要基于贝叶斯公式:

 (1)

其中P(c)是类”先验”概率；P(x|c)是样本x相对于类标记c的类条件概率；P(x)是用于归一化的”证据”因子。对给定样本x，证据因子P(x)与类标记无关，因此估计P(c|x)的问题就转化为如何基于训练数据D来估计先验P(c)和条件概率P(x|c)。朴素贝叶斯分类器采用了”属性条件独立性假设”：对已知类别，假设所有属性相互独立，换言之，假设每个特征属性独立地对分类结果发生影响。

基于”属性条件独立性假设”，上述公式可重写为：

 (2)

其中1<=i<=d，d为特征属性个数，xi为x在第i个属性上的取值。

* 1. KNN

K近邻(k-Nearest Neighbor,简称KNN)学习是一种常用的监督学习方法，其工作机制非常简单：给定测试样本，基于某种距离度量找出训练集中与其最靠近的k个训练样本，然后基于这k个”邻居”的信息进行预测。通常，在分类任务中可使用”投票法”，即选择这k个样本中出现最多的类别标记作为预测结果；在回归任务中可以使用”平均法”，即将这k个样本的实值输出标记的平均值作为预测结果；还可基于距离远近进行加权平均或加权投票，距离越近的样本权重越大。

* 1. Softmax

Softmax分类器又称多元逻辑回归，是逻辑回归的扩展，众所周知逻辑回归仅二值分类不可对多值分类，为解决此问题，诞生了softmax分类器。

Logistic regression中：

 (3)

 (4)

Softmax中：

 (5)

1. 实验分析

通过使用UCI数据集中的iris和wine对以上三种分类模型进行实验对比，使用的评价指标为准确率。

表1三种分类模型分类结果

|  |  |  |
| --- | --- | --- |
| data | iris | wine |
| Softmax | 97.53% | 68.89% |
| KNN | 90.24% | 62.22% |
| Baive Bayes | 90.12% | 60% |

通过实验结果可以看出softmax相对于其他两种分类器效果更好，通过KNN与Naïve Bayes对比，发现尽管KNN算法相对简单，但效果还不错。

代码地址：

<https://github.com/Emmitte/ClassificationModel>

http://download.csdn.net/detail/u013473512/9684562