ML概述_AI07_20190424

一、多选题

1、机器学习是构建一个映射函数，应用场景有哪些（ ）

A、语音识别

B、图像识别

C、围棋

D、对话系统

【答案】ABCD

2、ROC曲线的横、纵坐标分别表示？( )

A、TNR, TPR

B、FNR, FPR

C、FPR, TPR

D、FPR, FNR

【答案】C

3、属于机器学习问题的有哪些？( )

A、分类

B、回归

C、聚类

D、降维

【答案】ABCD

4、ML的三要素有哪些（ ）

A、模型

B、学习准则

C、优化

D、迭代

【答案】ABC

模型：线性方法，广义线性方法

学习准则：期望风险 E(L(f(x)),y)

优化：梯度下降

5、模型选择方法有哪些？（ ）

A、留出法

B、交叉验证法

C、自助法

D、过滤法

【答案】ABC

> 1、首先评估泛化误差需要考虑一个问题，我们评价的时候需要一个训练集来训练模型，同时也需要一个测试集来评价泛化误差。所以我们要解决的第一个问题是区分训练集和测试集。
>
> A、留出法
>
> 直接将数据集分为两个互斥集合，一个作为训练集另一个作为测试集。测试集大概只占四分之一到三分之一。
>
> 而测试集和训练集的采样过程应注意分层采样，否则正例负例比例不对称，会很忙麻烦。
>
> B、交叉验证法
>
> 在保证数据一致的情况下（分层采样），将数据集分为K个大小相同的互斥子集。然后每次用K-1个子集来训练，最后一个子集用来测试。从而可以进行k次循环。我们将其称为K折交叉测试。常用的k有5,10,20。
>
> 进一步讲，当一共有m个元素，然而我们将互斥子集分为M份之后，每个子集只有一个元素，这样就变成了一个特例，我么称为留一法。留一法准确性还算好，但是计算开销很大。
>
> C、自助法
>
> 自助法一般在数据集比较小的时候用。
>
> 方法是从原始数据集中每次取一个，放到集合里面，这个集合称为训练集，经过计算，原始集合中总有占总数的0.368的数据没有被采集到。可以用原始集-训练集得到测试集用来进行测试。

D选项的过滤法是特征处理过程。

6、模型评估的方法包括哪些（ ）

A、Confusion Matrix

B、ROC

C、AUC

D、Lift

E、Gain

【答案】ABCDE

7、以下说法正确的是（ ）

A、过拟合：训练误差和泛化误差的差距太大。

B、欠拟合：是指模型不能再训练集上获得足够小的误差（训练误差太大)

C、泛化错误：一般表现为一个模型在训练集和测试集上错误率的。可以衡量一个ML模型是否可以很好地泛化到未知数据。

D、ML目的：减少泛化误差。

【答案】ABCD

8、当模型欠拟合或者过拟合了，以下哪种说法是对的？（ ）

A、欠拟合处理：增加模型复杂度

B、过拟合处理：降低模型复杂度

C、欠拟合处理：正则化处理

D、过拟合处理：增加数据集

（圈里说法；好的数据集比好的模型更重要。）

【答案】ABD

解析：

欠拟合处理：增加模型复杂度。

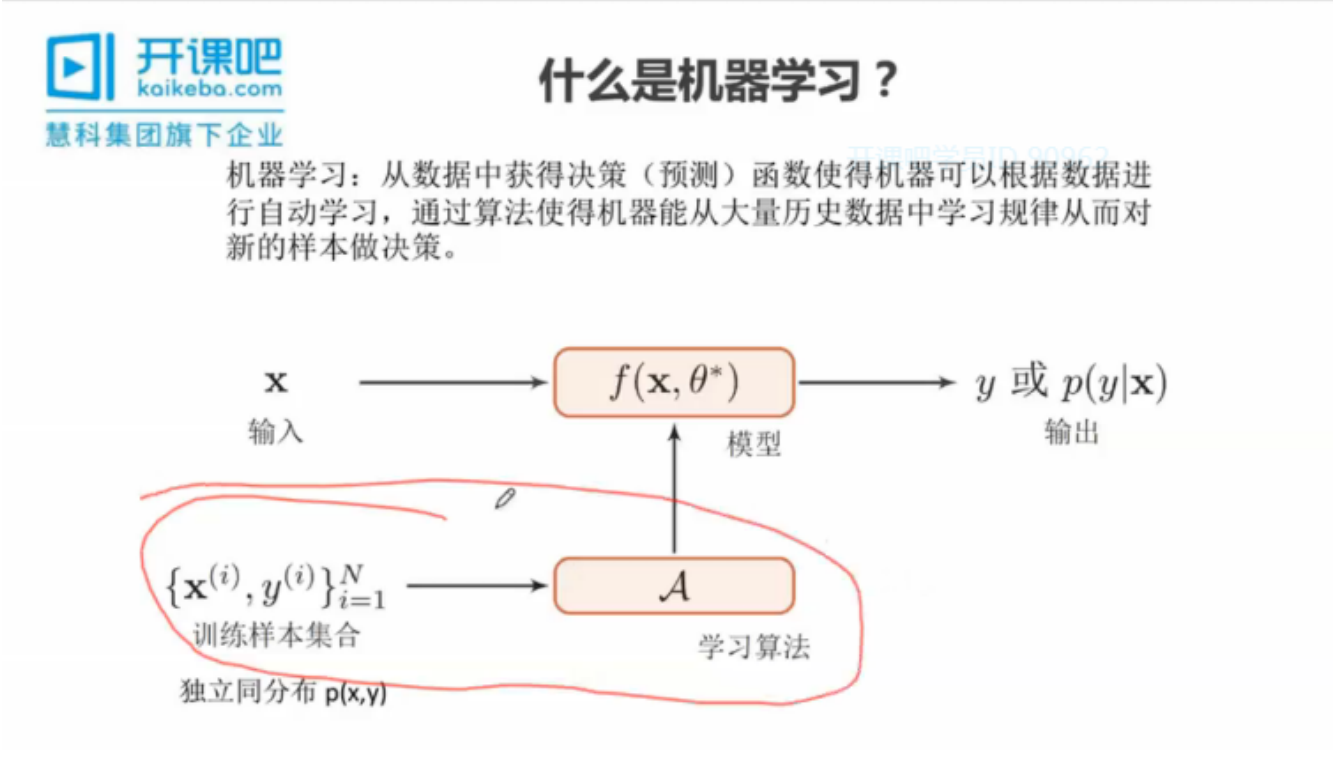
过拟合处理：降低模型复杂度 --正则化方法，损失一定的期望风险来降低模型复杂度方法/ 增加数据集 /

二、简答题

1、判断一个模型是否过拟合的依据是什么？

【答案】使用一个测试集（Test Dataset）来测试每一次迭代的参数在测试集上是否最优。

2、什么是ML？



3、留出法交叉验证（holdout cross-validation）整个过程描述。

【答案】老师课件上的图片流程图。王存存同学做的作业，非常棒，这里采用他的流程图，帮大家回顾一下，在这里谢谢王同学。

1)、训练集、验证集和测试集的划分；

将数据集的2/3作为训练集，将剩余数据集作为测试集和验证集；验证集的作用是参与监督过程和调参指导。具体如图1所示：

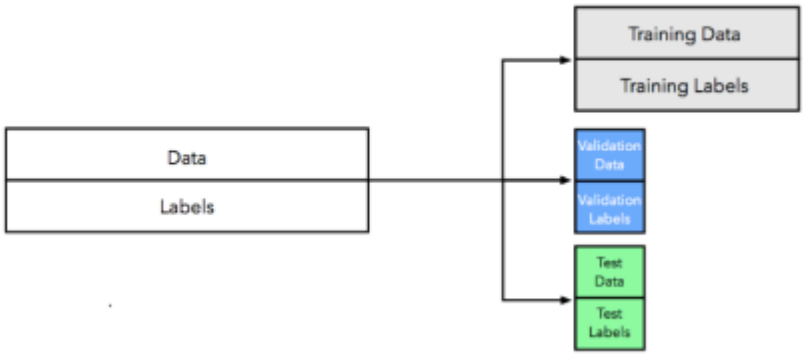


图1 训练集，测试集和测试集的划分

2)、调节超参数，获得训练模型；

在训练集中，通过不断调节超参数，从而获得训练模型。亦可采用不同的超参数，在训练集上训练同一模型。具体如图2所示：

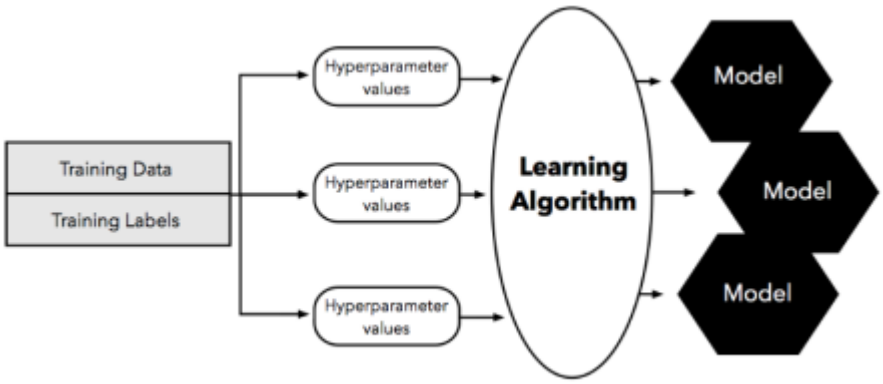


图2 调节超参数进行训练模型

3)、超参数（模型）选择；

为了防止偶然误差，在训练过程中利用验证集挑选出最优超参数，从而选择最佳的一个模型。具体如图3所示：

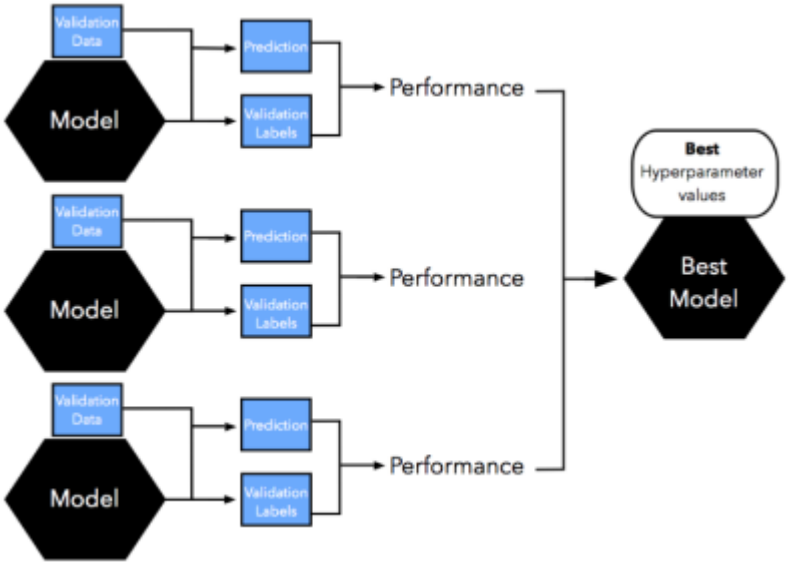


图3 验证集进行超参数选择

4)、通过最优超参数，重新获取模型；

挑选出最优的超参数，则将验证集和测试集合在一起训练最优模型。具体如图4所示：

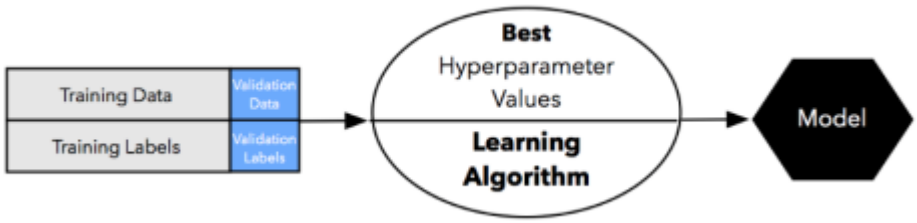


图4 验证集和测试集训练模型

5)、测试集测试

使用测试集测试模型的表现性能。具体如图5所示：

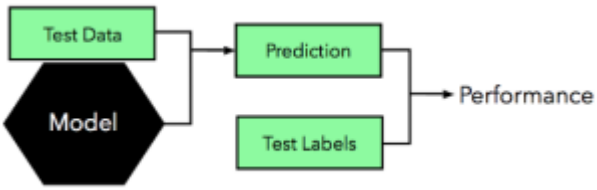


图5 模型在测试集上的表现

6)、所有数据集进行模型训练

将所有数据集放在一起，重新拿到模型中进行训练，从而获得最终的模型。具体如图6所示：

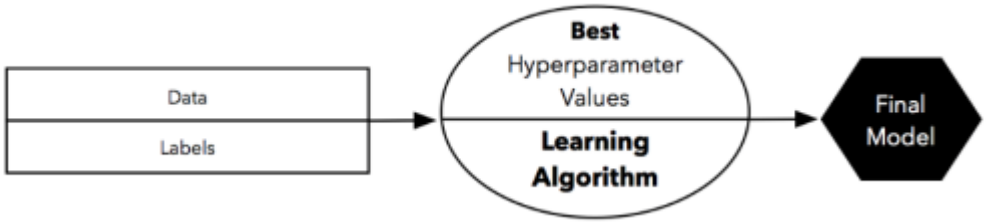


图6 所有数据集进行模型训练

4、期望误差是越小越好？

期望误差并不只是越小越好，若期望误差太小，则容易发生过拟合现象,模型泛化能力差。

5、超参数和学习参数的区别

超参数：ML模型里面框架参数，如聚类个数，正则化系数，KNN中k的选取，一般人工手动调节。

学习参数：有数据学习到的权重，不需要手动调节，由数据决定。

6、留出法交叉验证（holdout cross-validation）和k-fold交叉验证的优缺点。

1）存在一定偶然性，并不能监督过程。引入three-way holdout method.分为训练集、验证集、测试集（5%）（2/3,1/6,1/6）或（1/2,1/4,1/4）；测试集作用：参与监督过程和调参过程。

Holdout优点：是划分数据集简单,可避免了过拟合的现象

Holdout缺点：没有用所有数据来参与训练，存在一定的偶然性，并不能监督过程

2）k-fold交叉验证（用的多）：

k-fold优点：利用全部数据集更加鲁棒。

k-fold缺点：就是计算比较繁琐，需要训练k次，测试k次。

7、Boostrap在什么场合上用到，采样过程是什么样子？

集成学习中用到

有放回采样。采集到的数据，是训练集。未采集到的数据是测试集

8、写出Confusion Matrix模型评估中准确率（Accuracy）、精准率（Precision）、召回率（Recall）、F值的表达式。

真实情况	**预测结果**	**预测结果**
	正例	反例
正例	TP	FN
反例	FP	TN

准确率（Accuracy) = (TP+TN)/(TP+FP+TN+FN)

精准率（Precision）=TP/(TP+FP)

召回率（Recall）=TP(TP+FN)

F值：不偏袒，调和平均值

F1 Score=(2RecallPrecison)/(Recall+Precison)

9、ROC的全名是什么？ROC和AUC曲线横纵坐标分别表示什么？

“受试者工作特征”——ROC

AUC用于衡量“二分类问题”ML算法性能（泛化能力）

横坐标是“假正例率”（False Postive Rate, FPR）(FPR=FP/（FP+TN）），

纵坐标是“真正例率”（True Postive Rate, TPR）(TPR=TP/（TP+FN））
