第 10 章

1：

分类通常存在一个目标分类变量，该变量被划分为预定义的类或类别。只有分类变量是明显有序的才可以为此变量赋数字值。

1 等宽分箱法将数值型预测因子分为宽度相等的 k 个分类,其中 k 的取值由客户或者分析人员确定.2 等频分箱法将数值型预测因子分为 k 个分类,其中每个分类有 n/k 条记录,n 为记录的总数.3 通过使用一种聚类算法进行分箱,例如 k-均值聚类,以自动计算“最佳”划分.4 基于预测值的分箱.

2：

对数据集进行分析，学习哪些变量组合与目标变量有关。

数据探索答:又称质量分布图,是一种统计报告图,由一系列高度不等的纵向条纹或线段表示数据分布的情况. 一般用横轴表示数据类型,纵轴表示分布情况.

2,如何分析类型变量对预测值的影响.答:使用类型变量的条形图,列联表,对应聚类条形图,对比饼图等,通过分析图表来分析类型变量对预测值的影响.

3,如何分析数值变量对预测值的影响.答:通过使用覆盖直方图,规范化直方图来分析数值变量对预测值的影响.

4,如何对变量进行分桶.答:按照预测值对数值变量进行分类(两类、三类).

5,如何对数组变量进行变换.答: 分别对数组变量做 z-score 变化，然

后再去平均值,形成的变量具有一定的预测性

.6,相关变量如何处理.分哪两种情况.答:应该注意避免为数据挖掘和统计模型输送相关变量.1 找到相关系数为 1 或者-1 的变量(强相关),删除其中一个，看效果.2 找到相关的变量集合,采用主成分分析法(PCA)降维.

6:

权重投票可以给比较重要的信息更多的投票权，这使得投票结果更接近真实情况，但同时权重比例的分配会比较困难，如果权重分配不够完美，可能得到与真实情况相悖的结论。

简单权重投票方式:1 在执行算法前,确定 k 值,也就是说在确定新纪录的类别时需要多少记录参与.2 然后将新纪录与 k 个最近邻的记录比较,也就是说按照欧氏距离或者用户希望采用的其他距离度量方式选择 k 个与新记录最近的记录.3 一旦选择了 k 个记录,将采用简单的权重投票方式,k 个记录与新记录的距离已经不重要,每个记录具有一票.加权投票:距离近的近邻在分类决策中将比距离远的近邻获得更大的投票权,特定记录对分类新记录的影响与其和新记录的距离成反比.基于加权的投票将会大大减小平局发生的可能性.

7:

为了能够预测不常见的分类。

1.多元分类先说多元分类，在 python 中，svm 构建支持向量机可以对多分类做很好的预测。svm 是强大的模型，可以来回归，预测，分类，而根据选取不同的核函数，模型可以是线性和非线性的。

2.二元分类可以说在建模，或者其实可以更逼格的一点叫：机器学习中，二元分类是比较常见的模型。首先是 logistics 回归模型。逻辑回归应该是这行最为掌握的的模型了，spss 就可以做，而且 spss 可以直接输出各元素的权重系数。那么放到 python 中呢？很多的。第一种，同样的逻辑回归，先训练模型，score 得分，特征筛选，

在重新由有限特征值训练模型。model.predict()预测分类 model.predict_probel()输出概率 from sklearn.linear_model import LogisticRegression as LR

from sklearn.linear_model import RandomizedLogisticRegression as RLR

第二种 cart 决策树，训练模型，model.predict()预测分类 model.predict_probel()输出概率决策树是基于分类讨论，逐步细化思想的分类模型。from sklearn.tree import DecisionTreeClassifier

tree = DecisionTreeClassifier()

#建立决策树模型

第三种随机森林，训练模型，model.predict()预测分类 model.predict_probel()输出概率思想和决策树类似，精度通常比决策树高，缺点：随机性丧失了决策树的的易观性 from sklearn.ensemble import RandomForestClassifier

第四种 LM 神经网络，训练模型，model.predict()预测分类 model.predict_probel()输出概率注意 LM 神经网络是基于 tensorflow（一种基于谷歌的机器学习框架）。在 windows 需要切换后端神经网

络具有强大的拟合能力，可以用于拟合，分类等，他有很多个增强版本，如递神经网络，卷积神经网络，自编码等朴素贝叶斯分类是基于概率思想的简单有效的分类模型，能够给出容易理解的概率解释

```
from keras.models import Sequential
from keras.layers.core import Dense,Activation
```

第五种 朴素贝叶斯　朴素贝叶斯分类是基于概率思想的简单有效的分类模型，能够给出容易理解的概率解释我知道这种模型，

8：

减小最近邻的权重

提高其余数据权重

反函数

该方法最简单的形式是返回距离的倒数，比如距离 d，权重 1/d。有时候，完全一样或非常接近的商品权重会很大甚至无穷大。基于这样的原因，在距离求倒数时，在距离上加一个常量：

```
weight = 1 / (distance + const)
```

这种方法的潜在问题是，它为近邻分配很大的权重，稍远一点的会衰减的很快。虽然这种情况是我们希望的，但有时候也会使算法对噪声数据变得更加敏感。

高斯函数

高斯函数比较复杂，但克服了前述函数的缺点，其形式：

$$f(x) = ae^{-\frac{(x-b)^2}{2c^2}}$$

其中 a,b,c∈R

9：

K 取较小值，容易受到离群点和异常观察值的影响，可能导致算法过度拟合；

K 取较大值，局部呈现的比较有趣的行为将会被忽视。

如果当 K 的取值过小时，一旦有噪声得成分存在们将会对预测产生比较大影响，例如取 K 值为 1 时，一旦最近的一个点是噪声，那么就会出现偏差，K 值的减小就意味着整体模型变得复杂，容易发生过拟合；

如果 K 的值取的过大时，就相当于用较大邻域中的训练实例进行预测，学习的近似误差会增大。这时与输入目标点较远实例也会对预测起作用，使预测发生错误。K 值的增大就意味着整体的模型变得简单；

如果 K==N 的时候，那么就是取全部的实例，即为取实例中某分类下最多的点，就对预测没有什么实际的意义了；

K 的取值尽量要取奇数，以保证在计算结果最后会产生一个较多的类别，如果取偶数可能会产生相等的情况，不利于预测。

10：

使用轴伸缩可以让算法考虑对分类新纪录重要的字段，让算法不考虑那些不相关的字段.

可能存在一些相关记录,其所有重要的变量都与新记录相似,而在不重要的方面却与新记录距离甚远,导致其总体上与新记录有相当大的记录,由此在分类决策中未被考虑,因此要对 KNN 的各属性做轴伸缩.轴伸缩:确定某个字段的重要程度的问题与找到系数 Zj 并用该系数乘以第 j 各轴等价,用较大的 Zj 与更重要的变量轴关联.

11：

局部加权平均是将最近的几条记录进行加权平均，优点是系统可以不用记录所有数据，同时可以有效显示出最近邻对数据的影响。

改进的起因在于普通的线性回归努力寻找的是一个使得全局损失函数最小的模型（全局最优），这个模型对于整体而言是最好的，但是对于局部点来说，可能不是最好的。

改进的核心思想：设计损失函数时，待预测点附近的点拥有更高的权重，权重随着距离的增大而减小。

改进的具体关键环节：计算损失函数时，多乘一个权重函数 w，这个 w 保证：越靠近待测点附近权值越大，越远离待测点权值越小

定义：

给定我们要预测的样本的特征，求该样本与训练样本集之间的距离作为相应训练样本权重，距离越小，权重越大。然后模型训练和线性回

归一样

特点：

这样我们可以用局部线性函数近似非线性函数，

但是，每个预测样本都要训练一个模型，时间复杂度较高。