03-Machine Learning Fundamentals

课件：数据分析与机器学习\_2022年5月27日.pdf

试验环境链接：https://pan.baidu.com/s/1C43EqMM9oV66GLT-ikkFKg

提取码：n24t

培训录制链接：https://cisco.webex.com/recordingservice/sites/cisco/recording/playback/7594456fbf8a103abf1c00505681061a

密码：59JcU2SD

本文档所在的子目录machine learning下，有二本关于机器学习的书籍。由于机器学习也是当前非常热门的一门学科分支和职业，网上相应的书籍也非常多，可以自行搜索。

请回答下面题目

1. Python语言中关于机器学习涉及到哪些第三方工具库？请列举出来（自行在网上搜索）

Numpy, Pandas, sk-learn, Matplotlib

1. 机器学习会用到大量数据运算，特别是矩阵运算。请用Python演示一个4x4矩阵的加法，减法，乘法（包括矩阵点乘和矩阵相乘）

x=np.array([[1,2,3,4],[1,2,3,4],[1,2,3,4],[1,2,3,4]])

y=np.array([[3,2,5,4],[2,3,6,9],[4.3.7.6],[2,1,4,3]])

print(x+y)

print(x-y)

print(x@y)矩阵相乘必须满足：(m行,L列) \* (L行,n列) = (m行,n列)

1. 在Python代码的调试中，我们经常采用print输出值查看的方法，请问在Python中，print有几种格式化输出？。

|  |  |
| --- | --- |
| **占位符%格式化** | name = ‘李成文’  Print(“大家好，我是%s” %name) |
| **format方法格式化** | name = ‘李成文‘  age = 28  print(“大家好，我是{},今年{}岁”.format(name, age)) |
| **f-strting格式化** | name = ‘李成文‘  age = 28  print(f”大家好，我是{name},今年{age}岁“) |

1. 请列举出在Python中机器学习的7大步骤。

|  |
| --- |
| 获取数据 |
| 清洗与预处理 |
| 拆分数据集 |
| 特征工程 |
| 根据需求选择合适的类型 |
| 模型调优 |
| 模型的保存与加载 |

1. 机器学习是需要从大量的用户数据中找出普通规律，但是数据来源的多样化，用户数据可能存在问题，需要进行预处理，请列出出数据预处理的8大策略。

|  |
| --- |
| 抽样 |
| 数据处理 |
| 连续数据离散化 |
| 数据相关性与降维 |
| 标准化让数据落入相同范围 |
| 样本数量不够，不均衡问题 |
| 非结构化数据预处理 |
| 文本向量化 |

1. 本培训使用sk-learn机器学习库，在sk-learn中有几种常用的算法？回归类算法要对数据如何处理才能满足回归要求？

分类算法

回归算法

聚类算法

降维算法

文本挖掘算法

回归算法：试图寻找一个函数 使得参数之间的关系**拟合性**最好

1. 针对问题6的各种算法，列出算法结果好坏的评估指标。
2. 什么是归一化？什么是标准化？在机器学习的过程中，它们有什么作用？

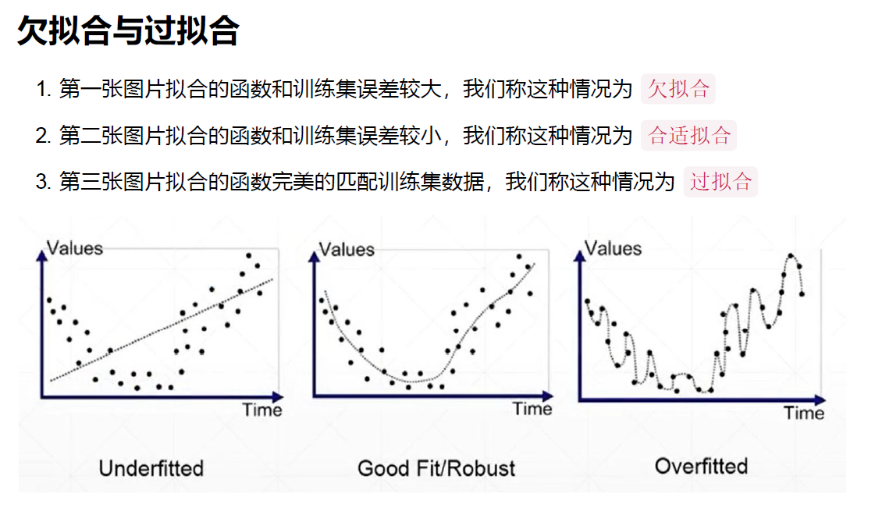
归一化：通过对原始数据进行变换，把数据映射到(0~1)之间X’=(x-min)/(max-min)

标准化：X’=(x-mean)/ δ,数据会在0左右，数据越接近0稳定性越高

1. 请使用提供的数据house.csv，使用回归类算法找出房价和其它自变量之间的关系。（按照7大步骤的顺序），并输出评估指标。

|  |  |
| --- | --- |
| 获取数据 | df=pd.read\_csv(filepath\_or\_buffer=“../data/house.csv”) |
| 清洗与预处理 | y=df[‘price’]  X=df.drop(label=”price”,axis=1) |
| 拆分数据集 | from sklearn.model\_selection import train\_test\_split  X\_train,X\_test,y\_train,y\_test =train\_test\_split(X,y,test\_size=0.25, random\_state=1) |
| 特征工程 |  |
| 根据需求选择合适的类型 | from sklearn.linear\_model import LinearRegression  lr=LinearRegression()  lr.fit(X\_train,y\_train)  print(f’权重为{lr.coef\_},偏值为:{lr.intercept\_}’)  y\_pred=lr.predict(X\_test)  from sklearn.metrics import mean\_absolute\_error  print(mean\_absolute\_error(y\_test,y\_pred)) |
| 模型调优 |  |
| 模型的保存与加载 | import joblib  joblib.dumb(value=lr, filename = “../data/lr.pkl”) |

1. 回归类算法的本质就是数据拟合，数据拟合有几种情况，请用图示说明。



1. 请使用sk-learn库提供的boston数据，分析房价和其它自变量之间的关系，并输出评估指标。

from sklearn.datasets import load\_boston

df = load\_boston()

X,y=df[‘data’],df[‘targat’]

from sklearn.preprocessing import PolynomialFeatures

pf=PolynomialFeatures(include\_bias=False, degree=2)

X=pf.fit\_transform(X)

y\_pred = lr.predict(X)

print(lr.score(X,y))

1. 重复问题10，如何提高评估指标？并用新的方法重新训练，比较评估指标是否提高？（提示：采用升维的方式，使用多项式拟合）

特征工程

from sklearn.preprocessing import PolynomialFeatures

pf = PolynomailFeatures(include\_bias=False, degree=2)

X\_train = pf.fit\_transform(X\_train)

X\_test = pf.fit\_transform(X\_test)

1. 什么是K-means算法，它属于机器学习中的哪一类？

K-means是基于距离的聚类算法，没有目标值的机器学习算法，属于机器学习里的非监督学习

1. 请使用提供的数据mobile.xls，对用户数据进行分析，并结合可视化方法，判断这些用户生活在什么社区圈？

要求：

1，以mobile.xls为数据来源

2，数据进行归一化或标准化处理

3，先对数据进行聚类

4，对每一类数据进行可视化，给出结论