**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？

统计建模和机器学习使用方法是相似的，最明显的例子就是线性回归，这种相似性使人们误认为两者是同一件事。统计学方法和机器学习方法存在本质上的不同，是两种完全不同的范式，两者差异可归纳如下：

（1）统计建模是通过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，试图通过观察输人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。

# 加载所需包

library(ade4)

library(vegan)

library(caret)

library(randomForest)

# --------------------------

# 1) 处理 doubs 数据：计算 Shannon 指数并新增 mpg 列

# --------------------------

data(doubs)

fish\_data <- doubs$fish

# 计算每个样地的 Shannon 多样性指数

shannon\_index <- diversity(fish\_data, index = "shannon")

# 新增 mpg 列（存储 Shannon 指数）

fish\_data$mpg <- shannon\_index

# 查看更新后的数据

head(fish\_data) # 确认新增的 shannon\_index 列存在

2）利用train()，训练随机森林（randomForest）模型

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

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)

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'),

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))

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

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

trControl =fitControl,

\_\_\_\_\_tuneGrid\_\_\_\_ = \_\_\_\_grid\_\_\_\_)

3. 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？

（1）递归消除选择：选择特征的一个较好方法是递归消除选择，递归消除特征法使用一个基模型来进行多轮训练，首先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。

（2）内置特征重要性评估：树模型在训练时自动计算特征重要性（如基尼不纯度减少量），可通过varImp()直接获取；抗高维噪声：随机森林通过自助采样和随机子空间法，天然对冗余特征和噪声具有鲁棒性。

要求：

回答问题时，将答案写为Word文档，包括填空内容，并命名为hwk-05，一周内上传代码至自己GitHub托管的homework中。