澳大利亚降雨预测

金笑缘 1800016611

▼ 背景

天气预报

天气预报是使用现代科学技术对未来某一地点地球大气层的状态进行预测。今天的天气预报主要是使用收集大量的数据（气温、湿度、风向和风速、气压等等），然后使用目前对大气过程的认识（气象学）来确定未来空气变化。由于大气过程的混乱以及今天科学并没有最终透彻地了解大气过程，因此天气预报总是有一定误差的。

天气预报就时效的长短通常分为三种：短期天气预报（2～3天）、中期天气预报（4～9天），长期天气预报（10～15天以上）。中央电视台每天播放的主要是短期天气预报。

▼ 数据集

本数据包含了来自澳大利亚多个气候站的日常共15W的数据

澳大利亚气象数据字段说明

Aa 特征	含义
Date	观察日期（2011～2017）
Location	获取该信息的气象站的名称
MinTemp	以摄氏度为单位的低温度
MaxTemp	以摄氏度为单位的高温度
Rainfall	当天记录的降雨量，单位为mm
Evaporation	到早上9点之前的24小时的A级蒸发量(mm)
Sunshine	白日受到日照的完整小时
WindGustDir	在到午夜12点前的24小时中的强风的风向
WindGustSpeed	在到午夜12点前的24小时中的强风速(km/h)
WindDir9am	上午9点时的风向
WindDir3pm	下午3点时的风向
WindSpeed9am	上午9点之前每个十分钟的风速的平均值(km/h)
WindSpeed3pm	下午3点之前每个十分钟的风速的平均值(km/h)
Humidity9am	上午9点的湿度(百分比)
Humidity3am	下午3点的湿度(百分比)
Pressure9am	上午9点平均海平面上的大气压(hpa)
Pressure3pm	下午3点平均海平面上的大气压(hpa)
Cloud9am	上午9点的天空被云层遮蔽的程度，这是以“oktas”来衡量的，这个单位记录了云层遮挡 天空的程度。0表示完全晴朗的天空，而8表示它完全是阴天.
Cloud3pm	下午3点的天空被云层遮蔽的程度
Temp9am	上午9点的摄氏度温度
Temp3pm	下午3点的摄氏度温度

这个特征矩阵由一部分分类变量和一部分连续变量组成，其中云层遮蔽程度虽然是以数字表示，但是本质却是分类变量。大多数特征都是采集的自然数据，比如蒸发量，日照时间，湿度等等。还有一些是单纯表示样本信息的变量，比如采集信息的地点，以及采集的时间。

因变量

RainTomorrow 目标变量

▼ **数据预处理方法**

数据概况

```
weather = pd.read_csv("weatherAUS5000.csv",index_col=0)

X = weather.iloc[:,:-1]
Y = weather.iloc[:,-1]

X.shape

# 数据类型
X.info()

#缺失值所占总值的比例 isnull().sum(全部的True)/X.shape[0]
X.isnull().mean()

np.unique(Y) #标签是二分类
```

```
#分训练集和测试集
Xtrain, Xtest, Ytrain, Ytest = train_test_split(X,Y,test_size=0.3,random_state=420) #随机抽样

#是否有样本不平衡问题？
Ytrain.value_counts()
Ytest.value_counts()
```

特征工程

▼ 处理日期特征和Rainfall

- 时间序列分析

  存在的问题：

  - 时间不连续
  - 空间不连续
- Rainfall
  - RainToday
- 日期抽出月份

▼ 处理地点特征

不同的地点因为气候不同，所以对“明天是否会下雨”有着不同的影响。

如果我们能够将地点转换为这个地方的气候的话，我们就可以将不同城市打包到同一个气候中，而同一个气候下反应的降雨情况应该是相似的。所以处理问题的核心在于获取每个城市对应的气候带。

1. 首先通过爬虫，爬取数据集的每个样本城市对应的经纬度

2. 通过澳大利亚气象局网站，下载澳大利亚每个气候带对应的主要城市和其经纬度

3. 计算数据集的每个样本的城市与各个气候带主要城市的距离，找出与气候带主要城市距离最近的一个，即为该样本对应的气候，由此便可删除地点特征。

缺失值处理

- 分类变量
- 定比变量

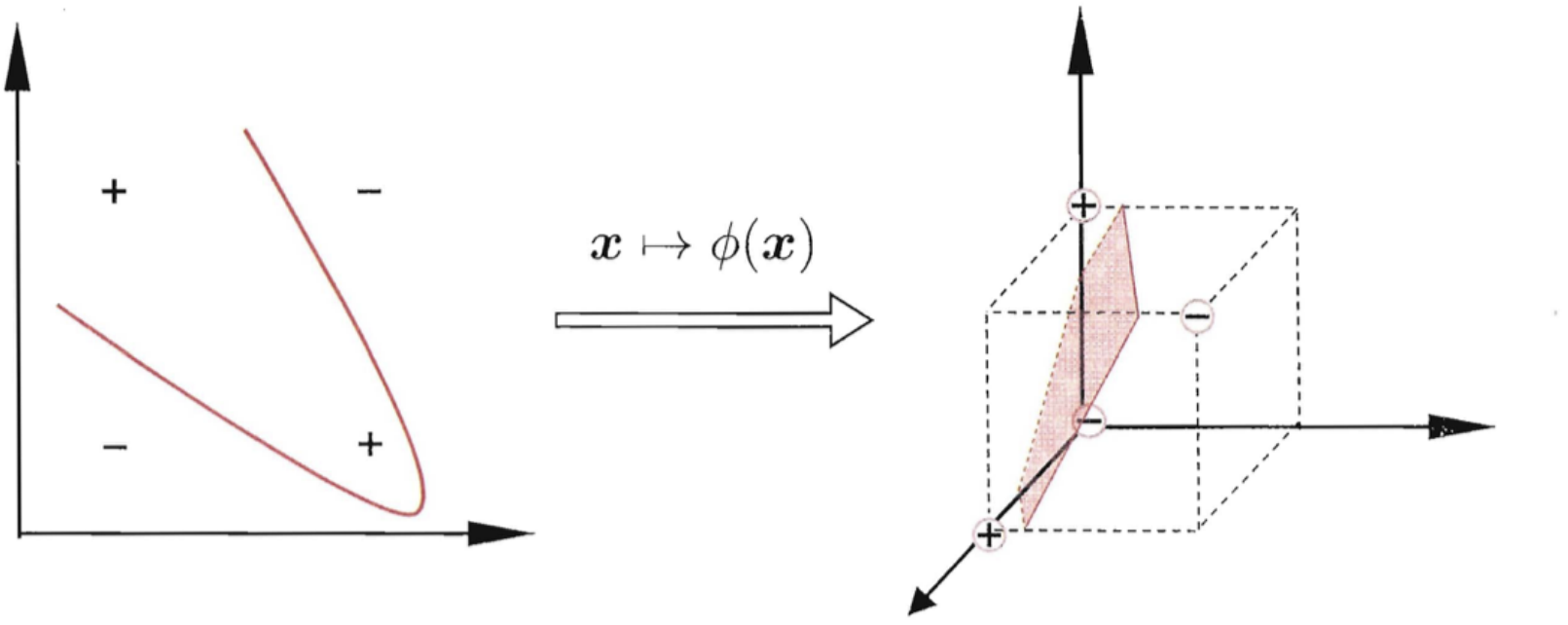
无量纲化

- 使用平均值和方差对数据进行标准化

编码

- 可识别&计算

▼ **机器学习算法**



SVM 支持向量机

- 核函数

调参数：两个调参数目标

- class_weight（少数类(雨天)的权重加大）
- 核函数（使用线性核函数）
  - 线性核函数本身参数C值

▼ **主要结果**

AUC 和 recall达到85+%

对于二分类效果一般

▼ **小结**

▼ 特征工程

时间、地点

▼ 更多模型

Model-1: Logistic Regression

Model-2: Decision Tree

Model-3: Neural Network

Model-4: Random Forest

Model-5: Light GBM

Model-6: CatBoost

Model-7: XGBoost