【译文】特征选择方法导论（如何选取合适的变量）

钱亦欣 (/space/181386) 发表于 2016-12-06 16:53 1214 阅读

钱亦欣 (/space/181386)
(/space/181386)
关注
短消息 (/message/index/181386)
统计界烤肉最好，烤肉界统计最棒。

【译文】特征选择方法导论（如何选取合适的变量）

作者 SAURAV KAUSHIK

译者 钱亦欣

引言

我时常以参加竞赛的方式来磨练自己的机器学习技能，它能让你更清楚地了解自己的水平。一开始，我以为算法就是机器学习的一切，知道采用哪种模型就能走上人生巅峰。但后来我发觉自己拿衣服了，竞赛的赢家们使用的算法和其他人并无二致。而后，我认为这些人一定有很牛逼的机器，但当我发现有的top选手建模用的仅仅是macbook air后，我知道自己又错了。最终，我发现真正使得这些人脱颖而出是两件事：特征构建和特征选择。

换句话说，他们创造并选取了恰能反应数据背后逻辑的特征进入预测模型。不知算好算坏，这个技能需要持之以恒地实战， 还包含着很强的艺术性，一些人有着特别技巧，而大部分人在这方面只能苦苦挣扎。

本文我将着重介绍特征选择这一重要技巧。我会详细介绍为什么它在训练有效的预测模型中扮演着如此重要的角色。

搞起！

目录

1. 特征选择的重要性

机器学习遵循一个简单法则，你输入的是垃圾，那么得到的输出也只能是垃圾，此处的垃圾指的就是数据中的噪声。

当特征的数量很大时，这个问题就更严重了。因此你没有必要使用所有的特征来建模，只需要放入那些真正重要的，本人亲测，很多时候用特征的子集反而能取得更好的效果（不换算法）。Rohan Rao也说“Sometimes, less is better!"。

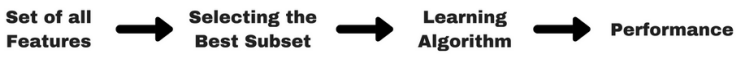
这一法则在工业级应用中也同样奏效，应用它不仅可以减少训练时间，也可以减少你所要担心的事。

使用特征选择的主要理由如下：

01. 更快的模型训练速度
02. 更低的模型复杂度和更好的解释性
03. 更高的精度（选对特征）
04. 减弱了过拟合

> 下一部分将讨论特征选择的几种方法，让我们开始吧。

2. 过滤法



过滤法时常应用于预处理阶段，不依赖于任何机器学习算法。它使用基于统计检验的得分作为筛选条件（检验特征和响应变量的相关性），这里所定义相关性带有一些主观

色彩。最基本地，你可以参考下面的表格来定义相关性。

Feature\Response	Continuous	Categorical
Continuous	Pearson's Correlation	LDA
Categorical	Anova	Chi-Square

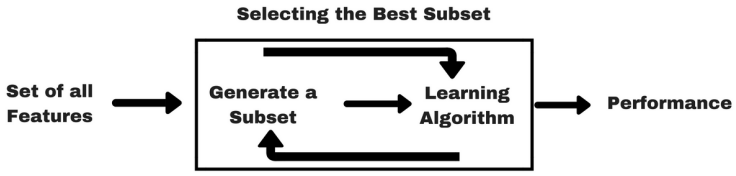
- Pearson's Correlation: 它能反应两个连续变量间的线性相关程度，取值在[-1, 1]上，计算方法如下（译者注：实际上，Pearson相关系数更多反应是两个服从正态分布的随机变量的线性相关性。如果变量虽然连续但是分布和正态分布相距比较远，建议采用非参数的spearman相关系数或kendall相关系数。）：

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

- LDA：全名线性判别分析，可以考察特征的线性组合能否区分一个分类变量。
- ANOVA：中文叫方差分析，原理和LDA类似只是它常用于特征是分类变量，响应变量是连续变量的情况下，它提供了不同组的均值是否相同的统计量。
- Chi-Square：卡方检验是基于频率分布来检验分类变量间的相关性的工具。（译者注：生物统计常用的列联表检验就是卡方检验，实际上和方差分析一毛一样，且都能通过回归形式来表示）

同时，请牢记过滤法不能减弱特征间的共线性，在训练模型前还需要针对特征的多重共线性做相应处理。

3. 包装法



包装法会仅用特征的一个子集来训练模型，并利用之前模型的结果来判断是否需要增删新的特征。这其实就是一个关于特征空间的搜索问题，但它的计算可能需要耗费大量时间空间。

常用的包装法包括了前向选择法，后向剔除法，迭代剔除法等。

- 前向选择法：这是一种基于循环的方法，开始时我们训练一个不包含任何特征的模型，而后的每一次循环我们都持续放入能最大限度提升

模型的变量，直到任何变量都不能提升模型表现。

- 后向剔除法：该方法先用所有特征建模，再逐步剔除最不显著的特征来提升模型表现。同样重复该方法直至模型表现收敛。
- 迭代剔除法：这是一种搜索最优特征子集的贪心优化算法。它会反复地训练模型并剔除每次循环的最优或最劣特征。下一次循环，则使用剩余的特征建模直到所有特征都被剔除。之后，按照剔除的顺序给所有特征排序作为特征重要性的度量。

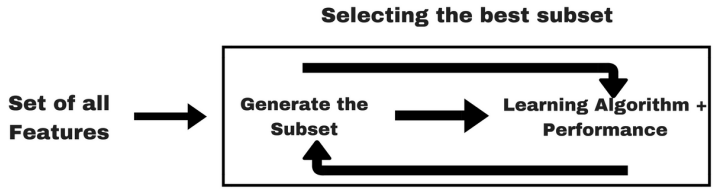
利用包装法选特征可以借助R语言中的Boruta包，它的工作原理如下：

01. 首先，它会把原来的特征打乱顺序作为新特征（称为影特征）添加到数据集中
02. 而后，基于所有特征训练随机森林模型，并评价每个特征的重要性（默认基于平均精度降低测度）
03. 每一次迭代中，该方法都会检测真实特征相对其影特征是否更重要，并移除哪些重要性差别最低的特征。
04. 最后，该算法会在所有特征都被判定为重要或无用之后停止（或者在达到给定的迭代次数后停止）

想进一步了解Boruta包的使用，请参考这篇文章(https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/)。

想了解Boruta在Python中的应用，请参考这篇文章(http://danielhomola.com/2015/05/08/borutapy-an-all-relevant-feature-selection-method/)。

4. 嵌入法



嵌入法综合了过滤法和包装法的特点，它要借助那些自带特征选择方法的算法。

最常用的嵌入法实例是LASSO和岭回归，他们的优化目标都带有惩罚项来减弱过拟合。LASSO使用L1正则，也就是对系数的绝对值大小加以惩罚。岭回归使用L2正则也

就是对系数的平方值加以惩罚。关于LASSO和岭回归的更多细节，可以参考这篇文章(https://www.analyticsvidhya.com/blog/2016/01/complete-tutorial-ridge-lasso-regression-python/)。

其他的方法则有正则树，文化基因算法和随机多项logit等。

5. 过滤法和包装法的不同

二者的主要不同包括：

- 过滤法测量特征和被解释变量的相关性，包装法则是基于模型测量特征的有效性。
- 过滤法由于不依赖于模型，速度更快。
- 过滤法基于统计检验选择特征，包装法基于交叉验证。
- 过滤法时常失效，但包装法常常被发现很有用。
- 使用包装法筛选的特征更容易导致模型过拟合。

6. 案例

让我们使用包装法来筛选变量，看看正确地使用特征子集后模型的精度会怎么变。

我们将用R来建模预测股价的涨跌。使用的数据集包含100个特征，代表股票的相关性质，响应变量y是一个二值变量，1代表股价上涨，-1代表下跌。

请点击此处 (https://drive.google.com/file/d/0ByPBn4rtMQ5HaVFITnBObXdtVUU/view)下载数据

让我们先基于所有特征训练一个随机森林模型。

```
library('Metrics')
library('randomForest')
library('ggplot2')
library('ggthemes')
library('dplyr')

# 设置随机数种子
set.seed(101)

# 导入数据
data<-read.csv("train.csv",stringsAsFactors= T)

# 检查数据维数
dim(data)

[1] 3000  101

# 将相应变量转化为因子
data$Y<-as.factor(data$Y)
data$Time<-NULL

# 把数据集划分为训练集和测试集
train<-data[1:2000,]
test<-data[2001:3000,]

# 训练随机森林
model_rf<-randomForest(Y ~ ., data = train)
preds<-predict(model_rf,test[,-101])
table(preds)

preds
 -1   1
453   547

# 检测进度
auc(preds,test$Y)

[1] 0.4522703
```

而后，为了降低模型复杂度，我们把入选特征的上限定为20，仅把随机森林模型判定最重要的20个特征纳入模型并看看预测精度。

```
importance(model_rf)

# 平均基尼下降值
##x1            8.815363
##x2           10.920485
##x3            9.607715
##x4           10.308006
##x5            9.645401
##x6           11.409772
##x7           10.896794
…
##x95           8.640581
##x96           9.368352
##x97           7.014134
##x98          10.640761
##x99           8.837624
##x100          9.914497


# 仅用20个最重要特征建模
model_rf<-randomForest(Y ~ X55+X11+X15+X64+X30
                           +X37+X58+X2+X7+X89
                           +X31+X66+X40+X12+X90
                           +X29+X98+X24+X75+X56,
                       data = train)
preds<-predict(model_rf,test[,-101])
table(preds)

preds
-1   1
218 782

# 检测精度
auc(preds,test$Y)

[1] 0.4767592
```

因此，仅仅使用20个特征我们反倒把预测精度从0.452提升到了0.476，这个例子说明，应用特征选取方法，我们不仅提高了预测精度，还

- 提升了模型的可解释性
- 降低了模型复杂度
- 减少了模型训练时间

结语

我相信本文对你而言应该是个不错的入门科普文，特征选择还有各种各样的方法和形式。我相信你将来会从特征选择中受益无穷。

注:原文刊载于Analytics Vidhya网站

链接:https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables/ (https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables/)

机器学习 (/tag/69) 特征选择 (/tag/35987)

3 推荐 ☆ 收藏 赏 感谢 分享长微博

请登录后发表评论

邮箱

密码

登 录 注册 (/register)

按时间 (/articlecomment/commentblock/273668?sort=new) 按推荐 (/articlecomment/commentblock/273668?sort=vote)

(/space/284931) google driver的测试数据无法下载，能否共享一下？ cumtlp@qq.com，谢谢

皇叔 (/space/284931) 发表于 2017-12-25 07:09:02 推荐

已发送 钱亦欣 (/space/181386) 发表于 2018-01-04 10:08:11