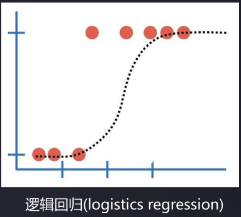
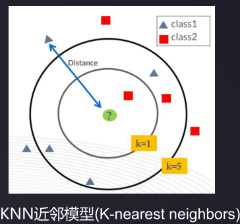
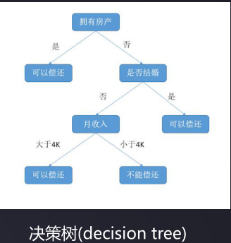
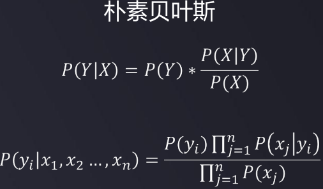
**监督学习-分类任务技术比对与总结**

**\*\*\*\*\* 没有最好的分类器，只有最适合的分类器 \*\*\*\*\***

* **Logistics regression（逻辑回归）**

1. **定义：逻辑回归根据数据特征，计算样本归属于某一类别的概率P(x)，根据概率数值判断其所属类别。其核心是基于逻辑回归方程，计算类别概率。**
2. **优点：**
   1. **输出值自然的落在0-1之间，并且有概率意义。**
   2. **参数代表每个特征对输出的影响，可解释性强。**
   3. **实施简单，高效（计算量小，存储占用低），可在大数据场景中使用。**
3. **缺点：**
   1. **本质上是一个线性的分类器，对于特征相关度高的情况效果不好。**
   2. **特征空间很大时，性能不好。**
4. **适用场景：需要较为清晰的理解每个属性对结果的影响。**

* **K-nearest neighbors（KNN近邻模型）**
  1. **定义：通过计算新数据与训练数据之间的距离，然后选取K（K >=1）个距离内的邻居进行分类判断，这些邻居的多数属于某个类，就把该新数据实例分类到这个类中。（即物以类聚，人以群分！）**
  2. **KNN距离计算方式：**
     1. **欧式距离：两点之间的直线距离（应用最多）**
     2. **曼哈顿距离：两个点在标准坐标系上的绝对轴距总和**
  3. **K值得选择：**
     1. **如果K较大，相当于使用较大领域中的训练实例进行预测，可以减小估计误差，但距离较远的样本也会对预测起作用，导致预测错误。**
     2. **如果K较小，相当于使用较小的领域进行预测，如果邻居恰好是噪声点，会导致过拟合。**
     3. **一般情况，K会倾向选取较小的值，并用交叉验证法选取最优K值。**
  4. **优点：**
     1. **简单、易于理解、易于实现，无需估计参数、无需求解训练。**
     2. **适合对稀有事件进行分类。**
     3. **在多分类场景中也不会增加训练复杂度，特别适合多分类问题。**
  5. **缺点：**
     1. **懒惰算法，对测试样本分类时的计算量大，需要扫描所有训练样本，内存开销大，评分慢。**
     2. **完全跟着数据走，没有数学模型可言，无法检查不同属性对结果的影响。**
     3. **当样本不平衡时，如一个类的样本容量很大，而其他类样本容量很小时，有可能导致当输入一个新样本时，该样本的邻居中大容量类的样本占多数。**
  6. **适用场景：需要一个特别容易向他人解释的模型的时候使用。**
* **Decision tree（决策树）**
  1. **定义：一种对实例进行分类的树形结构，通过多层判断区分目标所属类别。其本质是通过多层判断从训练数据集中归纳出一组分类规则。**
  2. **决策树求解方法（三种）：**
     1. **ID3：利用信息熵原理选择信息增益最大的属性作为分类属性，依次确定决策树的分枝，完成决策树的构造。**
     2. **C4.5：选择增益率来划分属性。从候选划分属性中找出信息增益高于平均水平的属性，在从中选择增益率最高的，从而解决ID3决策树的一个缺点——当一个属性的可取值数目较多时，那么可能在这个属性对应的可取值下的样本只有一个或者是很少个，那么这个时候它的信息增益是非常高的，这个时候纯度很高，ID3决策树会认为这个属性很适合划分，但是较多取值的属性来进行划分带来的问题是它的泛化能力比较弱，不能够对新样本进行有效的预测。**
     3. **CART：选择基尼系数来划分属性。**
  3. **优点：**
     1. **计算量小，运算速度快。**
     2. **判断步骤清晰，易于理解。**
  4. **缺点：**
     1. **忽略属性间的相关性。**
     2. **样本类别分布不均匀时，容易影响模型表现。**
  5. **适合场景：需要清晰地描述类别判断的前后逻辑时使用。（先依据哪个指标判断，后使用哪个指标…）**
  6. **参考网址：**
     1. **https://blog.csdn.net/dfly\_zx/article/details/107797864**
     2. **https://** **blog.csdn.net/dfly\_zx/article/details/107797695**
* **Naive Bayes（朴素贝叶斯）**
  1. **定义：基于训练数据集与贝叶斯概率公式，机器学习从输入到输出的概率分布，计算求出使得后验概率最大的类别作为预测输出。该分类器假设了各属性之间相互独立。**
  2. **朴素贝叶斯分类（朴素贝叶斯分类器的训练过程就是基于训练集D来估计类先验概率p(c),并为每个属性估计条件概率P(xi|c)。针对P(xi|c)计算的不同，朴素贝叶斯分类器出现了不同的变体**：
     1. **高斯朴素贝叶斯（GaussianNB）：假设p(xi|c)或p(xi|y)服从高斯分布。**
     2. **多项式朴素贝叶斯（MultinomialNB）：MultinomialNB实现了对多个分布数据的朴素贝叶斯算法，是文本分类中使用的两个经典朴素Bayes变体之一(其中数据通常表示为字向量计数)。对p(xi|c)或p(xi|y)采用平滑处理:p(xi|c)=|Dci|+α|Dc|+α∗n（n表示特征数量;α>=0）**
     3. **补码朴素贝叶斯（ComplementNB）：ComplementNB是标准多项式朴素贝叶斯(MNB)算法的一种自适应算法，特别适用于不平衡的数据集。**
     4. **类朴素贝叶斯（CategoricalNB）：CategoricalNB对分类分布的数据实施分类朴素贝叶斯算法。 它假定由索引描述的每个特征都有其自己的分类分布。对于训练集中的每个特征 XX，CategoricalNB估计以类y为条件的X的每个特征i的分类分布。**
     5. **伯努利朴素贝叶斯（BernoulliNB）：BernoulliNB实现了根据多元Bernoulli分布分布的数据的朴素贝叶斯训练和分类算法，即可能有多个特征，但每一个特征都被假定为一个二元值(Bernoulli，boole)变量。因此，这个类需要将样本表示为二进制值的特征向量；如果传递任何其他类型的数据，则BernoulliNB实例可以对其输入进行二进制化(取决于binarize参数)。**
  3. **优点：**
     1. **源于古典数学理论，分类逻辑清晰**
     2. **可清晰查看各个类别对应概率，观察数据改变的概率变化，帮助理解预测过程。**
  4. **缺点：**
     1. **在属性个数比较多或者属性之间相关性比较大时，分类效果不好**
     2. **对先验概率依赖度高，样本类别分布不均匀时，容易影响模型表现**
  5. **适用场景：数据不同维度之间相关性较小。朴素贝叶斯一般在小规模数据上的表现很好，适合进行多分类任务。**
  6. **参考网址：**
     1. [**https://www.cnblogs.com/B-Hanan/p/12871863.html**](https://www.cnblogs.com/B-Hanan/p/12871863.html)