**项目1**

习题

1.简述机器学习的流程

**答：1.问题定义；2.数据准备，一般包括：**（1）数据收集，（2）数据预处理，（3）数据集分割**；3.模型选择和开发；4.模型训练和调优；5.模型评估测试**

2.简述机器学习的分类

**答：1.监督学习：**监督学习又可分为“分类”和“回归”问题；**2.无监督学习；3.强化学习**

3.什么是过拟合和欠拟合？

答：过拟合(overfitting)和欠拟合(underfitting)是导致模型泛化能力不高的两种常见原因，都是模型学习能力与数据复杂度之间失配的结果。“欠拟合”常常在模型学习能力较弱，而数据复杂度较高的情况出现，此时模型由于学习能力不足，无法学习到数据集中的“一般规律”，因而导致泛化能力弱。“过拟合”则常常在模型学习能力过强的情况中出现，此时的模型学习能力太强，以至于将训练集单个样本自身的特点都能捕捉到，并将其认为是“一般规律”，同样这种情况也会导致模型泛化能力下降。

**项目2**

1. （1）对

（2）错

2. ABDE

3. train\_test\_split

**项目3**

1．判断对错。

（1）KNN算法中的*k*值最好是奇数。（ √ ）

（2）KNN算法中的*k*值越大越好。（ × ）

2．三角形

3．表3-1给出了一组鸢尾花花瓣长度（cm）、花瓣宽度（cm）和花的类别标签数据，请利用KNN算法判断花瓣长度和宽度分别为3.3和1的鸢尾花的类别。

表3-1 鸢尾花数据表

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | 花瓣长度（cm） | 花瓣宽度（cm） | 类别 |
| 1 | 1.4 | 0.2 | setosa |
| 2 | 1.7 | 0.4 | setosa |
| 3 | 1.4 | 0.3 | setosa |
| 4 | 1.5 | 0.2 | setosa |
| 5 | 1.4 | 0.1 | setosa |
| 6 | 4.7 | 1.4 | versicolor |
| 7 | 4.5 | 1.5 | versicolor |
| 8 | 4.9 | 1.5 | versicolor |
| 9 | 4 | 1.3 | versicolor |
| 10 | 4.6 | 1.5 | versicolor |

见代码”项目3 课后习题3.ipynb”

**项目4**

1．什么叫损失函数？

答：针对任何模型求解问题，最终都可以得到一组预测值，对比已有的真实值*y*，数据行数为*n*，损失函数定义如式所示：



2．如何用岭回归模型的正则化参数alpha调节权重系数？

答：增加alpha值，会降低特征变量的系数，使其趋于0，从而降低在训练集的性能，但更有助于泛化。

降低alpha值，会让特征变量的系数的限制变得不那么严格，如果用一个非常小的alpha值，那么系数的限制几乎可以忽略不计，得到的结果也会非常接近线性回归模型。

3．L2正则化和L1正则化的区别有哪些？

答：L2正则化会保留全部特征变量，只降低特征变量的系数来避免过拟合。L1正则化会把一部分系数变成0，有助于模型理解，可以突出体现模型中最重要的那些特征。

1. 表4-2是一个工作年限和对应年薪的简单数据集，请根据它绘出数据的散点图，并尝试应用线性回归模型拟合这些数据，求出直线方程。

表4-2 工作年限-年薪表

|  |  |
| --- | --- |
| 工作年限（年） | 年薪（万元） |
| 2 | 4 |
| 6 | 10 |
| 5 | 6 |
| 7 | 13 |

见代码”项目4 课后习题4.ipynb”

**项目5**

1. 判断对错

（1）逻辑回归和线性回归都是分类模型。 （× ）

（2）Logistic回归只能用于二分类。 （ × ）

2. 逻辑回归（Logistic Regression）采用（ LogisticRegression ）函数实现的。

**项目6**

1. （1）对

（2）错

1. 正态

**项目7**

1. （1）错

（2）对

1. ID3算法, C4.5算法, CART算法

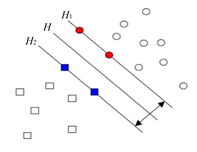
**项目8**

1．判断对错。

（1）正确

（2）错误

2．画图说明什么是“最大边界间隔超平面”。



*H*为最大边界间隔超平面。

3．SVM核函数的功能是什么？

核函数的作用就是隐含着一个从低维空间到高维空间的映射，而这个映射可以把低维空间中线性不可分的两类点变成线性可分的。

**项目 9**

1．关于聚类，下列说法错误的是（ B ）。

A．聚类属于无监督算法

B．聚类可用于数据预处理中的数据离散化

C．聚类的划分原则是样本距离最小化

D．聚类是根据数据相似度进行样本分组的方法

2．下列关于*k*-means算法初始聚类中心说法正确的是（ B ）。

A．样本中择优选出 B．样本中随机选出

C．互相距离最近的*k*个点 D．互相距离最远的*k*个点

**项目10**

1．判断对错。

（1）错误

（2）正确

2．神经网络具有哪些特性？

第一，具有自学习功能；第二，具有联想存储功能。第三，具有高速寻找优化解的能力。

3．常用的激活函数有哪些？

常用的激活函数有:sigmod函数、tanh函数、ReLU函数、softmax函数。

**项目11**

1．判断对错。

（1）错误

（2）正确

2．交叉验证的基本思想是什么？

交叉验证的基本思想是重复地利用同一份数据。将原始数据进行分组,一部分做为训练集,另一部分做为验证集,首先用训练集对分类器进行训练，再利用验证集来测试训练得到的模型,以此来做为评价分类器的性能指标。