工作记录9之交叉验证

2016-09-08 20:13:40

1.简单交叉验证Hold-Out Cross Validation 全部训练数据集为S，从S中随机选训练样本s作为训练集，其他的数据作为测试集testset，通过对s的训练，得到模型k，使用模型k预测testset的类别，根据testset的真正类别，求出分类正确率，以上重复n次，从中选择具有最大分类正确率的模型。 2.K折交叉验证Kfold Cross Validation 全部训练数据集为S，将S分为K个不相交的子集，假设S的训练样本数为m，那么每一个子集的样本数为m/k，分别称这些子集为{s1,s2,...,sk};每次从子集中取出一个子集作为测试集，剩下的K-1个子集作为训练集，通过对训练集的训练，得到模型ki，使用模型ki预测试集si的类别，根据si的真正类别，求出分类正确率，以上重复k次，计算k次求得的分类正确率的平均值，作为该模型分类率。 广义线性回归里Kfold交叉验证的使用： # leave-one-out and 6-fold cross-validation prediction error for # the mammals data set. data(mammals, package="MASS") mammals.glm <- glm(log(brain) ~ log(body), data = mammals) (cv.err.6 <- cv.glm(mammals, mammals.glm, K = 6)$delta) 3.Leave-One-Out Cross Validation 全部训练数据集为S，每次都只留下一个样本作为测试集，其他的所有样本作为训练集，如果样本数为m，那么需要训练m次，测试m次，计算m次求得的分类正确率的平均值，作为该模型分类率，LOOCV被证明比KfoldCV更接近无偏估计，但是计算成本过高，训练数据过多时，需要采用并行化计算减少计算时间。 广义线性回归里Leave-One-Out交叉验证的使用： # leave-one-out and 6-fold cross-validation prediction error for # the mammals data set. data(mammals, package="MASS") mammals.glm <- glm(log(brain) ~ log(body), data = mammals) (cv.err <- cv.glm(mammals, mammals.glm)$delta) (cv.err.6 <- cv.glm(mammals, mammals.glm, K = 6)$delta) # As this is a linear model we could calculate the leave-one-out # cross-validation estimate without any extra model-fitting. muhat <- fitted(mammals.glm) mammals.diag <- glm.diag(mammals.glm) (cv.err <- mean((mammals.glm$y - muhat)^2/(1 - mammals.diag$h)^2)) # leave-one-out and 11-fold cross-validation prediction error for # the nodal data set. Since the response is a binary variable an # appropriate cost function is cost <- function(r, pi = 0) mean(abs(r-pi) > 0.5)
nodal.glm <- glm(r ~ stage+xray+acid, binomial, data = nodal) (cv.err <- cv.glm(nodal, nodal.glm, cost, K = nrow(nodal))$delta) (cv.11.err <- cv.glm(nodal, nodal.glm, cost, K = 11)$delta)
http://robjhyndman.com/hyndsight/crossvalidation/里有一些证明内容。 附RStudio快捷键官方文档：
https://support.rstudio.com/hc/en-us/articles/200711853-Keyboard-Shortcuts
R最最基础的使用说明，来自gitbook： https://www.gitbook.com/book/joe11051105/r_basic/details