

**概率论与数理统计课程论文**

题目：统计与机器学习

|  |  |
| --- | --- |
| **班 号** | **2237102** |
| **学 号** | **2022110504** |
| **姓 名** | **蓝鸿翔** |
| **日 期** | **2023.11.24** |
| **成 绩** |  |

**摘 要**

统计学是一门研究数据收集、分析、解释和应用的学科，它通过使用概率和统计方法来推断和预测数据的特征和关系。在统计学中，我们研究如何收集数据、如何对数据进行建模和分析，以及如何从数据中得出结论。

机器学习是一种人工智能的分支领域，它关注如何通过计算机算法和模型，使计算机能够从数据中学习并进行预测和决策。机器学习通过训练算法来自动化地发现数据中的模式和规律，并根据这些模式和规律做出预测和决策。在人工智能的研究中，人们希望能够用机器 计算机 来模拟这种学习能力，学习找到数据内在的相互依赖关系从而对未知数据进行预测和判断。最终目标是使学习算法具有良好的泛化能力。

统计学在解决机器学习问题中起着基础的作用, 机器学习的日标在于使期望风险最小化。由于训练样本的有限性，而传统概率统计学所研究的主要是渐近理论即当样本趋于无穷多时的统计性质，为了解决这种问题，20世纪70年代Vapnik等人建立了一套新的理论体系即统计学习理论与传统统计学相 比，统计学习理论从更符合实际情况的有限样本假设出发，有完备的理论基础和严格的理论体系。

Minsky和Papert在20世纪60年代明确指出线性学习机计算能力有限[6]。总体上，现实世界复杂的应用需要比线性函数更富有表达能力的假设空间。多层感知器可以作为这个问题的一个解，由此导向了多层神经网络的反向传播算法。核函数表示方式提供了另一条解决途径，即将数据映射到高维空问来增强线性学习机的计算能力。核函数的引入最终使得在适当的特征空间中使用人们熟知的线性算法高效地检测非线性关系成为可能。

关键词：机器学习；统计学习理论；

## 课题背景及研究的目的和意义

机器学习致力于研究建立能够根据经验自我提高处理性能的计算机程序“ 。从最初的基于神经元模型以及函数逼近论的方法研究到以符号演算为基础的规则学习和决策树学习的产生以及之后的认知心理学中归纳、解释、类比等概念的引入直至最新的统计学习理论和支持向量机的兴起机器学习一直都在相关学科的研究与实际应用中起着重要的作用。在众多的机器学习方法中统计机器学习是近几年被广泛应用的机器学习方法。在众多的机器学习方法中，统计机器学习是近几年被广泛应用的机器学习方法。统计机器学习的发展历史大体上可以划分为四个阶段：Rosenblatt感知器模型、学习理论基础的创立阶段、人工神经网络、基于核的学习方法。

## 正文

机器学习的目的是根据已有数据或训练样本获取对系统输入输出之间关系的估计，使其可以对未知输入数据对输出数据做出尽可能准确的预测，机器学习问题可以被看作是利用有限数量的观测来寻找待求依赖关系的问题。一般可以形式化地表示为：已知变量x与输入y之间存在一定的未知的依赖关系，即遵循某一未知的联合概率F(x,y)，机器学习就是根据有限个独立同分布(x,y)观测样本，在一组函数f(x,w)中求一个最优的函数。对依赖关系进行估计使期望风险或称为真实风险最小。

学习问题的这种形式化表示的面是很广的包括 了很多特殊的问题 。其中三种主要的问题是 模式识别、回归估计和概率密度估计。

1， 模式识别，对于每一个测试样本，系统输出就是标号，可以通过y={0,1}或y={-1,1}构造指示函数（损失函数）：L(y,f(x,w)) ，风险泛函确定了系统的期望输出 和指示函数 f(x,w)的输出不同的概率 。因此，对于分类 问题而言，学习目标就是在概率测度 F(x,y)未知、但给定 了训练数据的情形下，寻找使得分类错误率最小的函数。

2， 回归估计。对于系统输出为实值，f(x,w)为实函数，可以利用回归方程构造损失函数：

L(y,f(x,w))=(y-f(x,w))^2

对于回归估计而言，学习目标就是在概率测度 F(x,y)未知、但给定了训练数据的情形下在允许的函数集合中寻找与实际回归函数 假如存在的话 最接近的函数表示。

3， 概率密度估计。概率密度估计。此时学习目标是根据训练样本确定x的概率分布。记估计的密度函数为p(x,w以定义为：

L(p(x,w))=-logp(x,w)

在上面的问题表述中，学习的日标在于使期望风险最小化。但是，我们并不知道概率分布 F(x,y)的具体形式,而只知道存在这样的一个概率分布,使得训练样本是遵循这个分布独立产生的。由于期望风险是依赖于概率分布 F(x,y)的，以要构造一个仅仅根据若千个训练样本就能找 出使期望风险最小的预测 函数是不可能的。因此在传统的学习方法中采用 了所谓的经验风险最小化归纳原则（ERM），即用样本定义经验风险：

用这种方式构造的风险等价于最大似然估计方法。

然而，通过经验最小化原则来替代期望最小化，并不完全合理。如在进行神经网络的学习过程中，开始训练都集中于将经验风险最小，但是很快发现经验最小化并不能达到最好的预测效果，训练误差过小反而会导致泛化能力的下降，即真实风险的增加，这就是过学习问题。虽然神经网络的学习能力足以记住每一个训练样本（只要网络足够复杂），经验风险也可以很快收敛到很小甚至为零，但却无法保证它对新的样本能够得到好的预测。统计学习理论被认为是 目前针对有限样本统计估计和预测的最佳理论，它从理 论上系统地研究了经验风险最小化原则成立的条件、有限样本下经验风险与期望风险之间的关系以及如何利用这些理论找到新的学习原则和方法等。

统计学习理论主要有四个方面：学习过程一致性的理论、学习过程收敛速度的非渐近理论、控制学习过程的泛化能力的理论、构造学习算法的理论。

学习过程一致理论。所谓学习过程一致性就是指当训练样木数 目趋于无穷大时，经验风险的最优值能够收敛到真实风险的最优值。因而原则一致性的条件是取决于预测函数集中 “最差” 的函数。学习理论关键定理把原则一致性的问题转化为了一致收敛的问题，虽然学习理论关键定理给出了 原则一致性成立的充分必要条件，但并没有给出什么样的预测函数能够满足这个条件为此统计学习理论定义了一些衡量函数集性能的指标这些指标多是从两分类函数即指示函数提出的，后来又推广到一般函数。

先介绍几个定义：指示函数集对某个样本集能实现的不同分类数日的对数称为函数集在这个样本集上的随机嫡，指示函数集在所有数日为 l的样本集上的随机嫡的期望值称为函数集在数目为l 的样本集上的 VC嫡 （或简称为嫡）记作 H(l)。

学习理论的第一个里程碑：指示函数集学习过程双边一致收敛 满足一致性的充分必要条件是随着观测的样本数目的增加嫡与样本数目的比值应该趋近于零。这一定理被认为是学习理论的第一个里程碑 ，它要求所有最小化经验风险的机器都要满足它。

除此之外，机器学习理论还有其他两个里程碑，这里就不再一一介绍了，他们建立学习机器收敛速度的界以及经验风险与真实风险之间误差的上界的基础，包括严格依赖于分布的界和与分布无关的界。一般认为，统计学习理论中最具指导性的结论是泛化能力的界。需要指出的是和学习理论关键定理一样，泛化能力的界也是对于最坏情况的结论。在很多情况下是比较松弛的。而且，这种界只在同一类学习函数进行比较时有效。可以指导我们从函数集中选择最好的函数，在不同函数集之间比较却不一定成立。事实上，寻找反映学习机器性能的更好的指标从而得到更紧的界是学习理论今后的重要研究方向之一。

统计学在机器学习问题中有着基础性的作用。从有限的训练样本中，通过构造出合适的损失函数，为了找到这个函数，我们需要给算法一些方法来“学习”如何最好地着手处理这个问题，而这由一个被称为“损失函数”的概念来提供。因此，对我们所有的每个假设（也即提议的函数），我们要通过比较所有数据下其预期风险的值来衡量这个函数的表现。利用原有的传统统计学原理，难以达到足够多的训练样本，通过利用统计学习理论，可以构造出更合理的损失函数，从而达到较好的预测效果。

**参考文献**

[1] V. Vapnik著.张学工 译. 统计学习理论的本质. 北京:清华大学出版社, 2000

[2] T.Mitchell著. 曾华军，张银奎等译. 机器学习. 北京:机械工业出版社, 2003

[3] 史忠植. 知识发现. 北京: 清华大学出版社, 2000

[4]M.Minsky and S.Papert. Perceptrons. Cambridge,MA:MIT Press,1969

[5] 李建民，张钹，林福宗.序贯最小优化的改进算法.软件学报， 2003

[6] 张学工.关于统计学习理论与支持向量机.自动化学报，2000