《机器学习技术导论》笔记——齐诏娣

写于 2018 年

Chapter1 机器学习概述

一、知识梳理

- **1. 机器学习目的**

  让机器通过经验积累来学习知识和掌握技能。

- **2. 机器学习基本框架**

  “**知识**“和“**经验**”是构造机器学习系统时常用的两个基本信息源。

  知识-经验的机器学习是一种将人类知识和实际经验相结合，以提高计算机处理某种特定任务能力的计算框架。这一框架包括学习目标、学习结构、训练数据和训练方法组成。基于这一框架，我们依赖先验知识设计合理的学习结构，设计相应的学习算法，利用实际经验对学习结构进行调整，实现既定学习目标最优化。

  学习目标：不同学习任务目标不同。学习任务分类，应用角度：感知任务、归纳任务、生成任务等。技术角度：预测任务（回归：MSE、分类：CE）、描述任务（聚类、概率估计）。

  学习结构：学习结构定义学习任务如何进行，一般称为“模型“。一些可能的学习结构包括：函数、网络（神经网络、概率图等）、规则集、有限状态自动机、语法结构。定义学习结构，本身即是对先验知识进行形式化的过程。

  训练数据：数据积累是机器学习研究的基础。形式多种多样，取值类型：二值、多值、连续数据等。复杂度：单值、向量、图、自然物等。收集和整理数据注意事项：数据是否完整、是否有动态性、不同数据间的相关性如何。另外，通过数据选择、特征提取、预处理等，抽取最有价值的数据进行学习。

  学习方法：学习方法是学习过程的具体表现，即算法。其分类，是否需认为标注：监督学习、无监督学习、半监督学习、增强学习。优化方法分类：直接求解（如 PCA 求解数据协方差矩阵的特征向量）、数值优化（如神经网络中梯度下降算法）、遗传算法（协同学习中鸟群算法）等。

- **3. 机器学习的流派**

  当前机器学习的研究主要受四门基础学科的启发：人工智能、概率与统计理论、生理学与神经学、仿生学与进化论。这些学科的研究对象和研究方法启发机器学习从不同方向思考学习问题，形成四个不同流派，即符号学派、贝叶斯学派、连接学派、进化仿生学派。

  符号学派：符号学派研究者认为所有智能行为都可以简化成在一个逻辑系统中的符号操作过程。局限性：对不确定性的描述能力不足。

  贝叶斯学派：贝叶斯学派研究学者从一开始就关注对不确定性的描述。局限性：在推

理过程中计算会比较复杂、为了推理简单，先验概率和条件概率一般会采用比较基础的函数形式（高斯分布、多项式分布等）、应用局限（两变量之间存在的关系只有领域专家才能确定）。贝叶斯方法在很大程度上依然是以知识为驱动的方法。

连接学派：也称神经网络学派，其基本思想是基于大量同质结点的连接网络来模拟智能行为。人工神经网络的结构多种多样，一般常用的结构是层次结构，有时会加入空间结构限制或时序递归连接。神经网络的连接权重一般采用随机初始化，并基于训练数据进行优化。预测任务：训练准则是使网络预测值与实际观值之间的误差最小，训练方法一般采用 BP 算法。记忆任务：训练准则是使网络生成训练数据的概率最大。

进化仿生学派：进化仿生学派没有自己特别的学习结构，而是一种学习方法，这一方法可以应用到各种学习结构上。优点：可以优化很多传统学习方法无法优化的模型（离散参数的神经网络、后验概率计算极为复杂的贝叶斯网络）。缺点：这种 Trial-and-Error 方法效率很低。为提高学习效率，提出各种遗传算法。遗传算法很难得到全局最优结果，但当局部最优点较多时，这一方法却容易摆脱局部最优。

- **4. 机器学习技术**

有趣的例子：从猴子摘香蕉到星球大战、集体学习的机器人、图片和文字理解、金融市场量化分析、Alpha Go

前沿：Michael Jordan 和 Tom Mitchell 2015 年在 Science 上发表了一篇文章，“Machine learning: Trends，perspectives, and prospects”，讨论了机器学习的进展和未来。同年，LeCun、Bengio 和 Hinton 在 Nature 发表“Deep learnign”一文，对深度学习的进展进行了总结，并对未来进行了展望。这两篇文章可以作为机器学习领域最近一段时间的研究方向指南。

- **5. 机器学习基础**

机器学习很大程度上注重“权衡”。

(1) 训练、验证与测试

同一模型在训练数据和测试数据具有不同的表达能力。

(2) 模型的表达能力与泛化能力

不同模型在训练数据和测试数据上的性能偏差。

(3) 没有免费的午餐

(4) 经典机器学习方法

a、 监督学习与非监督学习

监督学习：线性模型（Linear Regression, Logistic Regression）、非线性模型（SVM, NN）、参数模型（NaiveBayes, LDA, HMM, Probabilistic Graphical Models）、非参数模型（K-nearest Neighbors, Kernel Regression, Kernel Density Estimation, Local Regression, CART）、集体模型（Bagging/Bootstrap+Aggregation, Adaboost, Random Forest）

非监督模型：聚类（K-means Clustering, Spectral Clustering）、概率密度估计（GMM，Graphical Models）、降维和流行学习（PCA, Factor Analysis, MDS）

b、 线性模型与非线性模型

c、 参数模型与非参数模型

d、 生成模型与区分性模型

生成模型对每类数据分布进行建模 $p(x|C_k)$，再依贝叶斯公式得到分类模型；区分模型不考虑数据分布，仅关注分类面，直接对分类面建模。

e、 概率模型与神经模型

- **6. 机器学习过程**

  开始一个机器学习任务：

  (1) 设定目标函数

  (2) 设定模型结构

  (3) 设定约束方法：为防止过拟合

  (4) 设定训练算法

  (5) 设定推理算法

二、知识碎片

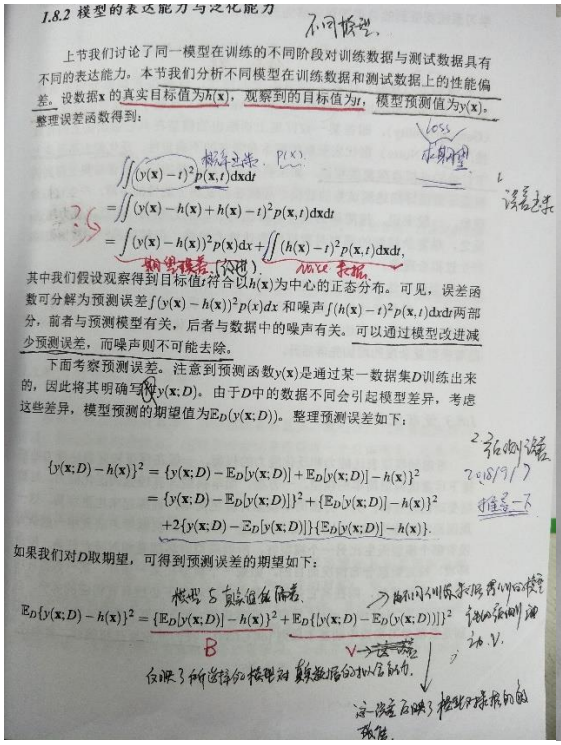
- **1. 同质**

  同质：节点的性质是一样的。

- **2. 符号学派**

  根据成功的经验，反向推出定理。

- **3. 模型的表达能力与泛化能力（不同模型）**



1.8.2 模型的表达能力与泛化能力

上节我们讨论了同一模型在训练的不同阶段对训练数据与测试数据具有不同的表达能力，本节我们分析不同模型在训练数据和测试数据上的性能偏差。设数据$\mathbf{x}$的真实目标值为$h(\mathbf{x})$，观察到的目标值为t，模型预测值为$y(\mathbf{x})$。整理误差函数得到：

$$\iint (y(\mathbf{x})-t)^2 p(\mathbf{x},t)\mathrm{d}\mathbf{x}\mathrm{d}t$$
$$=\iint (y(\mathbf{x})-h(\mathbf{x})+h(\mathbf{x})-t)^2 p(\mathbf{x},t)\mathrm{d}\mathbf{x}\mathrm{d}t$$
$$=\int (y(\mathbf{x})-h(\mathbf{x}))^2 p(\mathbf{x})\mathrm{d}\mathbf{x}+\iint (h(\mathbf{x})-t)^2 p(\mathbf{x},t)\mathrm{d}\mathbf{x}\mathrm{d}t,$$

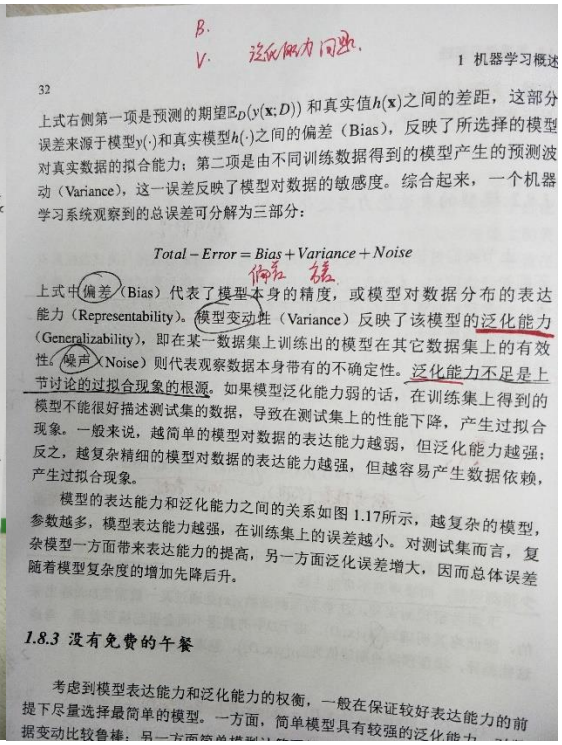
其中我们假设观察得到目标值t符合以$h(\mathbf{x})$为中心的正态分布。可见，误差函数可分解为预测误差$\int (y(\mathbf{x})-h(\mathbf{x}))^2 p(x)\mathrm{d}x$和噪声$\int (h(\mathbf{x})-t)^2 p(\mathbf{x},t)\mathrm{d}\mathbf{x}\mathrm{d}t$两部分，前者与预测模型有关，后者与数据中的噪声有关。可以通过模型改进减少预测误差，而噪声则不可能去除。

下面考察预测误差。注意到预测函数$y(\mathbf{x})$是通过某一数据集D训练出来的，因此将其明确写成$y(\mathbf{x};D)$。由于D中的数据不同会引起模型差异，考虑这些差异，模型预测的期望值为$\mathbb{E}_D(y(\mathbf{x};D))$。整理预测误差如下：

$$\{y(\mathbf{x};D)-h(\mathbf{x})\}^2=\{y(\mathbf{x};D)-\mathbb{E}_D[y(\mathbf{x};D)]+\mathbb{E}_D[y(\mathbf{x};D)]-h(\mathbf{x})\}^2$$
$$=\{y(\mathbf{x};D)-\mathbb{E}_D[y(\mathbf{x};D)]\}^2+\{\mathbb{E}_D[y(\mathbf{x};D)]-h(\mathbf{x})\}^2$$
$$+2\{y(\mathbf{x};D)-\mathbb{E}_D[y(\mathbf{x};D)]\}\{\mathbb{E}_D[y(\mathbf{x};D)]-h(\mathbf{x})\}.$$

如果我们对D取期望，可得到预测误差的期望如下：

$$\mathbb{E}_D\{y(\mathbf{x};D)-h(\mathbf{x})\}^2=\{\mathbb{E}_D[y(\mathbf{x};D)]-h(\mathbf{x})\}^2+\mathbb{E}_D\{[y(\mathbf{x};D)-\mathbb{E}_D(y(\mathbf{x};D))]\}^2$$



1 机器学习概述

32

上式右侧第一项是预测的期望$\mathbb{E}_D(y(\mathbf{x};D))$和真实值$h(\mathbf{x})$之间的差距，这部分误差来源于模型$y(\cdot)$和真实模型$h(\cdot)$之间的偏差（Bias），反映了所选择的模型对真实数据的拟合能力；第二项是由不同训练数据得到的模型产生的预测波动（Variance），这一误差反映了模型对数据的敏感度。综合起来，一个机器学习系统观察到的总误差可分解为三部分：

$$Total-Error=Bias+Variance+Noise$$

上式中偏差（Bias）代表了模型本身的精度，或模型对数据分布的表达能力（Representability）。模型变动性（Variance）反映了该模型的泛化能力（Generalizability），即在某一数据集上训练出的模型在其它数据集上的有效性。噪声（Noise）则代表观察数据本身带有的不确定性。泛化能力不足是上节讨论的过拟合现象的根源。如果模型泛化能力弱的话，在训练集上得到的模型不能很好描述测试集的数据，导致在测试集上的性能下降，产生过拟合现象。一般来说，越简单的模型对数据的表达能力越弱，但泛化能力越强；反之，越复杂精细的模型对数据的表达能力越强，但越容易产生数据依赖，产生过拟合现象。

模型的表达能力和泛化能力之间的关系如图 1.17所示，越复杂的模型，参数越多，模型表达能力越强，在训练集上的误差越小。对测试集而言，复杂模型一方面带来表达能力的提高，另一方面泛化误差增大，因而总体误差随着模型复杂度的增加先降后升。

1.8.3 没有免费的午餐

考虑到模型表达能力和泛化能力的权衡，一般在保证较好表达能力的前提下尽量选择最简单的模型。一方面，简单模型具有较强的泛化能力

- **4. 过拟合问题**

  分为：（1）参数过拟合（迭代次数过多）；（2）结构过拟合

Chapter2 线性模型

一、知识梳理

本章讨论几种简单的线性模型，一种是线性预测模型（有监督），包括线性回归和 Logistic 回归，在讨论它们概率意义的基础上引入贝叶斯方法；另一种是线性概率模型（无监督），基于隐变量和观察变量间的线性假设来推理数据的内在结构，如 PCA、LDA、PLDA 等。

- **1．线性预测模型**

(1) 多项式拟合

$$y(x;w) = w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M = \sum_{j=0}^{M} w_j x^j,$$

为什么要使用平方误差作为误差函数？

平方误差对应的是训练数据中的一种高斯不确定性。

(2) 线性回归

引入概率模型帮助描述不确定性，并由此得到基于概率的最优解。

$$\begin{aligned} t &= y(\mathbf{x};\mathbf{w}) + \varepsilon \\ &= \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}) + \varepsilon, \end{aligned}$$

其中$\varepsilon \sim N(0, \beta^{-1})$。

上式构造了一个由 x 到观察值 t 的生成模型：首先，由输入变量 x 经过非线性映射生成特征向量，再经过线性映射生成预测 y，最后加入一个高斯噪声得到目标的观察值 t。这一模型称为**线性回归模型**。

所谓**回归**，是指对输入变量 x 和目标变量 t 之间相互依赖关系的统计分析。

一个问题是，高斯分布是最优选择吗？

这需要结合数据的实际分布请况。比如当数据具有很强的长尾特性时，就可能考虑像 Student-t 分布或拉普拉斯分布。一般来说，如果我们对数据的特性了解的并不清楚，高斯分布通常是合理选择。然而，有一种情况需要我们必须考虑非高斯分布：当目标 t 是离散的，则高斯分布显然是不适合的，这时需要用离散分布来描述数据中的噪声。（Logistic 回归模型即是处理离散噪声的模型）

(3) Fisher 准则与线性分类

分类问题有三种可能的求解方法：

区分函数法；

生成性概率模型法；

区分性概率模型法。

线性拟合等价于 p(t|Φ)为高斯分布的线性回归，这说明 Fisher 方法事实上假设了分类任务中的类别标记是高斯的，这显然是不太合理的。这一高斯分布假设带来分类面上的偏差。

(4)Logistic 回归

首先对输入 x 经过一个非线性映射Φ(.)生成特征，再经由一个线性映射 W'Φ投影到一个标量空间，再经过 б(.)压缩到（0，1）之间，最后把该压缩值作为伯努利分布的参数生成目标 t，这一模型称为 Logistic 回归模型。

比较 Logistic 模型和线性回归模型，可见二者具有相似性，差别只是一个非线性映射函数 б(.)和伯努利分布假设。

(5)**Softmax 回归**

Softmax 回归将目标 t 由二分类问题中的伯努利分布扩展到多项式分布，相应的表示方法也由 0-1 表示扩展为 One-hot 表示。基于交叉熵准则，利用梯度下降法可实现对模型参数的优化。

(6)**小结**

最大似然准则，模型对数据的生成概率最大化；

最大后验准则，需要将知识和数据结合在一起时，这时就不仅要考虑训练数据概率最大化，还要考虑先验知识在目标函数中的比重。

- **2．线性概率模型**

线性概率模型和线性预测模型具有类似形式，区别在于线性概率模型中 x 是不可见的随机变量，而在线性预测模型中 x 是可见的确定输入。这一区别很重要：当 x 变成隐变量以后，我们能观察到的数据只有 t，因此学习方法由监督学习变成了无监督学习，推理过程也由前向预测变成了反向推理。

线性概率模型被广泛应用于因子分析和特征提取等任务中，是概率图的简单形式。

(1) 主成分分析(PCA)

考虑如下两种准则来评价主成分对数据的代表性：一是数据在主成分代表的映射空间里方差最大，二是由映射恢复成原始数据的损失最小。（这两种准则事实上是等价的）为取使 L(v1)最大的主成分，只需取协方差矩阵的最大特征值所对应的特征向量即可。取得第一个主成分后，依类似步骤可得到后续各主成分。若要求前 k 个主成分，只需选择特征值最大的 k 个特征向量即可。

(2) 概率主成分分析(PPCA)

提出原因：【PCA 是经典的无监督学习方法，广泛应用于降维、正则化、流行学习等任务中。然而，PCA 的优化函数（映射空间方差最大或恢复误差最小）和主成分之间的正交限制很大程度上是种人为定义，这使得 PCA 的适用性缺少明确解释。】

线性概率模型：$t=\mu + Wx +\varepsilon$

其中，$x\epsilon R^M$ 是符合正态分布的 M 维隐含变量：$p(x)=N(x|0,I)$。上式定义了数据 t 的生成模型：首先基于先验概率 p(x)生成隐变量的采样点 x，再通过线性变换生成 Wx，再加入一个高斯噪音 ε，最后加入位移 μ。

【PCA 是 PPCA 在 $б_{ML}\rightarrow 0$ 时的特殊形式。】

(3) 概率线性判别分析(PLDA)

提出原因：【PCA 描述数据的整体分布特性，不考虑不同类数据的不同分布。

$t=\mu+ Wx_{kj} +\varepsilon$】

LDA: μ_k 看作模型参数，$t=\mu_k+ Wx_{kj} +\varepsilon$。存在问题：一是如果数据中包括的类很多，则不仅计算开销增加，对数据的利用也不够充分（例如对那些样本很小的类，基于该模型得到的 μ_k 会产生偏差）；二是无法处理在测试时遇到的新类。

PLDA: μ_k 是随机变量(隐变量)，可形式化为：$t=\mu +Fu_k+ Wx_{kj} +\varepsilon$。下图给出了 PLDA 模型的生成过程。

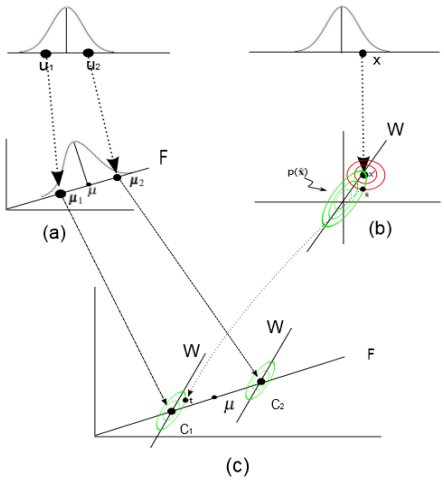


Fig. 2.7 PLDA模型，其中观察数据**t**是二维变量，类中心隐变量**u**和类内隐变量**x**都是一维。（a）首先基于正态分布采样得到类k的类别变量$\mathbf{u}_k$，经过线性变换得到在数据空间中的中心向量$\boldsymbol{\mu}_k$。（b）对每一类，依下述过程采样生成所有数据点：首先根据正态分布采样得到一维变量$\{\mathbf{x}_i\}$，经过线性变换得到数据空间中的采样$\{\mathbf{x}'_i = \mathbf{W}\mathbf{x}_i\}$，然后加入高斯噪声$\boldsymbol{\varepsilon}$，得到$\{\hat{\mathbf{x}}_i = \mathbf{x}'_i + \boldsymbol{\varepsilon}_i\}$。（c）将步骤a生成的类中心向量和步骤b生成的采样点相加，即得到该类的观察变量$\{\mathbf{t}_{ki} = \hat{\mathbf{x}}_i + \boldsymbol{\mu}_k\}$。

- **3．贝叶斯方法**

该方法将模型参数看作随机变量而非某个确定的值，通过设计这些变量的先验概率，可以将人为知识引入到模型中。更重要的是，贝叶斯方法对模型的优化不再是寻找一个最优参数（一般称为参数的点估计），而是对参数的后验概率进行估计，从而可以学习到数据分布的更多信息。

- **4．本章小结**

本章讨论了两种线性模型，一种是输入变量可预见的预测模型，一种是输入变量不可见的描述模型。

在线性预测模型中，线性回归模型假设目标变量的条件分布 p(t|x)是一个高斯分布，而 Logistic 回归模型假设 p(t|x)是一个伯努利分布。通过讨论发现，线性回归模型通过最大似然函数估计得到的回归参数与线性拟合参数是一致的，表明这两种模型具有等价性。同时，传统基于 Fisher 准则的线性分析模型在二分类问题上也等价于线性回归模型，这相当于对数据的类别标签假设了高斯分布。Logistic 回归模型将高斯分布假设修正为伯努利分布假设，更适合分类问题。

在线性概率模型中，讨论了基于非监督学习的 PPCA 方法和基于监督学习的 PLDA 方法。在 PPCA 方法中，观察数据是由一个正态分布的隐随机变量通过一个线性变换再加上一个高斯噪声生成，这意味着 PPCA（及其传统形式 PCA）只适用于符合高斯分布数据。PLDA 在 PCA 基础上考虑类间差异。这一模型假设每个类的中心向量由低维空间的一个正太分布经过线性变换得到；得到中心向量后，通过一个 PPCA 模型生成该类的所有数据。不同类数据共享同一个 PPCA 模型，因此该模型假设不同类的协方差矩阵是相同的。和 LDA 相比，PLDA 是将原来作为模型参数的类均值向量修正为随机变量。这一修正具有重要意义，它使得模型参数与数据无关，因此具有更强的泛化能力。

贝叶斯方法：模型参数不一定是一些确定的数值，还可以是一些随机变量。基于这些随机变量的后验概率，考虑各种可能的模型参数，由此可做出更合理的预测或推理。

二、知识碎片

- **1. 中心极限定理**

大数定律揭示了大量随机变量的平均结果，但没有涉及到随机变量的分布的问题。而中心极限定理说明的是在一定条件下，大量独立随机变量的平均数是以正态分布为极限的。

中心极限定理是概率论中最著名的结果之一。它提出，大量的独立随机变量之和具有近似于正态的分布。因此，它不仅提供了计算独立随机变量之和的近似概率的简单方法，而且有助于解释为什么有很多自然群体的经验频率呈现出钟形(即正态)曲线这一事实，因此中心极限定理这个结论使正态分布在数理统计中具有很重要的地位，也使正态分布有了广泛的应用。

Chapter3 神经模型

一、知识梳理

前一章我们介绍了线性模型，假设了变量之间存在简单的线性关系，忽略了现实。

非线性变换可以通过两种方式得到：一是利用先验知识手动设计；二是通过数据进行学习。总体来说，非线性变化学习方法可分为参数学习（如 ANN）和非参数学习（如 SVM）两种。

- **1. 基于映射的神经模型**

这一模型是上一章所述线性预测模型的非线性扩展。

经典一层感知器遇到线性不可分问题时，训练不能收敛。（1）解决因噪声导致的线性不可分问题时：a、阶跃输出→连续输出(如 sigmoid)；b、调整学习率保证训练收敛(及 Logistic 回归或 Softmax 回归)。（2）数据本身非线性：进行非线性扩展。几种典型的非线性变换和其对应的模型如下：

$$\phi_{nj}(\mathbf{x}) = \sum_i w_{ij}\phi_{n-1,i}(\mathbf{x}) \qquad \text{多层感知器}$$

$$\phi_j(\mathbf{x}) = \phi_j(\|\mathbf{x} - \mathbf{v}_j\|) \qquad \text{径向基函数}$$

$$K(\mathbf{x},\mathbf{y}) = \boldsymbol{\phi}(\mathbf{x})^T\boldsymbol{\phi}(\mathbf{y}) \qquad \text{核函数}$$

(1) 多层感知器(MLP)

一层扩展到多层即得到 MLP，除了结构上的扩展，标准 MLP 输出函数采用线性或 Logistic 函数。同时，这些输出函数对应不同数据分布假设，可分别对回归任务和分类任务进行建模。

训练方法：GD、SGD（最常用）、BP（BP 算法利用了 MLP 的层次结构、导数的链式法则和动态规划算法对参数进行顺序求导，避免了重复计算）。

训练技巧：（存在问题：层数增加时，BP 的消失爆炸、局部最优、过拟合）

输入/输出正规化和特征变换
选择合适的激发函数
连接权重初始化
二阶信息
使用动量
课程学习
迁移学习
正则化

(2) 径向基函数

- **2. 基于记忆的神经网络**

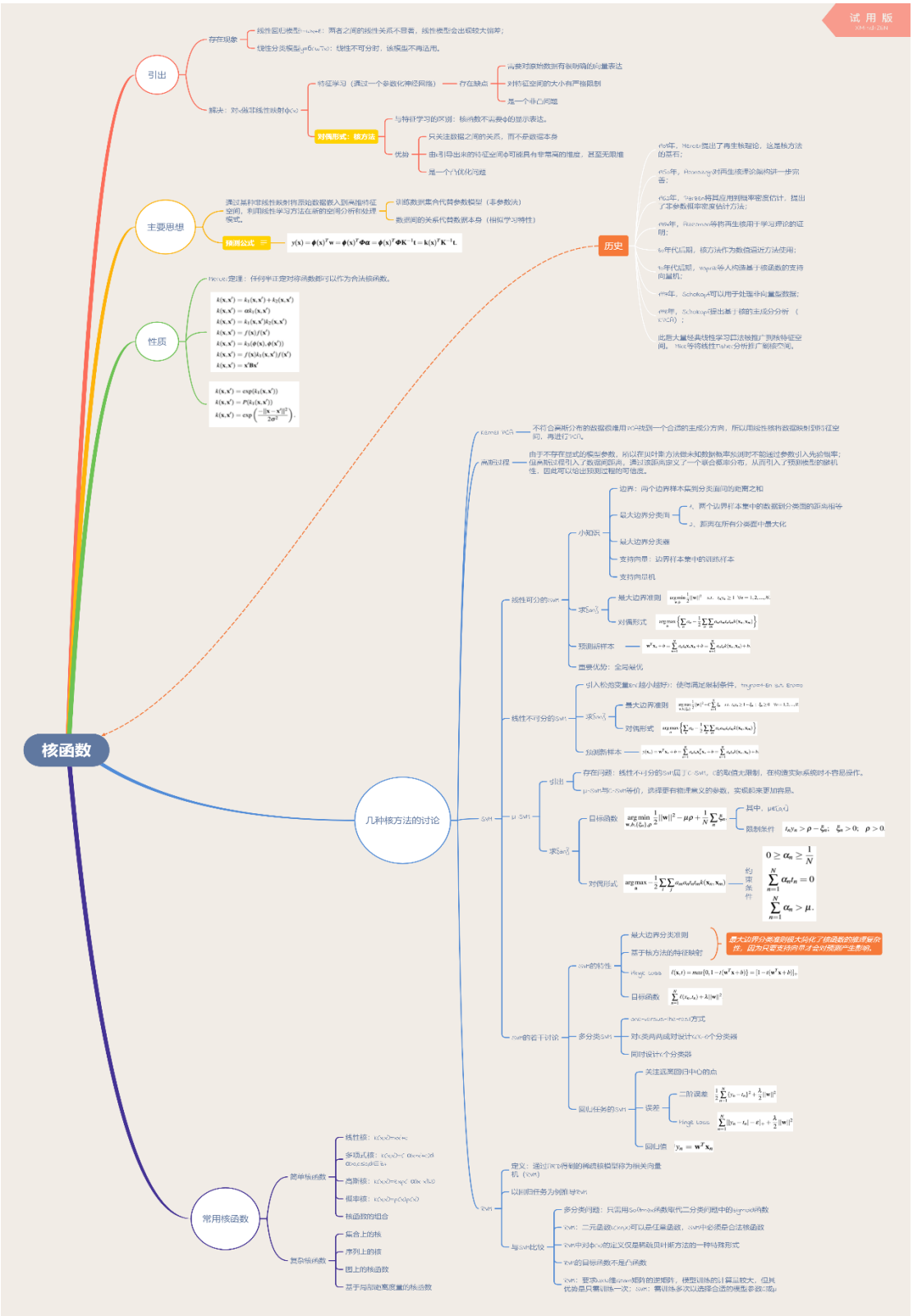
二、知识碎片

- **1. 交叉熵**
- **2. 动态规划算法**

Chapter4 深度学习

Chapter5 核方法

一、思维导图



Chapter6 图模型

神经模型：数据驱动模型，包括大量参数，需要足够的数据量以确定这些参数。

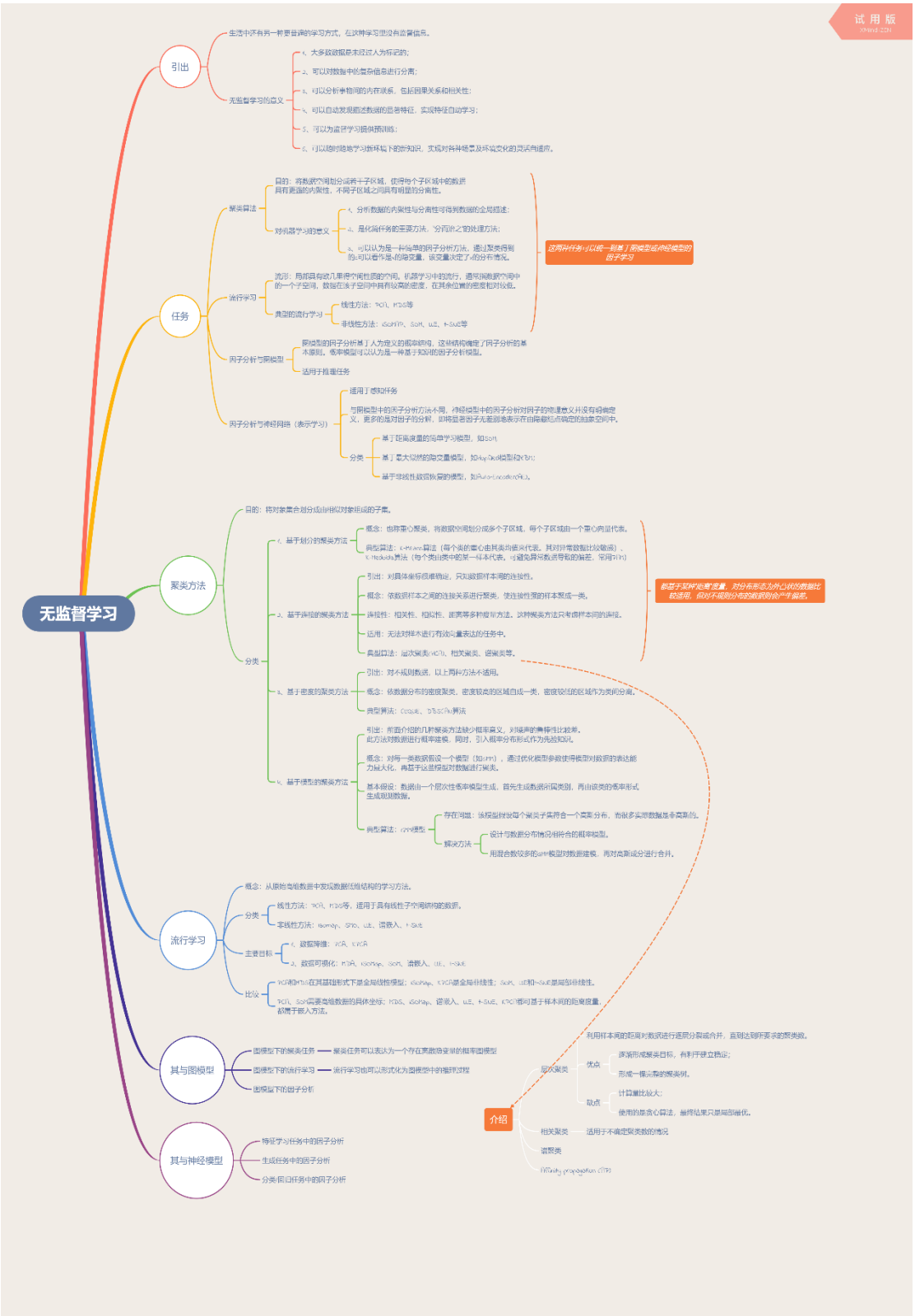
核方法：可调节的参数较少，不需要太多数据，但需要较强的先验知识来设计核函数的形式。（神经模型和核方法都缺少利用复杂先验知识的有效机制。它们很大程度是“盲学习”，对数据的因果关系分析不足。）

概率模型：是一种将领域知识和数据结合的有效工具。通过将先验知识形式化为变量间的相关性，可以对复杂系统进行有效刻画；同时，可基于数据学习模型中的参数，使之适应目标任务。

概率图模型：以图的形式来描述变量间的关系。（不仅可以更直观地描述目标问题，而且衍生出一套统一的推理方法和参数估计方法，极大简化了概率模型的构造和推理过程。）

Chapter7 无监督学习

一、 思维导图



二、知识碎片

- **1. 因子分析**

因子分析的基本目的就是用少数几个因子去描述许多指标或因素之间的联系，即将相关比较密切的几个变量归在同一类中，每一类变量就成为一个因子，以较少的几个因子反映原资料的大部分信息。运用这种研究技术，我们可以方便地找出影响消费者购买、消费以及满意度的主要因素是哪些，以及它们的影响力。运用这种研究技术，我们还可以为市场细分做前期分析。

- **2. SMO**

SMO 算法是一种解决二次优化问题的算法，其最经典的应用就是在解决 SVM 问题上。SVM 推导到最后，特别是使用了拉格朗日因子法求解之后便不难发现其最终等效为一个二次规划问题。二次规划问题有很多成熟的解法，在 SMO 算法出现之前这些解法就已经应用到了 SVM 问题的求解上。但是这些解法无论效果如何都有一个共同的缺点即是计算量太大，在小样本的情况下尚堪使用，数据量一大就变得难以奏效。1996 年，John Platt 发布了一个称为 SMO 的强大算法，用于训练 SVM 分类器。其基本思路就是一次迭代只优化两个变量而固定剩余的变量。直观地讲就是将一个大的优化问题分解为若干个小的优化问题，这些小的优化问题往往是易于求解的。

- **3. 多任务学习**

多任务学习（Multitask learning）是迁移学习算法的一种，迁移学习可理解为定义一个一个源领域 source domain 和一个目标领域（target domain），在 source domain 学习，并把学习到的知识迁移到 target domain，提升 target domain 的学习效果（performance）。

多任务学习（Multi-task learning）：由于我们的关注点集中在单个任务上，我们忽略了可能帮助优化度量指标的其它信息。具体来说，这些信息来自相关任务的训练信号。通过共享相关任务之间的表征，可以使我们的模型更好地概括原始任务。这种方法被称为多任务学习（MTL）。其也是一种归纳迁移机制，主要目标是利用隐含在多个相关任务的训练信号中的特定领域信息来提高泛化能力，多任务学习通过使用共享表示并行训练多个任务来完成这一目标。归纳迁移是一种专注于将解决一个问题的知识应用到相关的问题的方法，从而提高学习的效率。比如，学习行走时掌握的能力可以帮助学会跑，学习识别椅子的知识可以用到识别桌子的学习，我们可以在相关的学习任务之间迁移通用的知识。此外，由于使用共享表示，多个任务同时进行预测时，减少了数据来源的数量以及整体模型参数的规模，使预测更加高效。因此，在多个应用领域中，可以利用多任务学习来提高效果或性能，比如垃圾邮件过滤、网页检索、自然语言处理、图像识别、语音识别等。

归纳偏执(inductive bias)：归纳迁移的目标是利用额外的信息来源来提高当前任务的学习性能，包括提高泛化准确率、学习速度和学习的模型的可理解性。提供更强的归纳偏执是迁移提高泛化能力的一种方法，可以在固定的训练集上产生更好的泛化能力，或者减少达到同等性能水平所需要的训练样本数量。归纳偏执会导致一个归纳学习器更偏好一些假设，多任务学习正是利用隐含在相关任务训练信号中的信息作为一个归纳偏执来提高泛化能力。归纳偏置的作用就是用于指导学习算法如何在模型空间中进行搜索，搜索所得模型的性能优劣将直接受到归纳偏置的影响，而任何一个缺乏归纳偏置的学习系统都不可能进行有效的学习。不同的学习算法(如决策树，神经网络，支持向量机等)具有不同的归纳偏置，人们在解决实际问题时需要人工地确定采用何种学习算法，实际上也就是主观地选择了不同的归纳偏置策略。一个很直观的想法就是，是否可以将归纳偏置的确定过程也通过学习过程来自动地完成，也就是采用“学习如何去学(learning to learn)”的思想。多任务学习恰恰为上述思想的实

现提供了一条可行途径，即利用相关任务中所包含的有用信息，为所关注任务的学习提供更强的归纳偏置。

- **4. Laplacian matrix 的定义**

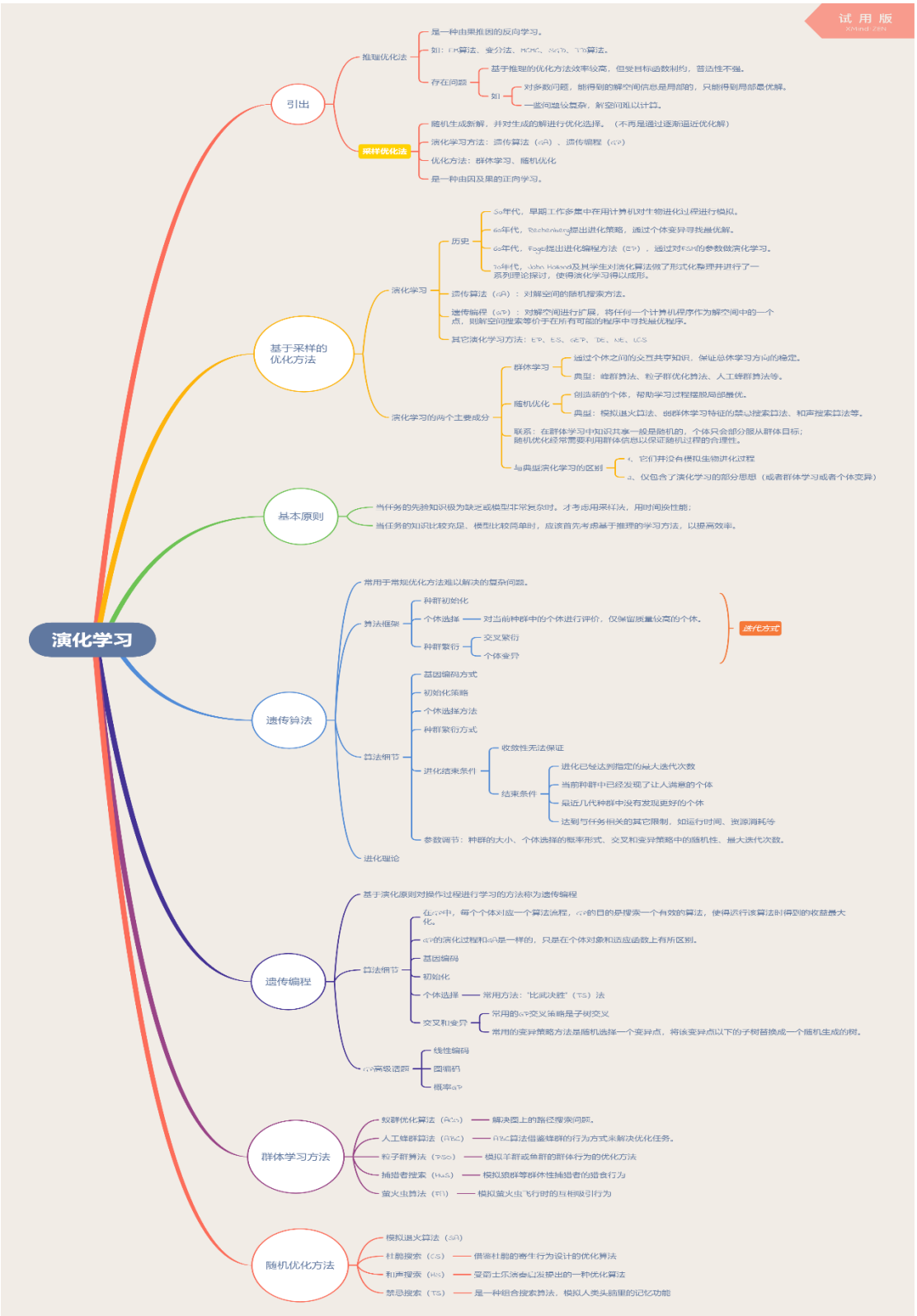
拉普拉斯矩阵（Laplacian matrix)，也称为基尔霍夫矩阵，是表示图的一种矩阵。给定一个有 n 个顶点的图 G=(V,E)，其拉普拉斯矩阵被定义为: L=D-W。其中 D 为图的度矩阵，W 为图的邻接矩阵。

Chapter8 非参数模型

一、思维导图

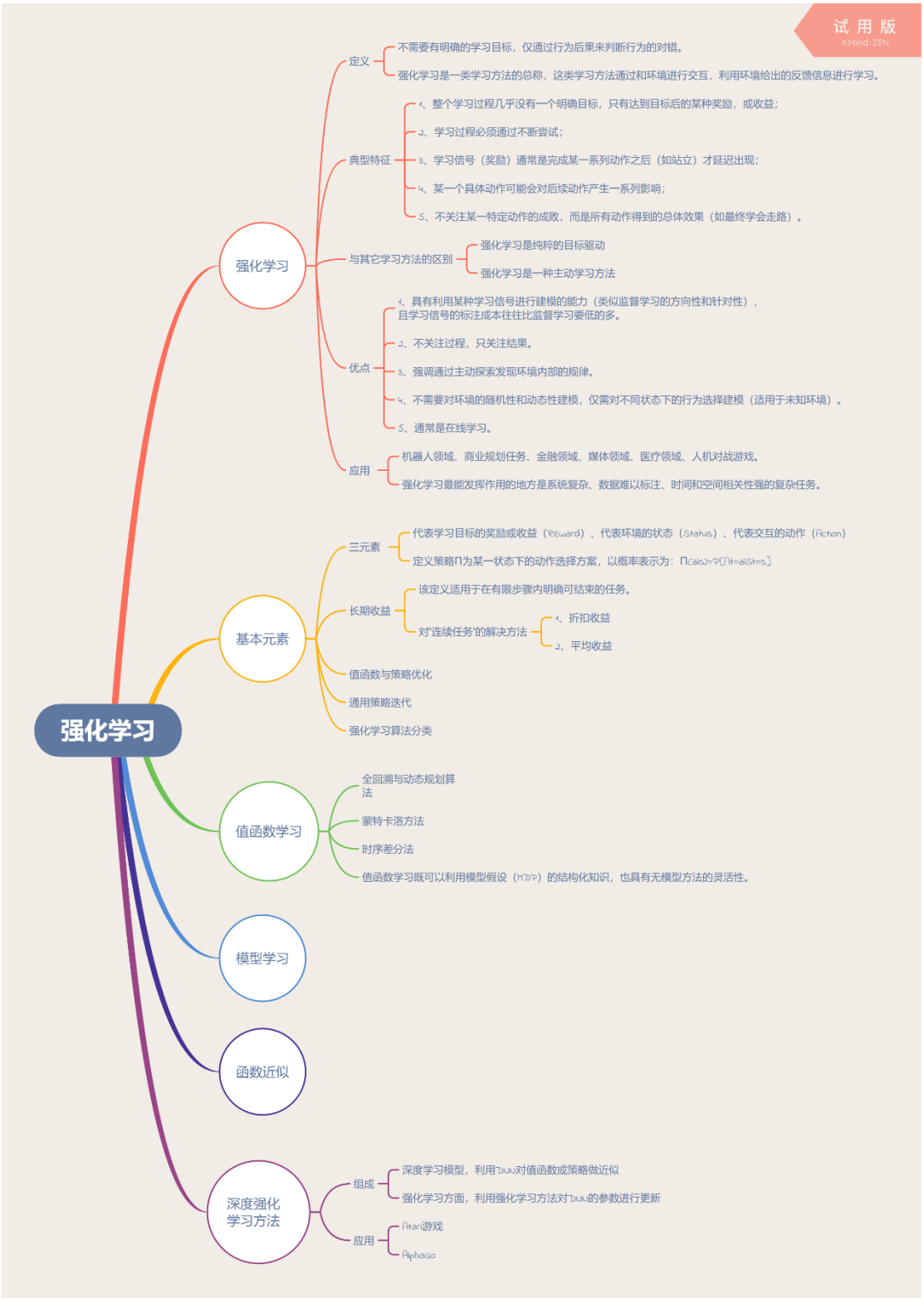
Chapter9 演化学习

一、思维导图



Chapter10 强化学习

一、思维导图



Chapter11 优化方法

一、思维导图