模式识别与机器学习期末考查
试　卷

研究生姓名：　　入学年份：　　导师姓名：

试题 1：简述模式识别与机器学习研究的共同问题和各自的研究侧重点。

答：（1）模式识别是研究用计算机来实现人类的模式识别能力的一门学科，是指对表征事物或现象的各种形式的信息进行处理和分析，以对事物或现象进行描述、辨认、分类和解释的过程。主要集中在两方面，一是研究生物体（包括人）是如何感知客观事物的，二是在给定的任务下，如何用计算机实现识别的理论和方法。机器学习则是一门研究怎样用计算机来模拟或实现人类学习活动的学科，是研究如何使机器通过识别和利用现有知识来获取新知识和新技能。主要体现以下三方面：一是人类学习过程的认知模型；二是通用学习算法；三是构造面向任务的专用学习系统的方法。两者关心的很多共同问题，如：分类、聚类、特征选择、信息融合等，这两个领域的界限越来越模糊。机器学习和模式识别的理论和方法可用来解决很多机器感知和信息处理的问题，其中包括图像/视频分析（文本、语音、印刷、手写）文档分析、信息检索和网络搜索等。

（2）机器学习和模式识别是分别从计算机科学和工程的角度发展起来的，各自的研究侧重点也不同。模式识别的目标就是分类，为了提高分类器的性能，可能会用到机器学习算法。而机器学习的目标是通过学习提高系统性能，分类只是其最简单的要求，其研究更侧重于理论，包括泛化效果、收敛性等。模式识别技术相对比较成熟了，而机器学习中一些方法还没有理论基础，只是实验效果比较好。许多算法他们都在研究，但是研究的目标却不同。如 SVM 在模式识别中研究所关心的就是其对人类效果的提高，偏工程。而在机器学习中则更侧重于其性能上的理论证明。

试题 2：列出在模式识别与机器学习中的常用算法及其优缺点。

答：（1） K 近邻法

KNN 算法作为一种非参数的分类算法，它已经广泛应用于分类、回归和模式识别等。在应用 KNN 算法解决问题的时候，要注意的两个方面是样本权重和特征权重。

优缺点：非常有效，实现简单，分类效果好。 样本小时误差难控制，存储所有样本，需要较大存储空间，对于大样本的计算量大。

（2） 贝叶斯决策法

贝叶斯决策法是以期望值为标准的分析法，是决策者在处理风险型问题时常常使用的方法。

优缺点：由于在生活当中许多自然现象和生产问题都是难以完全准确预测的，因此决策者在采取相应的决策时总会带有一定的风险。贝叶斯决策法就是将各因素发生某种变动引起结果变动的概率凭统计资料或凭经验主观地假设，然后进一步对期望值进行分析，由于此概率并不能证实其客观性，故往往是主观的和人为的概率，本身带有一定的风险性和不肯定性。虽然用期望的大小进行判断有一些风险，但仍可以认为贝叶斯决策是一种兼科学性和实效性于一身的比较完善的用于解决风险型决策问题的方法，在实际中能够广泛应用于组织系统改革、企业效益、市场开发、证券投资等诸多领域。使用时根据决策者的侧重点，结合变异系数，综合使用货币因素的贝叶斯决策、或效用函数的贝叶斯决策法，都会得到自己想要的结果。

（3） DES 加密算法

DES 是 Data Encryption Standard（数据加密标准）的缩写，它为密码体制中的对称密码体制，又被称为美国数据加密标准，是 1972 年美国 IBM 公司研制的加密算法。DES 是一个分组加密算法，他以 64 位为分组对数据加密。同时 DES 也是一个对称算法：加密和解密用的是同一个算法。它的密匙长度是 56 位（因为每个第 8 位都用作奇偶校验），密匙可以是任意的 56 位的数，而且可以任意时候改变。其中有极少量的数被认为是弱密匙，但是很容易避开他们。所以保密性依赖于密钥。

优缺点：具有极高安全性，分组比较短，密钥太短，密码生命周期短，运算速度较慢。

（4） 决策树学习算法

决策树算法是一种混合算法，它综合了多种不同的创建树的方法，并支持多个分析任务，包括回归、分类以及关联。决策树算法支持对离散属性和连续属性进行建模。

优缺点：决策树算法高效快速且可伸缩，可轻松实现并行化，这意味着所有处理器均可协同工作，共同生成一个一致的模型。这些特征使决策树分类器成为了理想的数据挖掘工具。在数据挖掘的各种方法中 ,决策树归纳学习算法以其易于提取显式规则、计算量相对较小、可以显示重要的决策属性和较高的分类准确率等优点而得到广泛应用。决策树的这种易理解性对数据挖掘的使用者来说是一个显著的优点。然而决策树的这种明确性可能带来误导。比如，决策树每个节点对应分割的定义都是非常明确毫不含糊的，但在实际生活中这种明确可能带来麻烦。对决策树常见的批评是说其在为一个节点选择怎样进行分割时使用“贪心”算法。此种算法在决定当前这个分割时根本不考虑此次选择会对将来的分割造成什么样的影响。

（5）C 均值算法

C 均值算法是通过不断调整聚类中心使得误差平方和准则函数取得极小值。优缺点：能够动态聚类，是一种无监督学习算法，算法简单，速度快，局部搜索能力强，能够有效处理大型数据库，与神经网络结合可极大地提高收敛性和精度。c-均值算法的一个主要问题是划分类别数必须事先确定，这种主观确定数据子集数目并不一定符合数据集自身的特点，所以对于随机的初始值选取可能会导致不同的聚类结果，甚至存在着无解的情况；在选取聚类中心点时采用随机选取易使得迭代过程陷入局部最优解，容易收敛于局部极小点；该算法对“噪音”和孤立点数据比较敏感，少量的该类数据能够对平均值产生极大的影响。

（6）遗传算法

遗传算法（Genetic Algorithm）是模拟达尔文的遗传选择和自然淘汰的生物进化过程的计算模型，是一种通过模拟自然进化过程搜索最优解的方法。

优缺点：遗传算法是一类可用于复杂系统优化的具有鲁棒性的搜索算法，与传统的优化算法相比，主要有以下特点： 1. 与问题领域无关切快速随机的搜索能力。2. 搜索从群体出发，具有潜在的并行性，可以进行多个个体的同时比较。3. 搜索使用评价函数启发，过程简单。4. 使用概率机制进行迭代，具有随机性。5. 具有可扩展性，容易与其他算法结合。6. 直接以适应度作为搜索信息，无需导数等其它辅助信息。7. 使用多个点的搜索信息，具有隐含并行性。8. 使用概率搜索技术，而非确定性规则。也存在一些问题：1. 没有能够及时利用网络的反馈信息，故算

法的搜索速度比较慢，要得要较精确的解需要较多的训练时间。2. 算法对初始种群的选择有一定的依赖性，能够结合一些启发算法进行改进。3. 算法的并行机制的潜在能力没有得到充分的利用，这也是当前遗传算法的一个研究热点方向。

(7)BP 神经网络算法

其学习过程由正向传播和反向传播组成。在正向传播过程中，输入信息从输入层经隐单元层逐层处理后，传至输出层。如果输出层得不到期望输出，那么就转为反向传播，把误差信号沿连接路径返回，并通过修改各层神经元的权值，使误差信号最小。

优缺点：BP 算法能够通过学习带正确答案的实例集自动提取“合理的”求解规则；具有一定的推广能力；学习过程有被“固化”的潜在可能性；它能以任意精度逼近任意非线性函数，而且具有良好的逼近性能，并且结构简单，是一种性能优良的神经网络。但也存在一些问题，BP 算法是按照均方误差的梯度下降方向收敛的，但均方误差的梯度曲线存在不少局部和全局最小点，这就使得神经网络易陷入局部最小；算法的收敛速度较慢，可能会浪费大量时间；神经网络隐层的结点个数难以确定合适的数值；如何选取合适的学习样本解决网络的推广（泛化）问题，即使网络能正确处理未学习过的输入。

（8）Hopfield 网络算法

Hopfield 网络算法作为典型的反馈神经网络，有下列特有的优点和缺点。

1. 只有不动点吸引子，没有其它类型的吸引子。Hopfield 同的这个性质被称为全局稳定性。2. 网络状态的演化趋于某个二次函数的局部最小点。3. 很难精确地分析 Hopfield 网的性能。4. 难于找到通用的学习算法。5. 这类阿络的动力学行为过于简单。5. Hopfield 问只有不动点吸子，是一种消极被动的神经网络。

试题 3：简述在模式识别与机器学习中解决问题的主要步骤。指出那些步骤涉及到学习？在数据的前处理中，特征选择起什么作用？

答： (1) 在模式识别与机器学习中解决问题的主要步骤：

1. 问题描述：准确分析研究目的，并对未来工作做出计划。

2. 数据选择：数据选择是根据用户需求从数据库中提取相关数据。

3. 知识发现过程： 归纳为 3 个步骤，即数据挖掘预处理、数据挖掘、数据挖掘后处理。数据预处理是对数据进行再加工，检查数据的完整性及一致性，对其中的噪音数据进行处理。对丢失的数据利用统计方法进行填补，形成发掘数据库。数据变换即从发掘数据库里选择数据，变换的方法主要是利用聚类分析和判别分析。数据挖掘是根据用户要求，确定知识发现的目标是发现何种类型的知识。运用选定的知识发现算法。从数据库中提取用户所需要的知识。知识评价主要用于对所获得的规则进行价值评定，以决定所得到的规则是否存入基础知识库。

4. 选择或设计模型：对同一个问题或许有许多不同的模型可以描述，不同的模型会导致识别和学习结果的不同，因此需要利用已有的经验和知识来选择或设计适当的模型。在确定了所建立的模型后，就可以估计模型的参数，需要注意的时，应该使得模型对未知数据有良好的适应性。

5. 训练所建立的模型：用前面所得的数据分成两组，一组作为训练数据，一组作为测试数据。设定目标误差，用训练数据对所建立的模型进行训练，达到目标误差，就停止训练，这样就确定了所建立模型的参数。

6. 测试、评估、验证模型：测试模型的目的是为了确定所建立模型是否满足实际应用要求。测试数据应该和训练用的样本数据不一致，否则，测试所得的结果永远都是满意的。用测试数据对所建立模型进行测试，观察测试结果是否与实际情况是相符合。若与实际情况相符合，所建立模型就可对未知数据做预测，从而得到进一步的验证。

(2)在这些步骤中，步骤 5 涉及到学习。

(3) 特征选取（也称作属性选择）是简化数据表达形式，是在模式识别中根据一定的原则，选取反映被识别模式本质的那些特征的方法或过程。模式识别和机器学习方法首先要解决的一个问题就是特征选择。在数据的前处理中，特征选择是一个非常重要的步骤，特征选择不合理，会影响识别和学习效果。通过特征选择和提取，我们才可得到所采集数据中最有效的信息，最有效的特征，选择出有利于分类或聚类建立模型的变量，从而实现特征空间维数的压缩，以降低后续处理过程的难度，才能基于这些特征对所建立模型进行训练和测试。同时特征选取也是降低存储要求，提高分类精度和效率的重要途径。

试题 4：在模式识别与机器学习的研究中，还不断有人提出新的算法。请列举一些可以用来比较算法好坏的方法？

答：算法是计算机科学中一个重要的研究方向，是解决复杂问题的关键。在计算机世界中，算法无处不在。同一问题可用不同算法解决，而一个算法的质量优劣将影响到算法乃至程序的效率。可以用来比较算法好坏的方法有：

1. 正确性

一个算法是否正确的，是指对于一切合法的输入数据，该算法经过有限时间（算法意义上的有限）的执行是否都能产生正确（或者说满足规格说明要求）的结果。

2. 时间复杂度和空间复杂度

一个算法的时间复杂性是指该算法的基本运算次数，记作 T(n)=O(f(n))。时间复杂度不断增大，算法的执行效率越低。空间复杂度是指算法在计算机内执行时所需存储空间的度量。记作 S(n)=O(f(n)) 。存储空间越大，算法效率也越低。

3. 占用空间

算法执行需要存储空间来存放算法本身包含的语句、常数、变量、输入数据和实现其运算所需的数据（如中间结果等），此外还需要一些工作空间用来对（以某种方式存储的）数据进行操作。

4. 可读性

可读性好的算法有助于设计者和他人阅读、理解、修改和重用。与此相反，晦涩难懂的算法不但容易隐藏较多的错误，而且增加了人们在阅读、理解、调试、修改和重用算法等方面的困难。

5. 坚固性

当输入数据非法时，算法能适当地作出合适的反应。

试题 5：在你所知道的模式识别与机器学习算法中，那些方法较合适用来解决纯数值型数据的问题，那些方法较适合用来解决包含大量非数值数据的问题。

答：（1）解决纯数值型数据问题的方法：贝叶斯决策法、神经网络算法等。贝叶斯决策法是基于概率统计的基本的判别函数分类法。只要知道先验概率和条件概率就可以对样本进行判断，由于数据是纯数值型数据，数据简单，样本间的空间距离

易计算，且先验概率和条件概率易求得。神经网络只能处理数值型数据。建立神经网络需要做的数据准备工作量很大。 要想得到准确度高的模型必须认真的进行数据清洗、整理、转换、选择等工作。对任何数据挖掘技术都是这样，神经网络尤其注重这一点。比如神经网络要求所有的输入变量都必须是 0—1(或-1—+1)之间的实数，因此像“地区”之类文本数据必须先做必要的处理变成数值之后才能用作神经网络的输入。

（2）对于非数值型数据可用方法：决策树、遗传算法等。决策树很擅长处理非数值型数据， 决策树的分类方法是从实例集中构造决策树，是一种有指导的学习方法。其算法的特点是通过将大量数据有目的分类，从中找到一些有价值的，潜在的信息，特别适合大规模的数据处理。遗传算法特点从解集合进行搜索，利于全局择优。该算法具有收敛性，通过选择、交叉、变异操作，能迅速排除与最优解相差极大的串。是非数值并行算法之一，解决了非数值数据及大量数据带来的计算量和存储量的问题。

试题 6：模式识别与机器学习最难解决的问题是什么？并说明理由。

答：我觉得模式识别与机器学习中最难解决的问题是：

（1）学习速率的确定。提出设计者应该从具体系统中获得的数据确定算法学习速率的上、下界数值，并选取最优学习速率。

（2）在处理具体的问题时，合适算法的选择。 在算法选择中没有天生优越的模式。识别与机器学习算法，各自算法的都有其对应的应用范围及应用中应注意的问题，只有充分了解不同模式识别算法，深入分析算法的使用条件，才能做到最佳选择。但目前算法很多，没有深入的话容易被遗忘，深入的话花得时间多，且在很多实际问题当中，常常不容易找到那些最重要的特征，或者受条件限制不能对它们进行测量，这使得特征选择和提取的任务复杂化，从而成为构造模式识别系统，提高决策精度的最困难的任务之一。

（3）相应的参数的选择。如何确定变量值，这是一个很关键的问题，但至今还没有快速而有效的规则，有的只是一些原则性的指导。而且选择参数值最终还应归结为每个用户对算法的体验，用户只能通过自己的编程实践，用各种不同的参数值进行调试，看结果会发生什么，并从中选取适合的值。

试题 7：请例举一些你认为应用得较好的算法及应用实例。

答： 我认为应用较好的算法如下：

（1）遗传算法

由于遗传算法的整体搜索策略和优化搜索方法在计算是不依赖于梯度信息或其它辅助知识，而只需要影响搜索方向的目标函数和相应的适应度函数，所以遗传算法提供了一种求解复杂系统问题的通用框架，它不依赖于问题的具体领域，对问题的种类有很强的鲁棒性，所以广泛应用于许多科学。

1、 函数优化

函数优化是遗传算法的经典应用领域，也是遗传算法进行性能评价的常用算例，许多人构造出了各种各样复杂形式的测试函数：连续函数和离散函数、凸函数和凹函数、低维函数和高维函数、单峰函数和多峰函数等。对于一些非线性、多模型、多目标的函数优化问题，用其它优化方法较难求解，而遗传算法可以方便的得到较好的结果。

2、 组合优化

随着问题规模的增大，组合优化问题的搜索空间也急剧增大，有时在目前的计算上用枚举法很难求出最优解。对这类复杂的问题，人们已经意识到应把主要精力放在寻求满意解上，而遗传算法是寻求这种满意解的最佳工具之一。实践证明，遗传算法对于组合优化中的 NP 问题非常有效。例如遗传算法已经在求解旅行商问题、背包问题、装箱问题、图形划分问题等方面得到成功的应用。

此外，GA 也在生产调度问题、自动控制、机器人学、图象处理、人工生命、遗传编码和机器学习等方面获得了广泛的运用。

（2）BP 神经网络算法

BP 神经网络模型有输入层、隐含层、输出层三个层次，通过误差反向后传算法来消除误差。它是一种具有模式变换能力、自组织、自适应、自学习特点的计算机制，它具有高度的并行结构和并行实现能力，具有高速寻找优化解的能力。应用也比较广。

1、 BP 人工神经网络模型在企业综合绩效管理评价体系中的应用

从输入层输入企业综合绩效评价的指标数据，经隐含层处理后传入输出层，输出结果即为评价结果。在正向传播阶段，每一层神经元的状态只影响到下一层神经元的状态。如果输出层所得到的输出结果与期望输出结果的误差超过误差允许范围，则进入误差反向后传阶段，误差信号按原来的连接通路返回，将误差进行反向传播，求出隐含层单元的一般化误差，调整各层之间的连接权值以及隐含层、输出层的阀值，使输出期望值和神经网络实际输出值的均方误差趋于最小。以足够的样本运用优化 BP 模型学习算法来训练此网络，训练好的网络所持有的那组权系数就是所要确定的企业综合绩效评价指标的权重。最后，将目标企业综合绩效评价指标的具体值作为训练好的 BP 模型的输入，可得目标企业的绩效评价。

2、应用到高校的学生就业工作中

通过收集已毕业的学生信息，对数据信息进行合并，形成结构统一的就业信息数据源。对数据源进行数据预处理，去掉与决策无关的属性和高分支属性、处理含空缺值的属性，然后根据随机算法，在训练样本数据库中，抽取其中 2/3 的数据用于训练网络，剩余 1/3 的数据用于测试模型的准确率。采用三层 BP 网络来进行建模。BP 神经网络具有很强的自适应性和学习能力，将其应用于毕业生就业预测中精度较高。经过对实际毕业生就业信息的预测，其结果与实际情况吻合理想，因此对毕业生就业指导有着现实的意义。BP 网络模型完全可以用来预测计算。