燕山大学

Python 机器学习实验指导书

Python machine learning Experiment Instruction Book

实验二：朴素贝叶斯和 SVM

教 务 处

2021 年 2 月

实验二　朴素贝叶斯和 SVM

一、实验目的

1. 理解并掌握经典的朴素贝叶斯和 SVM 分类算法。

2. 能够基于朴素贝叶斯和 SVM 分类算法分别实现印第安人分类、鸢尾花分类。

3. 能够举一反三，基于朴素贝叶斯和 SVM 分类算法实现肿瘤的分类与预测。

二、实验原理

（一）朴素贝叶斯

1. 朴素贝叶斯概述

简言之，朴素贝叶斯方法通过事件的先验概率预测事件的后验概率。

对于先验概率和后验概率，这里用一个例子来简单解释一下。假设今天开车出门不巧碰上堵车，我们想知道因为交通事故造成堵车的可能性是多少，这就是后验概率；但如果我们今天走之前看了看新闻，发现我们上班所必经的 XX 路发生了交通事故，这时我们如果坚持开车出门，那么堵车的可能性是多少，这就是先验概率。

- 先验概率：由因得出果 --> 由知道发生交通事故得出**堵车的可能性**
- 后验概率：由果推出因 --> 堵车后推断**发生交通事故的可能性**

2. 基本思想：概况为先验概率+数据=后验概率。实际过程中通过后验概率完成预测或分类，要求参数较少的算法。

（1）先验概率为高斯分类的朴素贝叶斯

应用场景：大部分特征分布为连续的情况。

```
from sklearn.naive_bayes import GaussianNB  #导入先验概率为高斯分布的朴素贝叶斯
```

（2）先验概率为多项式分布的朴素贝叶斯

应用场景：样本特征是多元、离散的情况。

```
from sklearn.naive_bayes import MultinomialNB  #导入先验概率为多项式分布的朴素贝叶斯
```

（3）先验概率为伯努利的朴素贝叶斯

应用场景：样本特征是二元离散或多元离散情况.

```
from sklearn.naive_bayes import BernoulliNB  #导入先验概率为伯努利的朴素贝叶斯
```

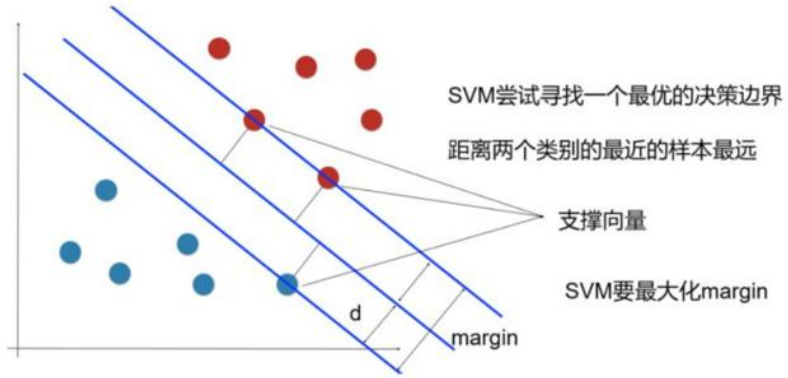
（二）SVM 分类

1. SVM 概述

支持向量机（Support Vector Machine ，SVM）是常见的一种分类方法。

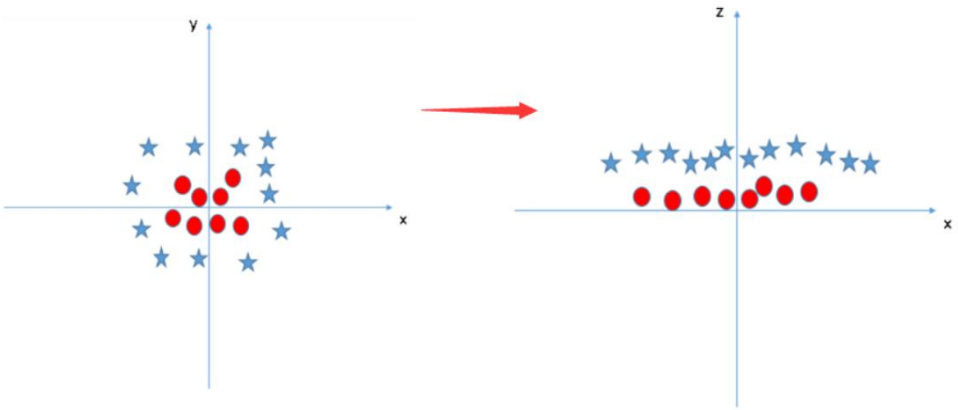
在机器学习中，SVM 是有监督的学习模型，是机器学习里面很经典的一个算法。

主要思想是建立一个最优决策超平面，使得该平面两侧距离该平面最近的两类样本之间的距离最大化，从而对分类问题提供良好的泛化能力。



2.SVM 的优点

- 相对于其他训练分类算法不需要过多样本，并且由于 SVM 引入了核函数，所以 SVM 可以处理高维样本。
- 结构风险最小。这种风险是指分类器对问题真实模型的逼近与问题真实解之间的累积误差。
- 非线性，是指 SVM 擅长应付样本数据线性不可分的情况，主要通过松弛变量（也叫惩罚变量）和核函数技术来实现，这一部分也正是 SVM 的精髓所在。



（三）鸢尾花分类数据集 iris

鸢尾花分类是机器学习中比较经典的入门式教学课程。

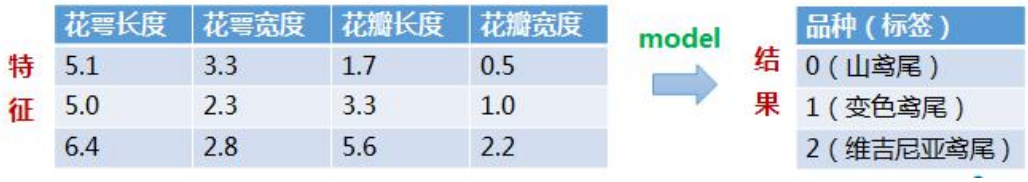
鸢尾花数据集总共包含 150 行数据。每一行数据由 4 个特征值及一个目标值组成。

4 个特征值分别为：萼片长度（SepalLengthCm）、萼片宽度（SepalWidthCm）、花瓣长度（PetalLengthCm）、花瓣宽度（PetalWidthCm）。

目标值为三种不同类别的鸢尾花，分别为：Iris-setosa（山鸢尾），Iris-versicolor（杂色鸢尾），Iris-virginica（维吉尼亚鸢尾）。

```
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
5.4,3.9,1.7,0.4,Iris-setosa
4.6,3.4,1.4,0.3,Iris-setosa
5.0,3.4,1.5,0.2,Iris-setosa
4.4,2.9,1.4,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
5.4,3.7,1.5,0.2,Iris-setosa
```

鸢尾花分类数据集



鸢尾花分类

（四）Pima（皮马）印第安人数据集

该数据集最初来自美国国立糖尿病/消化/肾脏疾病研究所。

数据集的目的是基于数据集中包含的某些诊断指标，来诊断性地预测患者是否患有糖尿病。

从较大的数据库中选择这些实例有几个约束条件。尤其是，这里的所有患者都是皮马印第安人血统、至少 21 岁的女性。

数据集由多个医学预测变量和一个目标变量组成。预测变量包括患者的怀孕次数、BMI、胰岛素水平、年龄等。目标变量是 class。

	preg	plas	pres	skin	test	mass	pedi	age	class
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Pima（皮马）印第安人数据集

数据集的特征值：

【1】preg：怀孕次数

【2】plas：葡萄糖（口服葡萄糖耐量试验中血浆葡萄糖浓度）

【3】pres：血压（舒张压）（mm Hg)

【4】skin：皮层厚度（三头肌组织褶厚度）（mm)

【5】test：胰岛素（2 小时血清胰岛素）（mu U / ml）

【6】mass：BMI 体重指数 （体重/身高）^2

【7】pedi：糖尿病谱系功能（糖尿病系统功能）

【8】age：年龄（岁）

【9】class：类标变量（0 或 1）

（五）威斯康星乳腺癌数据集

威斯康星乳腺癌数据集来自美国威斯康星州的乳腺癌诊断数据集。医疗人员采集了患者乳腺肿块经过细针穿刺（FNA）后的数字化图像，并且对这些数字图像进行了特征提取，这些特征可以描述图像中的细胞核呈现。

该数据集中肿瘤是一个非常经典的用于医疗病情分析的数据集，包括 569 个病例的数据样本，每个样本具有 30 个特征。样本共分为两类：恶性(Malignant)和良性(Benign)。

字段	含义
ID	ID标识
diagnosis	M/B（M：恶性，B：良性）
radius_mean	半径（点中心到边缘的距离）平均值
texture_mean	文理（灰度值的标准差）平均值
perimeter_mean	周长 平均值
area_mean	面积 平均值
smoothness_mean	平滑程度（半径内的局部变化）平均值
compactness_mean	紧密度（=周长*周长/面积-1.0）平均值
concavity_mean	凹度（轮廓凹部的严重程度）平均值
concave points_mean	凹缝（轮廓的凹部分）平均值
symmetry_mean	对称性 平均值
fractal_dimension_mean	分形维数（=海岸线近似-1）平均值
radius_se	半径（点中心到边缘的距离）标准差
texture_se	文理（灰度值的标准差）标准差
perimeter_se	周长 标准差
area_se	面积 标准差
smoothness_se	平滑程度（半径内的局部变化）标准差
compactness_se	紧密度（=周长*周长/面积-1.0）标准差
concavity_se	凹度（轮廓凹部的严重程度）标准差
concave points_se	凹缝（轮廓的凹部分）标准差
symmetry_se	对称性标准差
fractal_dimension_se	分形维数（=海岸线近似-1）标准差
radius_worst	半径（点中心到边缘的距离）最大值
texture_worst	文理（灰度值的标准差）最大值
perimeter_worst	周长 最大值
area_worst	面积 最大值
smoothness_worst	平滑程度（半径内的局部变化）最大值
compactness_worst	紧密度（=周长*周长/面积-1.0）最大值
concavity_worst	凹度（轮廓凹部的严重程度）最大值
concave points_worst	凹缝（轮廓的凹部分）最大值
symmetry_worst	对称性 最大值
fractal_dimension_worst	分形维数（=海岸线近似-1）最大值

威斯康辛乳腺癌数据集特征

属性信息：

ID——身份证号码

diagnose——诊断结果（M=恶性，B=良性）

其中'B'代表良性，包含 357 例；'M'代表恶性，包含 212 例。

计算每个细胞核的 10 个实值特征：

1） radius_mean：半径（从中心到周界各点的平均距离）

2） texture_mean：纹理（灰度值的标准偏差）

3） perimeter_mean：周长

4） area_mean：面积

5） smoothness_mean ：平滑度（半径长度的局部变化）

6） compactness_mean：密实度/紧密度（周长^2/面积-1.0）

7） concavity_mean：凹度（轮廓凹陷部分的严重程度）

8） concave points_mea ：凹点（轮廓凹面部分的数量）

9） symmetry_mean：对称性

10）fractal_dimension_mean：分形维数（“海岸线近似值”-1）

Mean、se、worst：为每个图像计算这些特征，产生了 30 个特征。所有特征值用四个有效数字重新编码。

- 包含 mean 的数据——平均值。
- 包含 se 的数据——标准误差。
- 包含 worst 的数据——最差值或最大值（三者中的平均值最大值），是最严重的数据样例(最坏值)。

三、实验环境

计算机；网络环境。

四、实验内容及步骤

（一）实验内容

1.对于给定的 2 个例题，基于朴素贝叶斯和 SVM 分别进行印第安人分类、鸢尾花分类等练习。

2.对于给定的 2 个项目，自行编写程序，分别使用朴素贝叶斯和 SVM 分类算法对威斯康星乳腺癌数据集进行肿瘤的分类与预测。

（二）实验步骤

1.进入指定实验课程，可通过阅览“知识讲解”进行实验相关知识的查缺补漏。

- 知识讲解：朴素贝叶斯算法
- 知识讲解：SVM 算法

2.在熟悉了实验原理的基础上，进行“实验二：朴素贝叶斯和 SVM”例题部分的实操练习。

- 例题 1——基于朴素贝叶斯来实现印第安人数据集分类
- 例题 2——鸢尾花分类

3.在步骤 2 例题练习的基础上，对给定的 2 个实操项目，自行编写程序，分别使用朴素贝叶斯和 SVM 分类算法对威斯康星乳腺癌数据集进行肿瘤的分类与预测。

- 实操项目 1——肿瘤分类与预测（朴素贝叶斯）
- 实操项目 2——肿瘤分类与预测（SVM）

3.完成实验报告。

请严格基于实验报告的模板撰写实验报告。

本课程所有实验全部结束后再统一打印。

附：实操项目要求

- **实操项目 1——肿瘤分类与预测（朴素贝叶斯）**

采用朴素贝叶斯方法，对美国威斯康星州的乳腺癌诊断数据集进行分类，实现针对乳腺癌检测的分类器，以判断一个患者的肿瘤是良性还是恶性。

【实验要求】

1.导入 sklearn 自带的数据集：威斯康星乳腺肿瘤数据集（load_breast_cancer）。

2.打印数据集键值（keys），查看数据集包含的信息。

3.打印查看数据集中标注好的肿瘤分类（target_names）、肿瘤特征名称（feature_names）。

4.将数据集拆分为训练集和测试集，打印查看训练集和测试集的数据形态（shape）。

5.配置高斯朴素贝叶斯模型。

6.训练模型。

7.评估模型，打印查看模型评分（分别打印训练集和测试集的评分）。

8.模型预测：选取某一样本进行预测。（可以进行多次不同样本的预测）

参考方法：可以打印模型预测的分类和真实的分类，进行对比，看是否一致，如果一致，判断这个样本的肿瘤是一个良性的肿瘤，否则结果相反。也可以用其他方法进行预测。

9.扩展（选做）：绘制高斯朴素贝叶斯在威斯康星乳腺肿瘤数据集中的学习曲线。

- **实操项目 2——肿瘤分类与预测（SVM）**

采用 SVM 方法，对美国威斯康星州的乳腺癌诊断数据集进行分类，实现针对乳腺癌检测的分类器，以判断一个患者的肿瘤是良性还是恶性。

【实验要求】

参考实现步骤：（具体实现可以不同）

1.加载 data 文件夹里的数据集：威斯康星乳腺肿瘤数据集（数据集路径：data/data74924/data.csv）。

2.查看样本特征和特征值，查看样本特征值的描述信息。

3.进行数据清洗（如删除无用列，将诊断结果的字符标识 B、M 替换为数值 0、1 等）。

4.进行特征选取（方便后续的模型训练）。用热力图呈现 features_mean 字段之间的相关性，从而选取特征。

注：（1）热力图中，颜色越浅代表相关性越大。

（2）通过热力图找到相关性大的几个属性，每组相关性大的属性只选一个属性做代表。这样就可以把 10 个属性缩小。

5.进行数据集的划分（训练集和测试集），抽取特征选择的数值作为训练和测试数据。

6.进行数据标准化操作（可采用 Z-Score 规范化数据）。

7.配置模型，创建 SVM 分类器。

8.训练模型。

9.模型预测。

10.模型评估。

封面设计： 贾丽

地 址：中国河北省秦皇岛市河北大街 438 号
邮 编：066004
电 话：0335-8057068
传 真：0335-8057068
网 址：http://jwc.ysu.edu.cn