线性回归和随机森林的操作方法

首先, 将数据集分为训练集 (2018年以前) 和验证集 (2019年).

接着, 在训练集 (假设自变量为x\_train, 因变量为y\_train) 上训练模型, 即拟合模型中的参数, 使得自变量x\_train输入模型, 并根据一定的逻辑与参数进行运算后, 输出值y\_train尽可能地接近因变量.

MATLAB中拟合模型通常使用fit方法, 如线性回归函数为fitlm, 随机森林函数为fitrensemble. 使用help命令可以查看如何向这些函数赋值. 一些参数可以手动调节, 使得结果尽可能好, 但有时复杂的模型要消耗极长的运算时间.

然后, 模型返回model变量, 这就是训练好的模型. 若x\_valid是验证集中的自变量, 使用predict(model, x\_valid) 函数能够给出验证集上的预测值y\_predict.

最后, 将y\_predict与y\_valid比较即可分析模型的表现好差.

教学案例中线性回归的模型代码:

model = fitlm(x\_train, y\_train)

y\_pred = predict(model, x\_valid);

histogram(y\_pred – y\_valid, 'NumBins', 21)

教学案例中随机森林的模型代码:

model = fitrensemble(x\_train, y\_train)

y\_pred = predict(model, x\_valid);

histogram(y\_pred – y\_valid, 'NumBins', 21)

而fitrensemble包含很多可手动调节的参数, 例如:

'NumLearningCycles' 决定迭代次数,

'Method' 使用LGBoost或者Bag作为基模型,

'LearnRate' 在使用LGBoost时决定学习率 …

更多的使用方法参考MATLAB官方文档:

<https://www.mathworks.com/help/stats/fitrensemble.html#d120e334185>