朴素贝叶斯

黑马程序员
www.itheima.com

传智教育旗下
高端IT教育品牌

目录

Contents

学习目标

Learning Objectives

1. 复习常见概率的计算
2. 知道贝叶斯公式
3. 了解朴素贝叶斯是什么
4. 了解拉普拉斯平滑系数的作用

朴素贝叶斯算法-利用概率值进行分类的一种机器学习算法

- 什么是概率？

  一件事情发生的可能性，取值在【0,1】之间。比如：抛硬币正面向上的概率、6面骰子抛出5这一面的概率

- 例子：判断女神对你的喜欢情况

样本数	职业	体型	女神是否喜欢
1	程序员	超重	不喜欢
2	产品	匀称	喜欢
3	程序员	匀称	喜欢
4	程序员	超重	喜欢
5	美工	匀称	不喜欢
6	美工	超重	不喜欢
7	产品	匀称	喜欢

朴素贝斯算法 - 概率数学基础复习

- **条件概率**：表示事件A在另外一个事件B已经发生条件下的发生概率，P(A|B)
  - 比如：在女神喜欢的条件下，职业是程序员的概率？
    - 女神喜欢条件下，有 2、3、4、7 共 4 个样本
    - 4 个样本中，有程序员 3、4 共 2 个样本
    - 则 P(程序员|喜欢) = 2/4 = 0.5
  - 【思考】在女神不喜欢的条件下，职业是程序员的概率是多少？
- **联合概率**：表示多个条件同时成立的概率，P(AB) = P(A) * P(B|A) = P(B)* P(A|B)
  - 比如：职业是程序员并且体型匀称的概率？
    - 数据集中，共有 7 个样本
    - 职业是程序员有 1、3、4 共 3 个样本，则其概率为：3/7
    - 在职业是程序员，体型是匀称的有样本3，共 1 个样本，则其概率为：1/3
    - 则即是程序员又体型匀称的概率为：3/7 * 1/3 = 1/7
  - 【思考】：体型匀称并且是程序员的概率是多少？ P(B) P(A|B)

朴素贝斯算法 - 概率数学基础复习

- 联合概率 + 条件概率
  - 比如：在女神喜欢的条件下，职业是程序员、并且超重的概率？ P(程序员, 超重|喜欢)
    - 在女神喜欢的条件下，有 2、3、4、7 共 4 个样本
    - 在这 4 个样本中，职业是程序员有 3、4 共 2 个样本，则其概率为：2/4=0.5
    - 在这 2 个样本中，体型超重的有样本4， 共 1 个样本，则其概率为：1/2 = 0.5
    - 则 P(程序员, 超重|喜欢) = 0.5 * 0.5 = 0.25

- 相互独立：如果P(A, B) = P(A)P(B)，则称事件A与事件B相互独立
  - 比如：女神喜欢程序员的概率，女神喜欢产品经理的概率，两个事件没有关系

- 简言之

  条件概率：在去掉部分样本的情况下，计算某些样本的出现的概率，表示为：P(B|A)

  联合概率：多个事件同时发生的概率是多少，表示为：P(AB) = P(B)*P(A|B)

朴素贝斯算法

- 贝叶斯公式

$$P(C \mid W) = \frac{P(W \mid C)P(C)}{P(W)}$$

  - P(C) 表示 C 出现的概率，一般是目标值
  - P(W | C) 表示 C 条件 W 出现的概率
  - P(W) 表示 W 出现的概率

编号	职业	体型	喜欢的概率?
1	程序员	超重	?

1. P(C | W) = P(喜欢 | (程序员，超重))
2. P(W | C) = P((程序员，超重) | 喜欢)
3. P(C) = P(喜欢)
4. P(W) = P(程序员，超重)

朴素贝斯算法

1.根据训练样本估计先验概率P(C)： P(C) = P(喜欢) = 4/7

2.根据条件概率P(W | C)调整先验概率： P(W | C) = P((程序员,超重) | 喜欢) = 1/4

3.此时我们的后验概率P(W | C) * P(C)为： P(W | C) * P(C) = P((程序员,超重) | 喜欢) * P(喜欢) = 4/7 * 1/4 = 1/7

4.那么该部分数据占所有既为程序员，又超重的人中的比例是多少呢?

P(W) = P(程序员,超重) = P(程序员) * P(超重 | 程序员) = 3/7 * 2/3 = 2/7

$$P(C \mid W) = \frac{P(W \mid C)P(C)}{P(W)}$$ = P(喜欢 | (程序员, 超重)) = 1/7 ÷ 2/7 = 0.5

朴素贝斯算法

朴素贝叶斯在贝叶斯基础上增加：**特征条件独立假设**，即：特征之间是互为独立的。

此时，联合概率的计算即可简化为：

1. P(程序员,超重|喜欢) = P(程序员|喜欢) * P(超重|喜欢)
2. P(程序员,超重) = P(程序员) * P(超重)

拉普拉斯平滑系数

为了避免概率值为 0，我们在分子和分母分别加上一个数值，这就是拉普拉斯平滑系数的作用

$$P(F_1 \mid C) = \frac{N_i + \alpha}{N + \alpha m}$$

α 是拉普拉斯平滑系数，一般指定为 1

N_i 是 F1 中符合条件 C 的样本数量

N 是在条件 C 下所有样本的总数

m 表示**所有独立样本**的总数

总结

1 朴素贝叶斯算法概念

用概率值进行分类的一种机器学习算法。贝叶斯基础上增加**特征条件独立假设。**

2 概率数学基础

- 概率：一件事情发生的可能性。
- 条件概率：表示事件A在另外一个事件B已经发生条件下的发生概率，P(A|B)
- 联合概率：表示多个条件同时成立的概率，P(AB) = P(A) * P(B|A) = P(B)* P(A|B)

3 贝叶斯公式

$$P(C \mid W) = \frac{P(W \mid C)P(C)}{P(W)}$$

4 拉普拉斯平滑系数

为了避免概率值为 0，我们在分子和分母分别加上一个数值，这就是拉普拉斯平滑系数的作用

练习

1、下列关于朴素贝叶斯实现原理的描述正确的是：（） **【多选】**

A）它是一种经典的概率统计方法

B）它假设样本之间是相互独立的

C）使用它需要计算出相应的联合概率和条件概率

D）为了避免概率值为 0，我们在分子和分母分别加上一个数值，这就是拉普拉斯平滑系数的作用。

答案：ABCD

目录

Contents

学习目标

Learning Objectives

1. 知道朴素贝叶斯的API
2. 能够应用朴素贝叶斯实现商品评论情感分析

案例：商品评论情感分析

- **朴素贝叶斯API**

sklearn.naive_bayes.MultinomialNB(alpha = 1.0)

- 朴素贝叶斯分类
- alpha：拉普拉斯平滑系数

商品评论情感分析

- 需求　已知商品评论数据，根据数据进行情感分类（好评、差评）

	Unnamed: 0	内容	评价
0	0	从编程小白的角度看，入门极佳。	好评
1	1	很好的入门书，简洁全面，适合小白。	好评
2	2	讲解全面，许多小细节都有顾及，三个小项目受益匪浅。	好评
3	3	前半部分讲概念深入浅出，要言不烦，很赞	好评
4	4	看了一遍还是不会写，有个概念而已	差评
5	5	中规中矩的教科书，零基础的看了依旧看不懂	差评
6	6	内容太浅显，个人认为不适合有其它语言编程基础的人	差评
7	7	破书一本	差评
8	8	适合完完全全的小白读，有其他语言经验的可以去看别的书	差评
9	9	基础知识写的挺好的！	好评
10	10	太基础	差评
11	11	略_嗦。。适合完全没有编程经验的小白	差评
12	12	真的真的不建议买	差评

商品评论情感分析流程

1 获取数据

2 数据基本处理

  # 2-1 处理数据y

  # 2-2 加载停用词

  # 2-3 处理数据x 把文档分词

  # 2-4 统计词频矩阵 作为句子特征

  # 2-5 准备训练集测试集

3 模型训练

  # 4-1 实例化贝叶斯 添加拉普拉斯平滑参数

4 模型预测

5 模型评估

案例：商品评论情感分析

```
# 1.导入依赖包
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import jieba
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.naive_bayes import MultinomialNB  # 多项分布朴素贝叶斯
def dm02_模型训练():
    # 2.获取数据
    data_df = pd.read_csv('./data/书籍评价.csv', encoding='gbk')
    print('data_df-->\n', data_df)

    # 3.数据基本处理
    # 3-1 处理数据y
    data_df['评论标号'] = np.where(data_df['评价'] == '好评', 1, 0)
    y = data_df['评论标号']
    print('data_df-->\n', data_df)

    # 3-2 加载停用词
    stopwords = []
    with open('./data/stopwords.txt', 'r', encoding='utf-8') as f:
        lines = f.readlines()
        stopwords = [line.strip() for line in lines]
    stopwords = list(set(stopwords))  # 去重
```

案例：商品评论情感分析

```
# 3-3 处理数据x 把文档分词
comment_list = [','.join(jieba.lcut(line)) for line in data_df['内容']]
# print('comment_list-->\n', comment_list)

# 3-4 统计词频矩阵 作为句子特征
transfer = CountVectorizer(stop_words=stopwords)
x = transfer.fit_transform(comment_list)
mynames = transfer.get_feature_names()
x = x.toarray()

# 3-5 准备训练集测试集
x_train = x[:10, :]       # 准备训练集
y_train = y.values[0:10]
x_test = x[10:, :]        # 准备测试集
y_test = y.values[10:]
print('x_train.shape-->',x_train.shape)
print('y_train.shape-->',y_train.shape)
```

```
# 4.模型训练
# 4-1 实例化贝叶斯# 添加拉普拉修正平滑参数
mymultinomialnb = MultinomialNB()
mymultinomialnb.fit(x_train, y_train)

# 5.模型预测
y_pred = mymultinomialnb.predict(x_test)
print('预测值-->', y_pred)
print('真实值-->', y_test)

# 6.模型评估
myscore = mymultinomialnb.score(x_test, y_test)
print('myscore-->', myscore)
```


黑马程序员
www.itheima.com
传智教育旗下高端IT教育品牌