# 台湾大学林轩田机器学习基石课程学习笔记16(完结) -- Three Learning Principles

作者:红色石头

微信公众号: AI有道(ID: redstonewill)

上节课我们讲了一个机器学习很重要的工具——Validation。我们将整个训练集分成两部分: $D_{train}$ 和 $D_{val}$ ,一部分作为机器学习模型建立的训练数据,另一部分作为验证模型好坏的数据,从而选择到更好的模型,实现更好的泛化能力。这节课,我们主要介绍机器学习中非常实用的三个"锦囊妙计"。

#### — Occam's Razor

奥卡姆剃刀定律(Occam's Razor),是由14世纪逻辑学家、圣方济各会修士奥卡姆的威廉(William of Occam,约1285年至1349年)提出。奥卡姆(Ockham)在英格兰的萨里郡,那是他出生的地方。他在《箴言书注》2卷15题说"切勿浪费较多东西去做用较少的东西同样可以做好的事情。"这个原理称为"如无必要,勿增实体"(Entities must not be multiplied unnecessarily),就像剃刀一样,将不必要的部分去除掉。

Occam's Razor反映到机器学习领域中,指的是在所有可能选择的模型中,我们应该选择能够很好地解释已知数据并且十分简单的模型。



上图就是一个模型选择的例子,左边的模型很简单,可能有分错的情况;而右边的模型非常复杂,所有的训练样本都分类正确。但是,我们会选择左边的模型,它更简单,符合人类直觉的解释方式。这样的结果带来两个问题:一个是什么模型称得上是简单的?另一个是为什么简单模型比复杂模型要好?

简单的模型一方面指的是简单的hypothesis h,简单的hypothesis就是指模型使用的特征比较少,例如多项式阶数比较少。简单模型另一方面指的是模型H包含的hypothesis数目有限,不会太多,这也是简单模型包含的内容。

#### simple hypothesis h

- small Ω(h) = 'looks' simple
- specified by few parameters

#### simple model ${\cal H}$

- small  $\Omega(\mathcal{H})$  = not many
- contains small number of hypotheses

其实,simple hypothesis h和simple model H是紧密联系的。如果hypothesis的特征个数是I,那么H中包含的hypothesis个数就是 $\mathbf{2}^{l}$ ,也就是说,hypothesis特征数目越少,H中hypothesis数目也就越少。

所以,为了让模型简单化,我们可以一开始就选择简单的model,或者用 regularization,让hypothesis中参数个数减少,都能降低模型复杂度。

那为什么简单的模型更好呢?下面从哲学的角度简单解释一下。机器学习的目的是"找规律",即分析数据的特征,总结出规律性的东西出来。假设现在有一堆没有规律的杂乱的数据需要分类,要找到一个模型,让它的 $E_{in}=0$ ,是很难的,大部分时候都无法正确分类,但是如果是很复杂的模型,也有可能将其分开。反过来说,如果有另一组数据,如果可以比较容易找到一个模型能完美地把数据分开,那表明数据本身应该是有某种规律性。也就是说杂乱的数据应该不可以分开,能够分开的数据应该不是杂乱的。如果使用某种简单的模型就可以将数据分开,那表明数据本身应该符合某种规律性。相反地,如果用很复杂的模型将数据分开,并不能保证数据本身有规律性存在,也有可能是杂乱的数据,因为无论是有规律数据还是杂乱数据,复杂模型都能分开。这就不是机器学习模型解决的内容了。所以,模型选择中,我们应该尽量先选择简单模型,例如最简单的线性模型。

#### 二、Sampling Bias

首先引入一个有趣的例子:1948年美国总统大选的两位热门候选人是Truman和 Dewey。一家报纸通过电话采访,统计人们把选票投给了Truman还是Dewey。经过大量的电话统计显示,投给Dewey的票数要比投个Truman的票数多,所以这家报纸就在选举结果还没公布之前,信心满满地发表了"Dewey Defeats Truman"的报纸头版,认为Dewey肯定赢了。但是大选结果公布后,让这家报纸大跌眼镜,最终Truman赢的了大选的胜利。

为什么会出现跟电话统计完全相反的结果呢?是因为电话统计数据出错还是投票运气不好?都不是。其实是因为当时电话比较贵,有电话的家庭比较少,而正好是有电话的美国人支持Dewey的比较多,而没有电话的支持Truman比较多。也就是说样本选择偏向于有钱人那边,可能不具有广泛的代表性,才造成Dewey支持率更多的假象。

这个例子表明,抽样的样本会影响到结果,用一句话表示"If the data is sampled in a biased way, learning will produce a similarly biased outcome."意思是,如果抽样有偏

差的话,那么学习的结果也产生了偏差,这种情形称之为抽样偏差Sampling Bias。

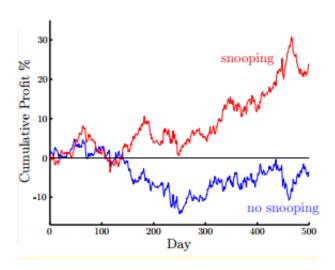
从技术上来说,就是训练数据和验证数据要服从同一个分布,最好都是独立同分布的,这样训练得到的模型才能更好地具有代表性。

#### 三、Data Snooping

之前的课程,我们介绍过在模型选择时应该尽量避免偷窥数据,因为这样会使我们人为地倾向于某种模型,而不是根据数据进行随机选择。所以, $\Phi$ 应该自由选取,最好不要偷窥到原始数据,这会影响我们的判断。

事实上,数据偷窥发生的情况有很多,不仅仅指我们看到了原始数据。什么意思呢? 其实,当你在使用这些数据的任何过程,都是间接地偷看到了数据本身,然后你会进 行一些模型的选择或者决策,这就增加了许多的model complexity,也就是引入了污染。

下面举个例子来说明。假如我们有8年的货比交易数据,我们希望从这些数据中找出规律,来预测货比的走势。如果选择前6年数据作为训练数据,后2年数据作为测试数据的话,来训练模型。现在我们有前20天的数据,根据之前训练的模型,来预测第21天的货比交易走势。



现在有两种训练模型的方法,如图所示,一种是使用前6年数据进行模型训练,后2年数据作为测试,图中蓝色曲线表示后2年的预测收益;另一种是直接使用8年数据进行模型训练,图中红色曲线表示后2年的预测收益情况。图中,很明显,使用8年数据进行训练的模型对后2年的预测的收益更大,似乎效果更好。但是这是一种自欺欺人的做法,因为训练的时候已经拿到了后2年的数据,用这样的模型再来预测后2年的走势是不科学的。这种做法也属于间接偷窥数据的行为。直接偷窥和间接偷窥数据的行为都是不科学的做法,并不能表示训练的模型有多好。

- snooping: shift-scale all values by training + testing
- no snooping: shift-scale all values by training only

还有一个偷窥数据的例子,比如对于某个基准数据集D,某人对它建立了一个模型H1,并发表了论文。第二个人看到这篇论文后,又会对D,建立一个新的好的模型H2。这样,不断地有人看过前人的论文后,建立新的模型。其实,后面人选择模型时,已经被前人影响了,这也是偷窥数据的一种情况。也许你能对D训练很好的模型,但是可能你仅仅只根据前人的模型,成功避开了一些错误,甚至可能发生了overfitting或者bad generalization。所以,机器学习领域有这样一句有意思的话"If you torture the data long enough, it will confess."所以,我们不能太"折磨"我们的数据了,否则它只能"妥协"了~哈哈。

在机器学习过程中,避免"偷窥数据"非常重要,但实际上,完全避免也很困难。实际操作中,有一些方法可以帮助我们尽量避免偷窥数据。第一个方法是"看不见"数据。就是说当我们在选择模型的时候,尽量用我们的经验和知识来做判断选择,而不是通过数据来选择。先选模型,再看数据。第二个方法是保持怀疑。就是说时刻保持对别人的论文或者研究成果保持警惕与怀疑,要通过自己的研究与测试来进行模型选择,这样才能得到比较正确的结论。

- be blind: avoid making modeling decision by data
- be suspicious: interpret research results (including your own) by proper feeling of contamination

#### 四、Power of Three

本小节,我们对16节课做个简单的总结,用"三的威力"进行概括。因为课程中我们介绍的很多东西都与三有关。

首先,我们介绍了跟机器学习相关的三个领域:

- Data Mining
- Artificial Intelligence
- Statistics

# Data Mining

- use (huge) data to find property that is interesting
- · difficult to distinguish ML and DM in reality

## Artificial Intelligence Statistics

- compute something that shows intelligent behavior
- ML is one possible route to realize Al

- use data to make inference about an unknown process
- statistics contains many useful tools for ML

#### 我们还介绍了三个理论保证:

- Hoeffding
- Multi-Bin Hoeffding
- VC

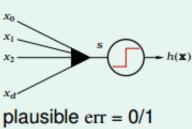
Hoeffding	Multi-Bin Hoeffding	VC
$P[BAD]$ $\leq 2 \exp(-2\epsilon^2 N)$	$P[BAD]$ $\leq 2M \exp(-2\epsilon^2 N)$	$P[BAD]$ $\leq 4m_{\mathcal{H}}(2N) \exp()$
<ul><li>one hypothesis</li><li>useful for verifying/testing</li></ul>	<ul><li><i>M</i> hypotheses</li><li>useful for validation</li></ul>	<ul><li>all H</li><li>useful for training</li></ul>

#### 然后,我们又介绍了三种线性模型:

- PLA/pocket
- linear regression
- logistic regression

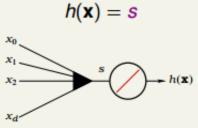
# PLA/pocket

## $h(\mathbf{x}) = \text{sign}(\mathbf{s})$



plausible err = 0/1 (small flipping noise) minimize specially

#### linear regression



friendly err = squared (easy to minimize) minimize analytically

#### logistic regression

$$h(\mathbf{x}) = \theta(s)$$

$$x_0$$

$$x_1$$

$$x_2$$

$$x_d$$

$$h(\mathbf{x})$$

plausible err = CE (maximum likelihood) minimize iteratively

#### 同时,我们介绍了三种重要的工具:

- Feature Transform
- Regularization
- Validation

### Feature Transform

$$E_{in}(\mathbf{w}) \rightarrow E_{in}(\tilde{\mathbf{w}})$$
  
 $d_{VC}(\mathcal{H}) \rightarrow d_{VC}(\mathcal{H}_{\Phi})$ 

- by using more complicated Φ
- lower E<sub>in</sub>
- higher d<sub>vc</sub>

#### Regularization

$$E_{\text{in}}(\mathbf{w}) \rightarrow E_{\text{in}}(\mathbf{w}_{\text{REG}})$$
  
 $d_{\text{VC}}(\mathcal{H}) \rightarrow d_{\text{EFF}}(\mathcal{H}, \mathcal{A})$ 

- by augmenting regularizer Ω
- lower d<sub>EFF</sub>
- higher E<sub>in</sub>

## Validation

$$E_{in}(h) \rightarrow E_{val}(h)$$
  
 $\mathcal{H} \rightarrow \{g_1^-, \dots, g_M^-\}$ 

- by reserving K examples as D<sub>val</sub>
- fewer choices
- fewer examples

#### 还有我们本节课介绍的三个锦囊妙计:

- Occam's Razer
- Sampling Bias
- Data Snooping

# Occam's Razer simple is good

Sampling Bias class matches exam

Data Snooping honesty is best policy

#### 最后,我们未来机器学习的方向也分为三种:

More Transform

- More Regularization
- Less Label

#### More Regularization Less Label More Transform bagging support vector machine decision tree neural network kernel sparsity autoencoder aggregation coordinate descent AdaBoost deep learning nearest neighbor decision stump uniform blending dual kernel LogReg large-margin prototype quadratic programming SVR matrix factorization Gaussian kernel GBDT PCA random forest RBF network probabilistic SVM k-means OOB error soft-margin

#### 五、总结

本节课主要介绍了机器学习三个重要的锦囊妙计:Occam's Razor, Sampling Bias, Data Snooping。并对《机器学习基石》课程中介绍的所有知识和方法进行"三的威力"这种形式的概括与总结,"三的威力"也就构成了坚固的机器学习基石。

整个机器学习基石的课程笔记总结完毕!后续将会推出机器学习技法的学习笔记,谢谢!

#### 注明:

文章中所有的图片均来自台湾大学林轩田《机器学习基石》课程