## 监督学习

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **算法名称** | **简单原理** | **损失函数** | **优化方式** | **超参数** | **优点** | **缺点** | **应用场景** | **API** |
| 分类算法 | | | | | | | | |
| K – 近邻 | 欧氏距离，比较最近的样本类别 | null | null | K | **简单**，易于理解，易于实现。  **无需估计参数**，无需训练参数。 | 性能：懒惰算法，对测试样本分类时的计算量大，内存开销大。  必须指定K值，K值选择不当则分类精度不能保证。 | 小数据场景  多分类 | sklearn.  neighbors.  KNeighborsClassifier |
| 朴素贝叶斯  （生成模型） | 特征之间独立  联合概率与条件概率 | null | null | 无 | 朴素贝叶斯模型发源于古典数学理论，有稳定的分类效率。  对缺失数据不太敏感，算法也比较简单，常用于文本分类。  分类准确度高，速度快。 | 由于假设（各属性独立）的先验模型的原因导致预测效果不佳。 | 文本分类  多分类 |  |
| 决策树 | 信息熵  1. 信息增益  2. 增益率  3. 基尼系数 | null | null | 树深度 | 简单的理解和解释，树木可视化。  需要很少的数据准备，其他技术通常需要数据归一化。 | 决策树学习者可以创建不能很好地推广数据的过于复杂的树，这被称为过拟合。  决策树可能不稳定，因为数据的小变化可能会导致完全不同的树被生成 | 多分类 | sklearn.  tree.  DecisionTreeClassifier |
| 随机森林 | 决策树+随机有放回抽样 | null | null | 树深度  树数量  最大特征数  叶子节点的样本数 | 在当前所有算法中，具有极好的准确率。  能够有效地运行在大数据集上。  能够处理具有高维特征的输入样本，而且不需要降维。  能够评估各个特征在分类问题上的重要性。  对于缺省值问题也能够获得很好得结果。 |  | 多分类 | sklearn.  ensemble.  RandomForestClassifier |
| 逻辑回归 | 广义线性模型+激活函数 | 对数似然损失 | 梯度下降 | 正则化方式L1，L2  正则化力度C |  |  | 二分类 | sklearn.  linear\_model.  LogisticRegression |
| 回归算法 | | | | | | | | |
| 线性回归 | ｙ＝ｗｘ＋ｂ | 均方误差（最小二乘法） | 正规方程求解  梯度下降 |  | 简单 | 容易过拟合 |  | sklearn.  linear\_model.  LinearRegression　正规方程  GDRegressor 梯度下降 |
| 岭回归 | 带有L2正则化的线性回归 | 均方误差 | 梯度下降 | 正则化力度 | 能让估计参数的波动范围变小，变的更稳定。在存在病态数据偏多的研究中有较大的实用价值。 |  |  | sklearn.  linear\_model.  Ridge |

## 非监督学习

K – means

1. 步骤：
2. 随机设置K个特征空间内的点作为初始的聚类中心
3. 对于其他每个点计算到K个中心的距离，未知的点选择最近的一个聚类中心点作为标记类别
4. 接着对着标记的聚类中心之后，重新计算出每个聚类的新中心点（平均值）
5. 如果计算得出的新中心点与原中心点一样，那么结束，否则重新进行第二步过程
6. API：sklearn.cluster.KMeans

３.优缺点：

优点：迭代式算法发，直观、简单且实用。

缺点：容易收敛到局部最优解。