**Homework-05**

问题：

1.从建模目标、建模算法和评估方面，阐述机器学习建模与传统统计模型有哪些不同？

1. 统计建模是通过一个假设一个合适的数据模型，然后根据数据估计模型参数。相反，机器学习方法不是从模型模型开始，而是使用一种算法来学习响应与预测变量之间的关系。机器学习方法是假设生成过程是复杂和未知的，试图通过观察输入和响应发现主导模式。
2. 传统统计方法选择了一个模型，没有检验任何其他假设。机器学习方法首先给出的是模型集合，包括多项式模型、指数模型等，目标是从不同模型选择最佳的一个。当然，最终的结果与回归算法结果可能是一致的。
3. 统计模型重点是描述数据与结果变量之间的关系，评估模型的合理性是通过置信区间、显著性检验和其他检验对回归参数进行分析，即统计推断。机器学习涉及训练集和测试集，模型优劣通过测试集评估。
4. 统计建模更多关注变量之间的关系和意义，是可解释的。然而，机器学习强调预测性能，而不在于模型是否具有可解释性。

2. 生态学上常用树模型，包括随机森林、提升回归树。建模步骤包括：数据预处理、拆分数据集、选择特征、算法和训练模型、模型评估等。caret为各种机器学习算法提供了统一模板，加载mtcars数据集，请根据问题填空。

1）对于doubs中的鱼群数据，按照样地，计算各样地鱼类Shannon多样性指数，并新增mpg列。

install.packages("vegan")

library(vegan) # 需要加载 vegan 包来计算 Shannon 指数

library(dplyr)

# 加载 fish 数据，并添加 site（样地编号）列

fish\_data <- doubs$fish %>%

mutate(site = row\_number()) # 添加 site 变量

# 按照 site 分组，计算 Shannon 指数

shannon\_index <- fish\_data %>%

group\_by(site) %>% #分组

summarise(Shannon = diversity(select(., -site), index = "shannon")) %>%

mutate(mpg = Shannon) # 添加 mpg 列

shannon\_index

2）利用train()，训练随机森林（randomForest）模型

model\_rf <- train(mpg ~ ., data = training\_data, method = " rf \_\_") #使用随机森林模型，训练training\_data 中的数据预测map结果

3）通过trainControl()，向train()添加重采样10-fold cross-validation，以优化参数

fitControl <- trainControl(method = " repeatedcv ", number = 10, repeats = 5)#使用重复的交叉验证，10折重复验证，重复5次

model\_rf <- train(mpg ~ ., data = training\_data, method = " rf ", trControl =fitControl)

3）在train()中，增加中心化和标准化等数据预处理，提高模型精度

model\_rf <- train(mpg ~ ., data = training\_data, method = " rf ",

preProcess = c('scale', 'center'),#指定预先处理，包括中心化（减去每列均值），标准化（除去每列的标准差）

trControl =fitControl)

4）rf有mtry和tree两个参数，可以通过expand.grid()设置调优，并在train()添加

grid <- expand.grid(.mtry=c(1:10))=#构建特征数量，从1到10进行尝试

model\_rf <- train(mpg ~ ., data = training\_data, method = " rf ",

preProcess = c('scale', 'center'),

trControl =fitControl,

tuneGrid = grid)

1. 什么是递归消除选择？在caret包中，为何选择随机森林等树模型时，没有特征选择这个过程？

递归特征消除（FRE）：使用一个基模型来进行多轮训练，首先使用所有特征训练模型得到每个特征权重，并剔除拥有最小权重特征，之后再基于其余特征训练，重复前面的过程，如此往复递归，直至剩余的特征数量达到所需的特征数量。

随机森林模型会从n个特征中选择m个特征，根据每个特征对分类的贡献，从m中选择一个最优的特征做分割点，具有处理大量特征的能力，可以通过自身的机制判断特征的重要性，所以不需要特征选择。