AI学习笔记--机器学习

- 机器学习概述

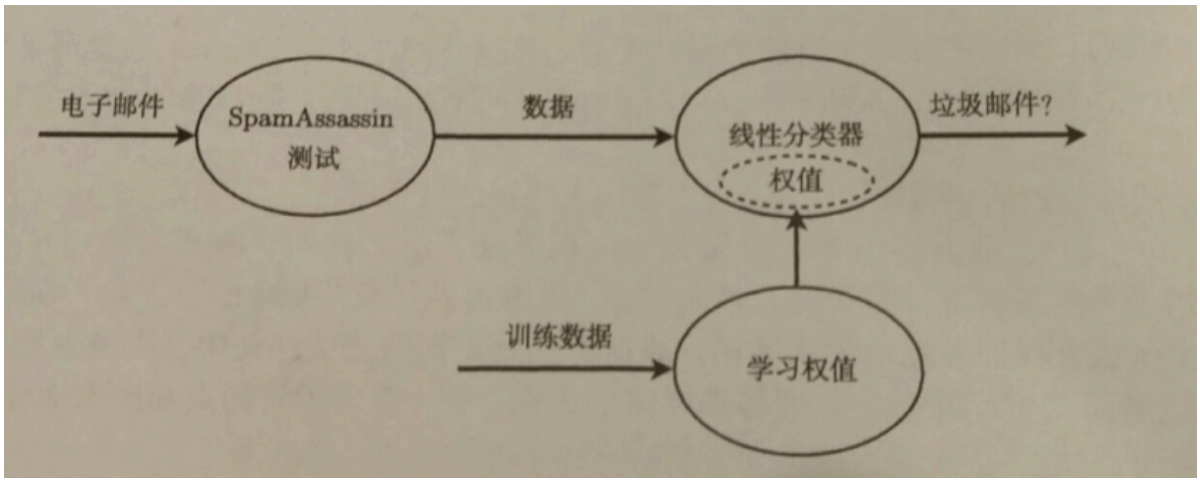
早期邮件系统中的客户端垃圾邮件过滤系统,早期都采用了Spam filter 的方式过滤垃圾邮件,这种过滤器是依据人工规则,例如:正则表达式的模式匹配方法。但是人们很快发现了，这样的做法非常不利于后续的维护、并且缺乏灵活度。从某些纬度上讲，绝对的垃圾邮件是不存在的，可能有些邮件对你来说没什么用，但是对其他人来说，可能十分有意义。如今，机器学习在邮件分类领域的应用，导致了垃圾邮件计算的广泛应用，这都使得分类变得更加可适配、灵活。

例如垃圾邮件的分类识别开源工具SpamAssassin，他能依据学习经验给出一定的权值，并且最后累加计算获取这封邮件的所属类。例如我们有4封邮件，一封垃圾邮件，3封正常邮件，用SpamAssassin做一个简单的训练数据，假定测试条件x1和x2是判定某项阈值条件的计算结果，则按照如上所说的，可以列出如下的一个表格：

邮件编号	x1	x2	垃圾邮件	4*x1+4*
1	1	1	1	8
2	1	0	0	4
3	0	1	0	4
4	0	0	0	0

以上的实例和机器学习有什么关系呢？以上说的只是一个数学问题，和学习存在什么样的关联。实际上，spamassassin通过正例和反例来学习如何识别垃圾邮件，这样和学习就有那么一点点沾边。如果能够获取更多的训练集合，那么我们就能够使得spamassassin算法获取更优异的表现，因此，我们定义机器学习是：依据以往的经验提升自身性能或者丰富自身知识的各种算法和系统的系统性研究。

在上例中，我们可以把正确标注的邮件训练集比做学习的经验，而把对应垃圾邮件的识别比做是学习算法的性能。下图是对机器学习如何在垃圾邮件归类这个历程中应用的任务原理图：



在不同的机器学习任务中，经验的获取往往具有不同的来源、形式。如对错误的纠正，实现某个目标后的奖励。此外还需要注意，机器学习未必能对某些性能得到提升，但是会在整体的知识体系当中得到提升。

- 注意：

我们已经知道，一个机器学习的过程可能存在多种解决方案，即便像垃圾邮件过滤这种简单问题上，可能也存在很多不同的解决思路。这就会导致一个问题，如何从众多的候选方法中做出一个明智的抉择，考虑这个问题的时候，我们有一个准则是当案例在算法训练的过程中所损耗的性能，往往不是我们关注的重点，正真关心的是在今后的使用场景中能不能尽可能的区分垃圾邮件和有用邮件，能否对2折进行正确的分类。要想知道垃圾邮件是否被正确分类，必须先要知道邮件的类别，如果我们已经知道了邮件的类别，则无需在进行分类决策。但是，必须记住一点：

在训练数据上取得优异性能知识手段，而非目的。实际上，如果一味求系统训练集上的性能，很容易导致存在一种貌似喜人，但是实际上存在很大隐患的现象—过拟合。

- 过拟合

过拟合是指为了得到一致假设而使假设变得过度严格。避免过拟合是分类器设计中的一个核心任务。通常采用增大数据量和测试样本集的方法对分类器性能进行评价。假设你现在在准备机器学习的考试，为了帮助你复习，教授准备了历年真题和答案共享给大家，做题时，你一味的开始尝试自己求解答案，然后与参考答案进行了比较。但是不幸的是，你过分的追求了答案的准确度，把大量的时间放在了记录参考答案上。如果即将到来的考试是历年的题目，你一定能够取得非常优异的成绩，但是如果出现了相同的知识点，只是不同的形式，那么你势必会因为准备不足，取得不了最好的成绩。这种情况就是一种过拟合的情况，机械记忆学习获取的知识，不能用在了新的考题上。

在程序的世界里，过拟合一般出现的条件是：

1. 建模样本选取有误，如样本数量太少，选样方法错误，样本标签错误等，导致选取的样本数据不足以代表预定的分类规则；

2. 样本噪音干扰过大，使得机器将部分噪音认为是特征从而扰乱了预设的分类规则；
3. 假设的模型无法合理存在，或者说是假设成立的条件实际并不成立；
4. 参数太多，模型复杂度过高；
5. 对于决策树模型，如果我们对于其生长没有合理的限制，其自由生长有可能使节点只包含单纯的事件数据(event)或非事件数据(no event)，使其虽然可以完美匹配（拟合）训练数据，但是无法适应其他数据集。
6. 对于神经网络模型：a)对样本数据可能存在分类决策面不唯一，随着学习的进行,，BP算法使权值可能收敛过于复杂的决策面；b)权值学习迭代次数足够多(Overtraining)，拟合了训练数据中的噪声和训练样例中没有代表性的特征。

常用的解决思路有：

1. 在神经网络模型中，可使用权值衰减的方法，即每次迭代过程中以某个小因子降低每个权值。
2. 选取合适的停止训练标准，使对机器的训练在合适的程度；
3. 保留验证数据集，对训练成果进行验证；
4. 获取额外数据进行交叉验证；
5. 正则化，即在进行目标函数或代价函数优化时，在目标函数或代价函数后面加上一个正则项，一般有L1正则与L2正则等。

- 推广性

推广性可能是机器学习过程中最基础的概念。机器学习是一门能够自适应你个人情况的技术，可以对个人的参数，训练，学习。从而找到推广到一个函数来判别，预测你未来的动作。在解决一个问题的过程中，可能会出现过拟合等难以避免的情况，可以有很多种类的分类器算法对同一个应用场景处理，每个分类器对应的性能也不尽相同。随着分类器的不断训练，我们可以避免很多噪点，计算出一条合理的分类边界函数，适应之后的场景分类，或者增设某些条件，比如上提到的，4*x1+4*x2>5判别条件，我们还可以增加判别条件为4*x1+4*x2<1 。

- 概率学

比如一封邮件在SpamAssassin中给予的测定条件出现了”伟哥”、”Free Ipad”或者”affirm your account details”等等词汇，这样的词汇都可以成为垃圾邮件的指示信号，但是其他的词汇指向了普通邮件。为此，很多邮件分类系统都才用了文本分类技术。通常这些技术都会维护一个可作为垃圾邮件和普通邮件指示信号的词汇、短语词典。假如我们的4封邮件和1封垃圾邮件中，同时发现了”伟哥”这个词汇。那么当新邮件来的时候，我们会推断垃圾邮件和普通邮件的比例是1：4.用概率学定律，垃圾邮件概率为0.2，普通邮件概率为0.8。

概率涉及到的事件结果的随机变量，这里的事件通常是以假设的形式出现的，因此需要用估计的方法来计算出事件本身的出现概率。例如，假定某候选人选举，在总得样本中抽取100人，有43人支持该名候选人，那么按照概率学理论，可以估算出整体有43%的选民支

持该名候选人。

条件概率模型P（A|B）刻画了当事件B发生时，事件A产生的概率，例如，男性选民和女性选民的支持率可能不同。若记P(M)为抽到公民支持候选人的概率，则有女性P(Woman)= P(PM,woman)/P(woman)，其中P(PM,woman)是支持该候选人并且是女性所占的比例,而P(woman)则是抽到女性被调者的概率。一个事件的概率是该事件发生和不发生的比值，如果一个事件发生的概率是0.8，那么发生的概率是4：1。比较有用的公式是贝叶斯公式。

- 贝叶斯公式

贝叶斯定理也称贝叶斯推理，早在18世纪，英国学者贝叶斯(1702～1763)曾提出计算条件概率的公式用来解决如下一类问题：假设H[1],H[2]…,H[n]互斥且构成一个完全事件，已知它们的概率P(H[i]),i=1,2,…,n,现观察到某事件A与H[1],H[2]…,H[n]相伴随机出现，且已知条件概率P(A|H[i])，求P(H[i]|A)。贝叶斯公式（发表于1763年）为：

$$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{\sum_{j=1}^{n} P(B_j)P(A|B_j)}$$

贝叶斯公式为利用搜集到的信息对原有判断进行修正提供了有效手段。在采样之前，经济主体对各种假设有一个判断（先验概率），关于先验概率的分布，通常可根据经济主体的经验判断确定（当无任何信息时，一般假设各先验概率相同），较复杂精确的可利用包括最大熵技术或边际分布密度以及相互信息原理等方法来确定先验概率分布。 [1]

- 联合概率

实际的情况往往并没有数学公式那样简单，假定我们收到了6封普通邮件和一封垃圾邮件，这意味着下一封邮件有1:6，如果我们知道有关键词”伟哥”，由于这个词汇在垃圾邮件中出现的概率多4倍。借助贝叶斯公式，我们可以把这两个的概率相乘，最后得到垃圾邮件的概率是0.4，及（1：6）*（4：1）=（4：6）。因此，如果一封邮件出现了这个词汇，我们就把他定义为垃圾邮件是不科学的。

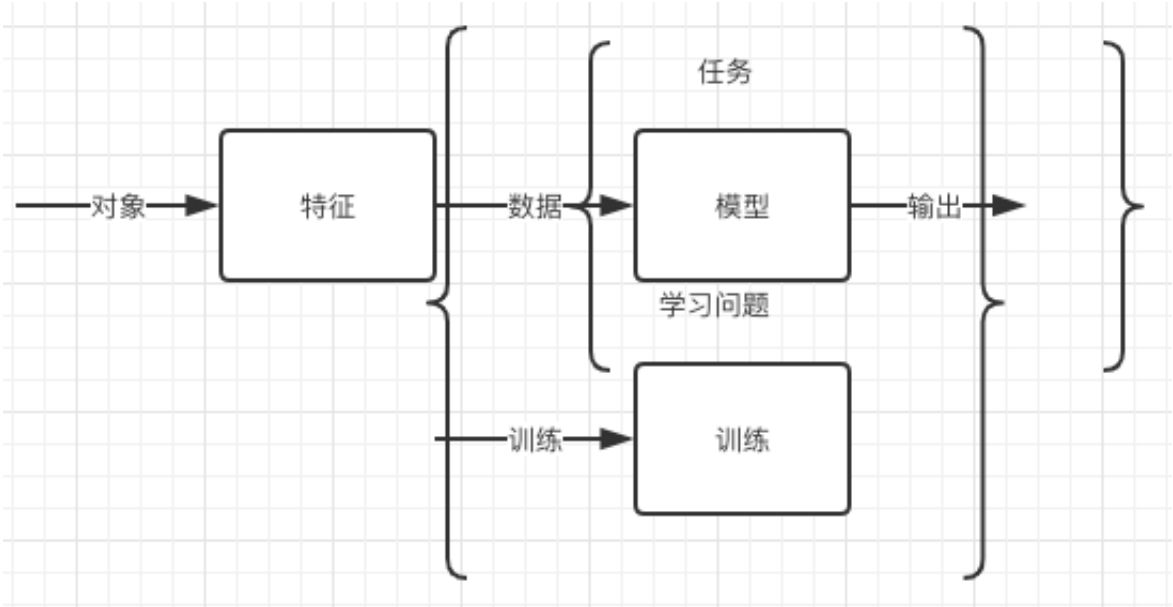
需要明确对两种独立事件的融合，其中一种证据是刻画一个垃圾邮件出现的概率，另一个证据是出现关键词的概率。这两个条件相互制约，相互牵制。两者条件不足以把邮件定义为垃圾邮件，只能是把邮件定义为普通邮件的概率从1/7降低到了6/10。

贝叶斯公式分类的好处是进一步的证据可以在原有的基础上使用，比如出现了一个新的关键词「blue pill」,普通邮件和垃圾邮件出现的几率是3:1,假定一封邮件出现了这两个关键词特征，综合二者得到的几率是4：1乘以3：1，得到是12：1，这足以压倒垃圾邮件1：6的出现几率。这样该邮件作为垃圾邮件的总几率为2：1，而其为垃圾邮件的概率从原先的0.4增加到了0.67。

联合概率使得处理大量高纬度随机变量成为可能。实际上一个高纬度的贝叶斯垃圾邮件过滤器可以包括1000个分类词汇以上。所以，我们不一定需要在专业的领域亲手打造，更多的是在较大规模的数据集合中借助分类器来找到哪些特征的重要性，以及组合的最优解。

- 总结

可见任务、模型以及特征是机器学习领域的三大原料。下图可以形象的展现这三着之间的联系：



要完成一项任务，需要建立从特征描述的数据到输出的恰当映射（模型），学习任务的中心就是如何训练数据来获取相对正确、合理的映射。结合上述的思想，可以用一段话来概括机器学习的主要内容：

机器学习所关注的问题是使用正确的特征来构建正确的模型，以便完成特定的任务。

我们往往需要用不同的组合搭配，来着手学习原料，构建模型，最后输出一个比较优化的结果。