模型评估和模型选择

HW4　于雪　15320171151914

Apr 20,2019

一、模型性能评估概述

对于模型性能的评估，通常分为以下三步：

1. 对数据集进行划分，分为训练集和测试集两部分

2. 对模型在测试集上面的泛化性能进行度量

3. 基于测试集上面的泛化性能，依据假设检验来推广到全部数据集上面的泛化性能

二、模型评估和模型选择

模型评估是模型开发过程不可或缺的一部分。它有助于发现表达数据的最佳模型和所选模型将来工作的性能如何。在数据挖掘中，使用训练集中的数据来评估模型性能是不可接受的，因为这易于生成过于乐观和过拟合的模型。数据挖掘中有两种方法评估模型，验证（Hold-Out）和交叉验证（Cross-Validation）。为了避免过拟合，这两种方法都使用测试集来评估模型性能。

（一）验证（Hold-Out）

使用这种方法时，通常大的数据集会被*随机*分成三个子集：

1.训练集：用于构建预测模型。

2.验证集：用于评估训练阶段所得模型的性能。它为模型参数优化和选择最优模型提供了测试平台。不是所有模型算法都需要验证集。

3.测试集：之前未遇到的样本用于评估模型未来可能的性能。如果模型与训练集拟合的好于测试集，有可能是过拟合所致。

（二）交叉验证（Cross-Validation）

当仅有有限数量的数据时，为了对模型性能进行无偏估计，可以使用 k折交叉验证（k-fold cross-validation）。使用这种方法时，数据被分成 k份数目相等的子集。构建 k次模型，每次留一个子集做测试集，其他用作训练集。如果 k 等于样本大小，这也被称之为留一验证（leave-one-out）。

三、模型选择具体介绍

不管是用线性回归、逻辑回归算法，还是用神经元网络等算法，模型的选择对于达到的效果是至关重要的。相同的算法，不同的模型，结果也可能是千差万别的。所以，接下来要介绍的便是如何选择模型。

在典型的机器学习应用中，为进一步提高模型在预测未知数据的性能，要对不同的参数设置进行调优和比较，该过程称为 模型选择 。指的是针对某一特定问题，调整参数以寻求最优超参数的过程。假设要在 10 个不同次数的二项式模型之间进行选择，显然越高次数的多项式模型越能够适应训练数据集，但是适应训练数据集并不代表着能推广至一般情况。应该选择一个更能适应一般情况的模型。如果在模型选择过程中不断重复使用相同的测试数据，这样的话测试数据就变成了训练数据的一部分，模型更容易陷入过拟合。

于是将数据集分为训练集，验证集和测试集。训练数据集用于不同模型的拟合，模型在验证集上的性能表现作为模型选择的标准，测试集作为最终的性能评估。使用 60%的数据作为训练集，使用 20%的数据作为交叉验证集，使用 20%的数据作为测试集。

（一）具体的模型选择方法为：

（a）使用训练集训练出 10 个模型

Train error:

$J_{train}(\theta)=\frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)})-y^{(i)})^2$

(b)用 10 个模型分别对交叉验证集计算得出交叉验证误差（代价函数的值）

Cross validation error:

$J_{cv}(\theta)=\frac{1}{2m_{cv}}\sum_{i=1}^{m_{cv}}(h_\theta\left(x_{cv}^{(i)}\right)-y_{cv}^{(i)})^2$

(c)选取代价函数值最小的模型:min。Jcv(θ)

(d)用上一个步骤中选出的模型对测试集计算得出推广误差（代价函数的值）

Test error:

$J_{test}(\theta)=\frac{1}{2m_{test}}\sum_{i=1}^{m_{test}}(h_\theta\left(x_{test}^{(i)}\right)-y_{cv}^{(i)})^2$

（二）数据集及其划分简介

- 训练集（60%）
- 测试集（20%）
- 交叉验证集（20%）

（1）其中训练集用于算法的训练，由此可以得到很多个不同的模型，再使用交叉验证集分别测试每个模型的泛化能力，选择其中最优的模型。

最后，使用测试集来测试最优模型的泛化能力，而不直接使用上一步的交叉验证集是因为， 这个交叉验证集误差是通过对比选择出来的， 它在这个数据集上肯定是最优的， 相当于已经看到了这些数据， 所以用它来代表对未知数据的泛化能力显然不行。

（2）对于模型来说，其在训练集上面的误差称之为“训练误差”或者“经验误差”，而在测试集上的误差称之为“测试误差”。因为测试集是用来测试学习期对于新样本的学习能力的，因此可以把测试误差作为泛化误差的近似（泛化误差：在新样本上的误差)。通常更关心的是模型对于新样本的学习能力，即希望通过对已有样本的学习，尽可能的将所有潜在样本的普遍规律学到手，而如果模型对训练样本学的太好，则有可能把训练样本自身所具有的一些特点当做所有潜在样本的普遍特点，这时候就会出现“过拟合”的问题。

因此在这里通常将已有的数据集划分为训练集和测试集两部分，其中训练集用来训练模型，而测试集则是用来评估模型对于新样本的判别能力。对于数据集的划分，通常要保证满足一下两个条件：

- 训练集和测试集的分布要与样本真实分布一致，即训练集和测试集都要

保证是从样本真实分布中独立同分布采样而得

- 训练集和测试集要互斥

基于以上两个条件主要由三种划分数据集的方式：留出法，交叉验证法和自助法

留出法：最基本的抽样方法，最好使用分层抽样保证数据分布一致性，也可以做多次划分，最后返回在每次划分上测试结果的平均值，可以避免偏差。包括两个互斥集合，分别为训练集和测试集并将 2/3-4/5 样本用于训练。

具体的，留出法是直接将数据集 D 划分为两个互斥的集合，其中一个集合作为训练集 S，另一个作为测试集 T，在划分的时候要尽可能保证数据分布的一致性，即避免因数据划分过程引入额外的偏差而对最终结果产生影响。

为了保证数据分布的一致性，通常采用分层采样的方式来对数据进行采样。假设数据中有 m1 个正样本，有 m2 个负样本，而 S 占 D 的比例为 p，那么 T 占 D 得比例即为 1-p，可以通过在 m1 个正样本中采 m1*p 个样本作为训练集中的正样本，而通过在 m2 个负样本中采 m2*p 个样本作为训练集中的负样本，其余的作为测试集中的样本。 但是样本的不同划分方式会导致模型评估的相应结果也会有差别，例如如果我们把正样本进行了排序，那么在排序后的样本中采样与未排序的样本采样得到的结果会有一些不同，因此通常会进行多次随机划分、重复进行实验评估后取平均值作为留出法的评估结果。

留出法的缺点：对于留出法，如果对数据集 D 划分后，训练集 S 中的样本很多，接近于 D，其训练出来的模型与 D 本身训练出来的模型可能很接近，但是由于 T 比较小，这时候可能会导致评估结果不够准确稳定；如果 S 样本很少，又会使得训练出来的样本与 D 所训练出来的样本相差很大。另外，还可能产生偏差，若采取多次取样，则训练成本过高。优点则是划分简单

交叉验证法：将数据集划分为 k 个大小相似的数据集，注意使用分层抽样。每次使用一个小数据集做测试集，其他 k-1 个做训练集，轮流进行 k 次，最后返回的是测试结果的平均值。

具体的,k 折交叉验证通常把数据集 D 分为 k 份,其中的 k-1 份作为训练集，剩余的那一份作为测试集，这样就可以获得 k 组训练/测试集，可以进行 k 次训练与测试，最终返回的是 k 个测试结果的均值。这里数据集的划分是依据分层采样的方式来进行。对于交叉验证法，其 k 值的选取往往决定了评估结果的稳定性

和保真性。通常 k 值选取 10。与留出法类似，通常会进行多次划分得到多个 k 折交叉验证，最终的评估结果是这多次交叉验证的平均值。

当 k=1 的时候，称之为留一法，而留一法并不需要多次划分，因为其划分方式只有一种。由于留一法中的 S 与 D 很接近，因此 S 所训练出来的模型应该与 D 所训练出来的模型很接近，所以通常留一法得到的结果是比较准确的。但是当数据集很大的时候，留一法的运算成本将会非常的高。

自助法：假设有 m 个数据的数据集，每次有放回的从其中抽取一个样本，执行 m 次，最终大概有 36.8%的数据未被抽取到，当做测试集，其余当做训练集。

具体的，留出法与交叉验证法都是使用分层采样的方式进行数据采样与划分，而自助法则是使用有放回重复采样的方式进行数据采样，即每次从数据集 D 中取一个样本作为训练集中的元素，然后把该样本放回，重复该行为 m 次，这样就可以得到大小为 m 的训练集，在这里面有的样本重复出现，有的样本则没有出现过，把那些没有出现过的样本作为测试集。进行这样采样的原因是每个样本不被采到的概率为 1-1/m，那么经过 m 次采样，该样本都不会被采到的概率为 $(1-\frac{1}{m})^m$，那么取极限有 $\lim_{m\to\infty}(1-\frac{1}{m})^m\to\frac{1}{e}\approx0.368$，因此可以认为在 D 中约有 36.8%的数据没有在训练集中出现过。

这种方法对于那些数据集小、难以有效划分训练/测试集时很有用，但是由于该方法改变了数据的初始分布导致会引入估计偏差。

四、模型优化

选好模型并不意味着能够顺利地得出想要的结果，因为特征数量、样本数量、正则化参数等都会对模型最终得到的结果产生影响。应考虑影响到底有多大？影响是从那些特征可以看出？用什么方法可以改善模型？

首先，要找出能够表明模型好坏的一些特征，这样有利于分析从哪里入手去改进、优化模型。其中，欠拟合和过拟合，它们分别代表着“高偏差”、“高方差”。那么可以通过什么方法来判断高偏差和高方差呢？利用选择模型时计算的训练代价函数值和验证代价函数值画出关于多项式阶数 d 的曲线图，从图中可以很容易判断高偏差和高方差之处。样本数、特征数、正则化参数等因素影响模型的高偏差、高方差。

模型优化的方法：

对于过拟合，可以寻找更多的数据；增大正则项的强度；减小模型结构的复杂度；减少特征个数

对于欠拟合，可以减小正则项的系数；找更多的特征；寻找更多的数据；换更好的模型

五、总结

通过对模型评估和模型选择的学习，更详细的地了解了数据集的三种划分方法，分别为留出法、交叉验证法和自助法。数据量充足的时候，通常采用 留出法 或者 k 折交叉验证法 来进行训练/测试集的划分；当数据集小且难以有效划分训练/测试集时使用 自助法 ；而对于数据集小且可有效划分的时候最好使用 留一法 来进行划分，此时最为准确。

此外，选好模型并不意味着能够顺利地得出想要的结果，因为特征数量、样本数量、正则化参数等都会对模型最终得到的结果产生影响。还要通过模型优化的方法改善模型。

六、参考文献

[1]达莫达尔 N.古扎拉蒂，道恩 C.波特（2010）.经济计量学精要.机械工业出版社
[2]茅家铭.slide Model Selection and Regularization.Xiamen University
[3]https://blog.csdn.net/batuwuhanpei/article/details/51884351