# 机器学习

6.1 介绍

机器学习算法是解决来自许多学科的问题的通用工具，没有详细的领域特定的知识。 它们已被证明在大量语境中非常有效，包括计算机视觉，语音识别，文档分类，自动驾驶，计算科学和决策支持。

核心问题：机器学习的核心问题是从带有标签的数据中学习出一个好的分类规则。这个问题由兴趣域X，（称它为实域空间，比如邮件信息，病人病例等等），而一个分类任务，比如将电子邮件进行垃圾分类识别哪些是垃圾邮件，哪些不是垃圾邮件或者确定哪些患者将需要药物治疗。我们通常将我们的实域空间表示为X={0,1}d或者X = Rd ，对应d个布尔值或者实值特征描述的数据。电子邮件消息的特征可以是存在或不存在各种类型的单词，而用于患者记录的特征可以是各种医学测试的结果。为了执行学习任务，我们的学习算法给定一组标记的训练示例，其是X中的点以及它们的正确分类。该训练数据可以是电子邮件消息的集合，每个标记都可以指出垃圾邮件或非垃圾邮件，或一组患者，每个患者标记为是否对给定的治疗反应良好的情况。我们的算法旨在使用训练示例以产生将在新数据上执行的分类规则。机器学习的一个关键特征，它区别于其他算法任务，我们的目标是泛化：使用一组数据，以便对我们还没有看到的新数据执行良好。我们专注于二分类，其中感兴趣的域中的项目被分类为两个类别，如在上面的医疗和垃圾邮件检测示例中。

怎么样学习：一个高效的解决这个问题的办法是使用许多的算法在训练数据上可以找到一个“简单”的规则同时有一个好的表现在训练数据之上。比如在邮件分类上，我们可以找到一些高度的指示性的词语，使得训练数据中的每个垃圾邮件都有这个词汇，而不是垃圾邮件的则一个也没有。在这样的情况下，规则就是“如果有这写词语就是垃圾邮件，没有就不是”这就是一个简单的规则且可以在训练数据上有则良好的表现。或者我们可以发现一种加权的方式来使用正负加权的方式来对这些词语进行加权，使得电子又见小夕中的词语的总加权在训练数据上是正的，在垃圾邮件是上负的。只要训练数据代表了以后需要预测的数据，那么我们有在训练数据上表现良好的规则，同时也会在未来预测的数据上有良好的表现。我们需要精确地描述我们的意思是“简单”，以及它对于训练数据是否代表未来数据意味着什么。事实上，我们将看到几个复杂性的概念，包括位计数和VC维度，这将允许我们做这种形式的数学语句。这些语句可以被视为形式化的Occam’s razor的直观哲学概念。

形式化问题：为了形式化学习问题，假设实例空间上存在一些概率分布D，使得（a）是随机从这个分布D中取出来的独立的点组成的集合S，而(b)则是我们需要预测的新的点，也是我们认为我们的训练数据未来数据的代表。让C\* 表示为目标概念集，它是X中的正的二分类。比如说，医疗患者中对药物治疗有良好反应的所有患者的集合，或者垃圾邮件中所有垃圾邮件的集合。所以，在我们的训练集的每一个点都是根据它是否属于C∗和我们的目标标记是产生一组h⊆X，称为我们的假设，这是接近C∗相对于分布D. H真正的错误是 在 表示D上面的对称差，和概率质量。换个说法就是说h的真实误差是其对从D随机抽取的数据点进行不正确分类的概率。我们的目标是让h的正式误差变小。其中h的训练误差为 ，他是集合S中的一部分，这里的h和C\*是不一样的。那么.训练误差也称为经验误差。请注意，即使是由随机从D点，一个假设h具有较低的训练误差甚至完全等同C∗的训练样本，它是可能的，但有很高的真实误差。这就是所谓的过度拟合训练数据。例如，假设h，就是在正例上市，这相当于一个规则，将训练样本和预测积极的一个例子，当且仅当它已经在训练样本表现的很好时，训练误差为0。然而，这种假设可能会有真正的高误差，因此将高度过度拟合训练数据。更普遍的是，过度关注因为算法通常会优化训练样本。学习和分析算法的设计，我们将解决过度拟合的问题。

为了能够正式的分析过拟合问题，我们引入一个假设类的概念，也被称为一个概念类或集合系统。一个关于H的假设，H是X子集的集合，称为假设。例如，区间在 类是集从 中收集的，在的线性分隔的类的集合为：



也就是说，它是集合的所有集合在RD是线性可分的补充。在这种情况下，X是在平面F也就是说，它是集合的所有集合在 是线性可分的补充。比如X是平面上的4个点的集合{(-1,-1),(-1,1),(1,-1),(1,1)}, 线性分隔符的类包含X的 个可能子集中的14个。给出一个假设H和训练数据集合S，我们通常在算法上要做的是在H中找到最接近于c的假设，为了解决过拟合问题，我们认为如果S与H的一些性质相比足够大，则以高概率所有 具有接近其真实误差的训练误差，使得如果我们找到训练误差低的假设，我们可以有信心 它的真实误差也会很低。

在给出我们的第一个结果之前，我们注意到将假设表示为{-1,1}的值可以作为指示器



这时的h的真实误差为 ，并且训练误差为

6.2过拟合和一致收敛性

我们现在提出两个方式，解释如何防止过度拟合。给定一个分类假设H，第一个结果只需要给定一个就可以了，所以只要我们的训练数据集比 大，那么我们产生的 将不会有0的训练误差，而真实误差会比 大。这意味着在高概率下，我们的算法发现与训练数据上的目标假设是一致的任何假设将具有低的真实误差。第二个结果表明，如果训练数据集比，那么就在对于任何关于H的假设中不可能会产生训练误差和真实误差超过，这意味着如果我们在训练误差较低的H中找到假设，我们可以确信，即使其训练误差不为零，其真实误差也将很低。

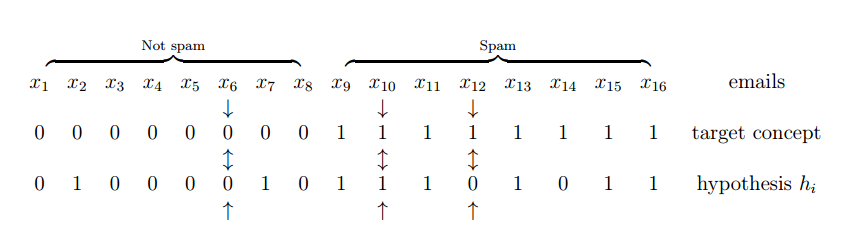
基本思想如下。 如果我们考虑h具有大的真实误差，并且根据D随机选择元素 ，则存在x将属于对称差的合理机会。 如果我们选择足够大的训练样本S，每个点从D中独立选取X，则S与完全不相交的机会将非常小。 这仅仅是针对单个假设h，但是当H是有限的时候，我们现在可以对大的真实误差的所有应用联合约束。 我们在下面形式化。

定理6.1 产生一个H假设类，并且让 ，如果S的训练集大小为



是从分布D绘制的，则在H中每h的概率大于或等于，具有真误差 具有训练误差 。等效地，以大于或等于 的概率，训练误差为零的每个 具有小于的真误差。

证明：让 …是从H中选出来的，且他们的真实误差大于或者等于。这是假设使我们不想要输出的。考虑选择一个样例S

 图6.1：假设 中有1/4是真的。因此，训练集在为|S|，那么是假设正确的可能性为 。