**统计学习**

**（Statistical Learning）**

# 统计学习理论基础

机器学习不仅要从已有的数据中学习得到规律，从而较好地解释已知的实例，而且更重要的是对还无法观测到的未来数据做出正确的预测和判断[1]。

设输入变量与输出变量遵循某一未知的联合概率，机器学习的目的就是在函数集中中求一个最优的函数，使得期望风险最小化，即



式中：是函数对预测造成的损失。

## 统计学习三要素

统计学习方法由三要素构成：方法=模型+策略+算法

## 统计学习方法的分类

统计学习分为监督学习，非监督学习，半监督学习和强化学习

监督学习的模型可以是概率模型或非概率模型，由条件概率分布或者决策函数表示。

## 损失函数

常见的损失函数有以下几种：

1. 0-1损失函数（模式识别问题）



1. 平方损失函数（函数逼近问题）



1. 对数损失函数（概率密度估计问题）



式中：

——为分类的个数；

——一个二值指标，表示类别是否是输入实例的真实类别；

——模型或分类器预测输入实例属于类别的概率；

对于二分类问题，上式变为：



经验风险最小化（Empirical Risk Minimization，ERM）：



经验风险与期望风险之间的联系与区别：

经验风险是局部的，基于训练集所有样本点损失函数最小化的。

期望风险是全局的，是基于所有样本点的损失函数最小化的。

经验风险函数是现实的，可求的；

期望风险函数是理想化的，不可求的；

对于模式识别问题，采用0-1损失函数，经验风险最小化就是最小化训练数据集的错误率；

对于函数逼近问题，采用平方损失函数，经验风险最小化就是最小二乘法；

对于概率密度估计问题，采用对数损失函数，经验风险最小化等价于极大似然估计。

只考虑经验风险的话，会出现过拟合的现象，过拟合的极端情况便是模型对训练集中所有的样本点都有最好的预测能力，但是对于非训练集中的样本数据，模型的预测能力非常不好。

结构风险是对经验风险和期望风险的折中。在经验风险函数后面加一个正则化项（惩罚项）。



表示的是是模型的复杂度。结构风险可以这么理解：

经验风险越小，模型决策函数越复杂，其包含的参数越多，当经验风险函数小到一定程度就出现了过拟合现象。也可以理解为模型决策函数的复杂程度是过拟合的必要条件，那么我们要想防止过拟合现象的方式，就要破坏这个必要条件，即降低决策函数的复杂度。也即，让惩罚项最小化，现在出现两个需要最小化的函数了。我们需要同时保证经验风险函数和模型决策函数的复杂度都达到最小化，一个简单的办法把两个式子融合成一个式子得到结构风险函数然后对这个结构风险函数进行最小化。

当所求的模型是概率模型的条件概率分布，损失函数是对数损失函数时，经验风险最小化就等价于极大似然估计。

当所求的模型是概率模型的条件概率分布，损失函数是对数损失函数时，模型复杂度由模型的先验概率表示时，结构风险最小化就等价于最大后验概率。

传统的统计学方法是在数据集的数目足够大的前提下进行研究的

统计学习理论是研究在有限数据集的情况下基于数据的机器学习问题

过拟合现象是因为试图用一个十分复杂的模型取拟合有限的样本，导致丧失了推广能力。

1971年，V。 Vapnik And A。 Chervonenkis在论文“On The Uniform Convergence Of Relative Frequencies Of Events To Their Probabilities”中提出VC维的概念。

1974年，V。 Vapnik提出了结构风险最小化原则。

1993年，Corinna Cortes和Vapnik等人提出了支持向量机(Support Vector Machine)。

# 统计学习，机器学习和数据挖掘的区别与联系

这三个领域或学科交叉和重叠部分很多，统计学习，机器学习和数据挖掘的概念一直有或多或少的混淆，希望同样有我这样的困惑的朋友读完下面的文字能够清晰一些。

　　数据库提供**数据管理技术**，机器学习和统计学提供**数据分析技术**。由于统计学界往往醉心于理论的优美而忽视实际的效用，因此，统计学界提供的很多技术通常都要在机器学习界进一步研究，变成有效的机器学习算法之后才能再进入数据挖掘领域。从这个意义上说，**统计学主要是通过机器学习来对数据挖掘发挥影响，而机器学习和数据库则是数据挖掘的两大支撑技术**。从数据分析的角度来看，绝大多数数据挖掘技术都来自机器学习领域。但能否认为数据挖掘只不过就是机器学习的简单应用呢？答案是否定的。一个重要的区别是，传统的机器学习研究并不把海量数据作为处理对象，很多技术是为处理中小规模数据设计的，如果直接把这些技术用于海量数据，效果可能很差，甚至可能用不起来。因此，数据挖掘界必须对这些技术进行专门的、不简单的改造。例如，决策树是一种很好的机器学习技术，不仅有很强的泛化能力，而且学得结果具有一定的可理解性，很适合数据挖掘任务的需求。但传统的决策树算法需要把所有的数据都读到内存中，在面对海量数据时这显然是无法实现的。为了使决策树能够处理海量数据，数据挖掘界做了很多工作，例如通过引入高效的数据结构和数据调度策略等来改造决策树学习过程，而这其实正是在利用数据库界所擅长的数据管理技术。实际上，在传统机器学习算法的研究中，在很多问题上如果能找到多项式时间的算法可能就已经很好了，但在面对海量数据时，可能连算法都是难以接受的，这就给算法的设计带来了巨大的挑战。

　　总结：

　　1）统计学习：是其它两门技术的基础，更偏重于理论上的完善；

　　2）机器学习：是统计学习对实践技术的延伸，更偏重于解决小数据量的问题提供算法技术支撑；

　　3）数据挖掘：更偏重于大数据的实际问题，更注重实际问题的解决，包括真实数据的数据清洗，建模，预测，等操作。

不管是传统的统计分析，还是现在火热的机器学习和深度学习，无数的数据分析员、数据工程师和数据科学家们其实都在追求一个终极梦想：找到或者能够无限接近一个“上帝函数” - 一个能够完美利用数据解决现实各种问题的模型或者方法。

这样的上帝函数有什么特点呢？我们说学习的目的无外乎两点：理解和预测，所以我们期盼它能拥有两个能力或者说两个必要条件：解释因果和预测未来。我需要知道问题的本质是什么，形成我们看到的结果的推动力究竟是什么（我是谁？我从哪里来？），这是因果分析；未来事物的发展会怎么样，是否按照某一个模式（我要到哪里去？)，这是预测分析。回到我们人类现在所处的阶段，对求解上帝函数的渴望，使得我们依靠计算硬件的发展和数学理论的支持，能越来越从其中一个方面逼近完美模型。

谈到统计推断（或者跟随机器学习的概念，叫统计学习），绕不开Hastie，Tibshirani和Friedman编写的《The Elements of Statistical Learning》。三位教授强大的统计+机器学习背景使得统计学家、计算机科学家和算法工程师们能够更好地从对方的角度理解Statistics和Machine learning的互相依托的关系。比如说，逻辑回归模型。在多元统计分析中，逻辑回归可以划做广义线性模型（GLM)的一种类型；在机器学习里面，它被当作是一类线性分类器。再比如主成分分析。同样是在回归分析中，主成分分析被看作是降维的一种方法；而在机器学习里面，它被看作是一种变量选择工具。机器学习和统计分析在大肆宣传的“大数据”时代来临之前，存在这么一种互相借鉴，互相学习的过程。交叉验证（cross-validation)这么一种带有浓厚算法工程味道的技术，在统计方法（主要是点估计）的研究中也被广泛使用并且效果不错。

统计学习依托背后的数学理论，在远早于机器学习大爆发的这十年，率先从解释因果的能力的角度，努力寻找上帝函数。在我看来，统计学习里最重要的两个部分就是回归分析和假设检验。其他的方法或者技术在统计学习这个大框架下，最终也是为了这两者服务的。回归分析提供了解释因果的武器，假设检验则给这项武器装上了弹药。单纯的线性回归用最小二乘法求解逼近事实的真相，再使用显著性检验，检测变量的显著性、模型的显著性、模型的拟合精度。当然是否属于线性，也可以使用假设检验的方法检测。非线性回归的问题，使用极大似然估计或者偏最小二乘回归求解模型，后续的显著性检验仍然是一样的思路。显著性检验有它的局限性，这本身是由统计学习的一些限定假设引起的，在没有更强大的解释因果的方法框架出现前，它依然是解释因果的第一选择。虽然显得粗糙，但是，能用。

我们从逻辑回归模型的角度来理解统计学习。从统计的方法论来看，逻辑回归脱胎于目标变量属于binary分布的非线性模型。所有的回归问题可以归结为确定稳定统计量（比如目标变量的期望或者中位值）和解释变量的函数关系。这种函数关系在目标变量服从指数分布族时，可以推导出它的结构，我们只需再求解结构中的未知参数即可。这种结构被称之为sigmoid函数，在信息学里面经常能用到。那么之所以从统计的角度逻辑回归可以得到严谨的数学解释和推断，全依赖于服从分布这个强假设。在这个假设下发展出的一整套理论，提供了现在这个通过数据学习世界的初级阶段，最优的解释因果框架。

我们再从逻辑回归出发来看看机器学习。机器学习，连带属于它的深度学习或强化学习，天生是为了解决大数据下的预测能力而提出并且发展的。目前最火热的图像识别、语音识别，包括金融领域里的借贷风险识别，为机器学习的发展提供了极其丰富的土壤- 极其大的数据量和极其多的（且极其稀疏的）数据特征。而众所周知的是，传统的回归分析在处理这种场景下的问题，收敛速度和预测精度都无法达到满意的程度。求解回归模型并非线性问题，在模型训练阶段只能在算法迭代过程中使用分布式系统提高运算速度，算法速度的提高受到限制。高维数据中经常碰到的稀疏问题，在回归模型中需要大量的预处理，也很难保证算法最后的收敛和估计精度。另一方面，逻辑回归可以看作是神经网络算法的一个特例- 删去隐含层，输入层和输出层直接用sigmoid函数连接。而加入了隐含层，通过参数调优，在原本逻辑回归的框架下，能够大大提高处理大样本，高稀疏的数据分类问题。虽然对于我们来说，隐含层输出的特征依然类似于上帝的选择- 机器根据训练过程中已有的输入和输出信息，得到了比人工选择更加优秀的数据特征，在最终的模型预测能力上大大胜过了原有的特征选择。

那么同样是逻辑回归模型，统计学习和机器学习的优势和劣势在哪里？在统计推断理论中，有这么一种妥协- bias - variance tradeoff，指的是我们在寻找一个统计估计量来接近真实值时，考虑到所拥有数据的健康程度，在估计精度和估计可信度之间需要做出平衡。打个比方，要得到偏离度很小的估计量必然会牺牲一部分可信度。而如果看重可信度，希望可信度的区间尽可能小，那么我们得到的估计量可能会和真实值有较大偏差。

这也是我们在逼近上帝函数的过程中遇到的问题。统计推断重解释，机器学习重预测。在小样本下，逻辑回归作为基础的线性分类器预测效果通常不比神经网络和其他ensembled算法差，且解释能力更强。当数据量越大，神经网络的预测能力就越强大，类似回归的统计推断方法越力不从心。在样本量不大的情况下，我们往往会比较重视模型的解释能力，因为数据量有限，特征之间是否有共线性不难发现，特征选择也只是在较少的维度下进行，模型的预测能力在我们的可控范围内不难做到最好。而在样本量大到超过我们的可控范围的情况下，预测能力是我们更看重的。因为这时，特征选择和特征间的相关性检测超过了我们的能力并且会极大伤害到我们的运算速度，牺牲特征的相关性检测，间接地等于放弃了模型的解释能力。从这个角度来讲，现在这个阶段，我们并没有一个很好的同时满足上帝函数两个必要条件的并且逼近上帝函数的统计或机器学习模型。

在统计推断发展的起步阶段，限制于数据采集能力和数据计算能力，往往是在解决小数据的问题。怎么在有限的资源里挖掘出更多的信息，怎么得到更精确的估计。精确性的要求远远大于对算法的速度和扩展性要求。机器学习一开始从工程的角度出发，首要解决的问题就是效率。这很契合大数据应用的初衷- 希望降低成本，优化配置，提高效率。那么随之而来的就是机器学习的模型必须在极大数据量的情况下尽可能得快，有极强的扩展性和可控性，在分布式系统出来之后，更要求易于部署。而精确性不再是最先需要解决的- 极其大的数据量导致机器学习的模型对结果的容忍度很高。银行的ToB贷款业务由于贷款金额巨大，且有严格的还贷能力要求，数学模型分析只能作为辅助，风险控制部分需要严格的公司财务分析和行业分析。而ToC的p2p贷款和消费金融这些本身在传统银行借贷体系里的长尾业务，由于自身交易量大，而单笔交易金额相对较低的特点，在保证一定预测精确度的前提下，可以充分利用机器学习模型进行风控，大幅度地提高贷前审查，批准贷款，还款这个业务闭环的效率。

相比较统计回归模型，机器学习最大的优势在于模型的可控性和延展性。怎么理解？boosting或者ensembled算法本身可以将逻辑回归当作它其中的一个弱分类器来使用。理论上，我们可以训练无数个分类器，直到全部收敛再归纳结果。同时，对于损失函数的选择，机器学习模型比统计回归模型由更大的自由度。比如分类决策树类型的模型，我可以选择GINIindex，也能用entropy来定义信息损失，由此可以建立不同的决策树算法。在具体算法参数迭代的过程中，机器学习专家为了提高运算速度，会对特征进行压缩，或者按照bagging用特征抽样的方法，降低特征维度。传统的回归模型在处理高维数据时，可以添加正则项压缩，而经典的通过显著性检验删减变量的方法其实无法大规模地使用。从模型的可控性来讲，对于回归模型来说，当模型的拟合程度不够，我们可以通过添加特征和特征转换的方法增加模型在训练集上的精确度，但代价是因模型的自由度（可以自由赋值的特征数量）增加过快导致的过拟合问题。对于机器学习模型比如说随机森林，当训练的结果不够好时，我们则是通过添加树的方法训练更多的树来降低训练的估计误差。这种方法每次的应用对模型的影响是相对弱的，但累积效应仍然能达到提高精确度的要求。而同时模型的自由度不会提升得很快，我们能有效控制模型的规模，防止过拟合的发生。另外由于现实数据或多或少存在很多噪音，传统的回归模型抗噪能力不强，需要大量的数据清洗工作才能保证模型的预测能力。而决策树模型和神经网络模型，从模型的设计角度来看更容易处理缺失值。