工作记录7之二分类准确性衡量

2016-09-08 11:14:49

机器学习中很常见的一个大类就是二元分类器。很多二元分类器会产生一个概率预测值，而非仅仅是0-1预测值。可以使用某个临界点（例如0.5），以划 分哪些预测为1，哪些预测为0。得到二元预测值后，可以构建一个混淆矩阵来评价二元分类器的预测效果。所有的训练数据都会落入这个矩阵中，而对角线上的数 字代表了预测正确的数目，即True Positive+True Nagetive。同时可以相应算出TPR（真正率或称为灵敏度）和TNR（真负率或称为特异度）。这两个指标越大越好，但实际上二者是此消彼涨的关系。除了分类器的训练参数，临界点的选择，也会大大的影响TPR和TNR。有时可以根据具体问题和需要，来选择具体的临界点。 如果选择一系列的临界点，就会得到一系列的TPR和TNR，将这些值对应的点连接起来，就构成了ROC曲线。 ROC曲线可以帮助我们清楚的了解到这个分类器的性能表现，还能方便比较不同分类器的性能。在绘制ROC曲线的时候，习惯上是使用（1-TNR）作为横坐 标，TPR作为纵坐标。 在R语言中绘制ROC曲线： # 做一个logistic回归，生成概率预测值model1 <- glm model1 <- glm(y~., data=newdata,family='binomial') pre <- predict(model1,type='response') # 将预测概率prob和实际结果y放在一个数据框中data <- data.frame(prob=pre,obs=newdata$y) # 按预测概率从低到高排序data <- data[order(data$prob),] n <- nrow(data) tpr<- fpr <- rep(0,n) # 根据不同的临界值threshold来计算TPR和FPR，之后绘制成 #for (i in 1:n) { threshold <- data$prob[i] tp <- sum(data$prob > threshold & data$obs == 1) fp <- sum(data$prob > threshold & data$obs == 0) tn <- sum(data$prob < threshold & data$obs == 0) fn <- sum(data$prob < threshold & data$obs == 1) tpr[i]<- tp/(tp+fn)# 真正率 fpr[i] <- fp/(tn+fp)# 假正率 } plot(fpr,tpr,type='l') abline(a=0,b=1) R中也有专门用来绘制ROC曲线的包，例如常见的ROCR包，它不仅可以用来画图，还能计算ROC曲线下面积AUC，以评价分类器的综合性能，该数值取0-1之间，越大越好。 library(ROCR) pred <- prediction(pre,newdata$y) performance(pred,'auc')@y.values #AUC值 perf <- performance(pred,'tpr','fpr') plot(perf) ROCR包画图函数功能比较单一，笔者比较偏好使用功能更强大的pROC包。它可以方便比较两个分类器，还能自动标注出最优的临界点。 library(pROC） modelroc <- roc(newdata$y,pre) plot(modelroc, print.auc=TRUE, auc.polygon=TRUE, grid=c(0.1, 0.2), grid.col=c("green", "red"), max.auc.polygon=TRUE, auc.polygon.col="skyblue", print.thres=TRUE)