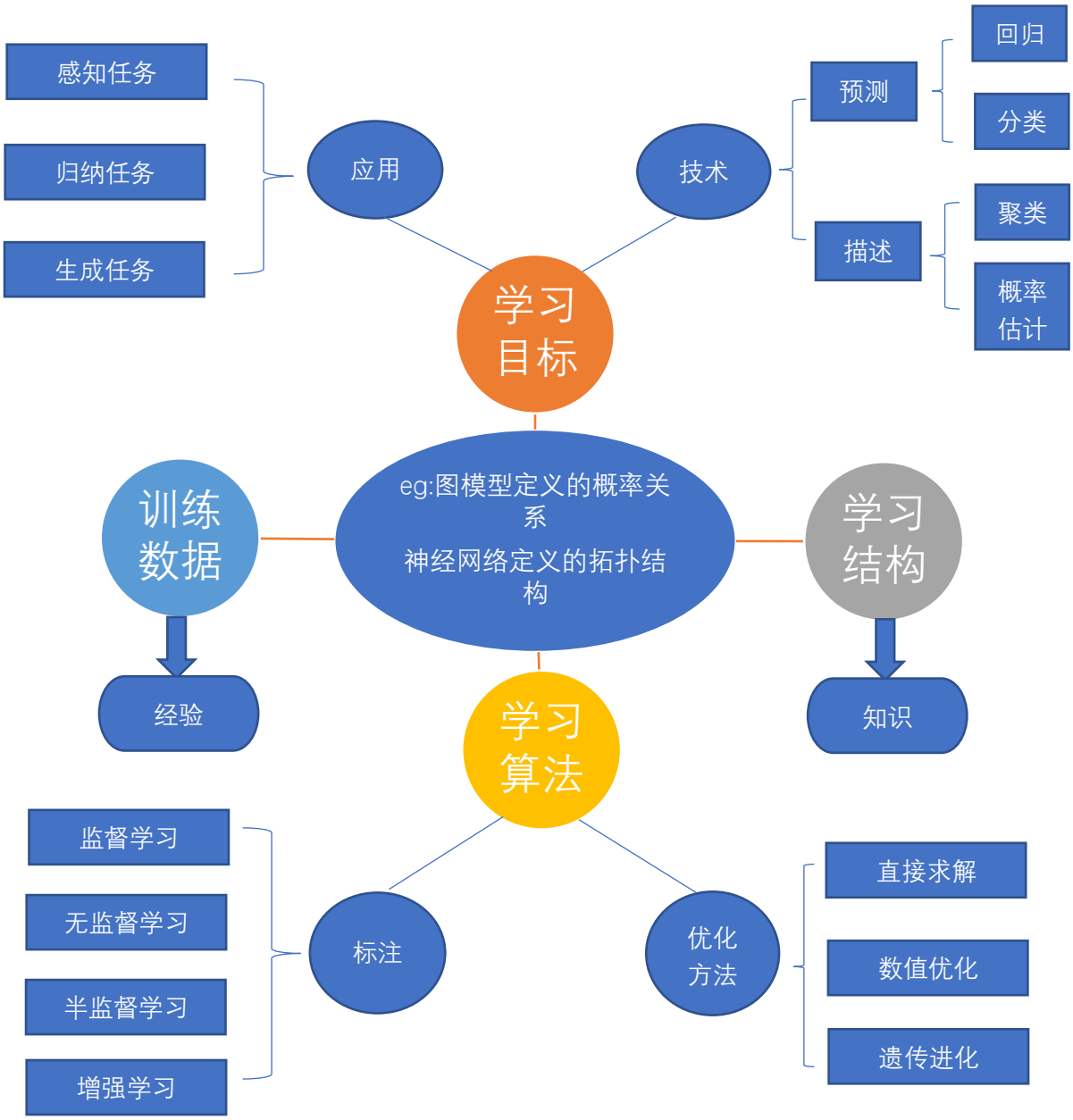
Chapter1 机器学习概述

机器学习的基本框架

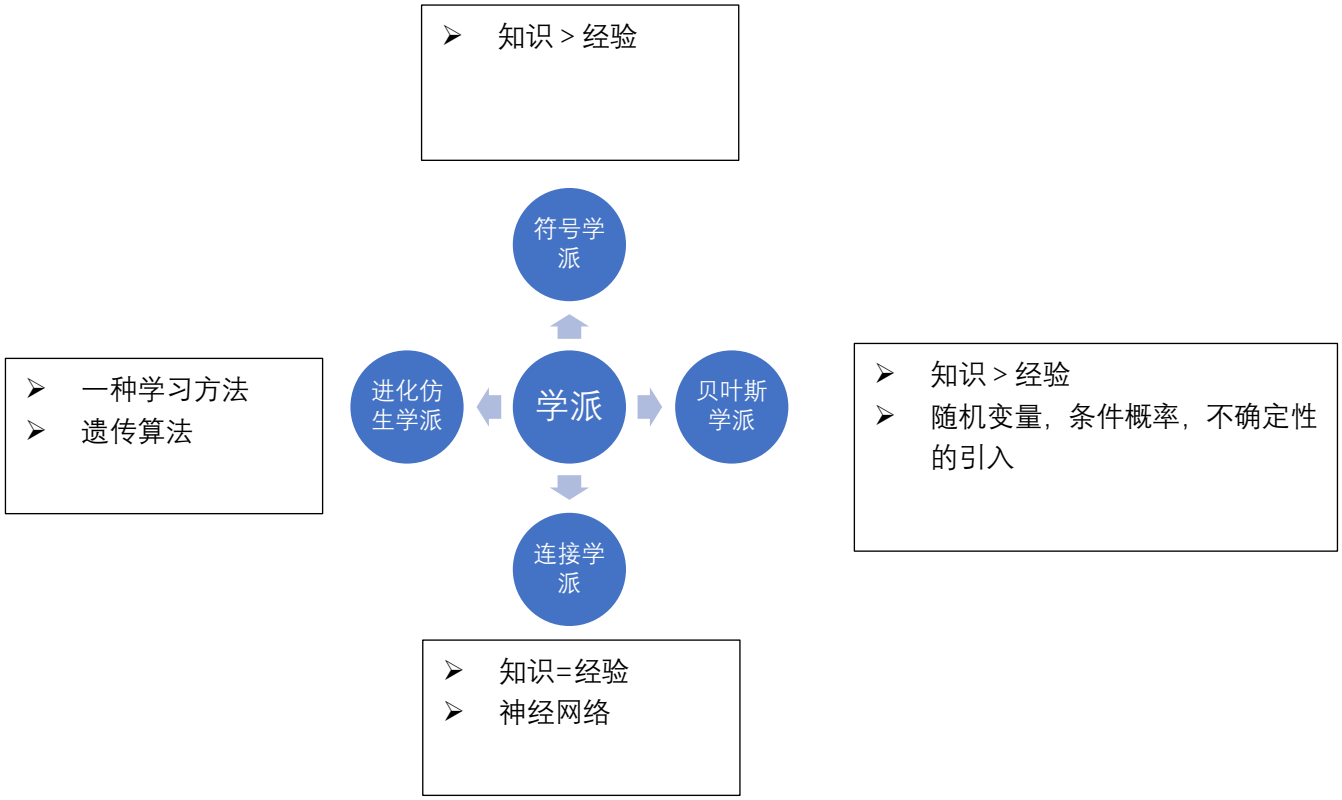


原文摘录

P2——计算机程序如果通过某种方法，利用经验 E，提高在任务 T 上的性能（以 P 为评价标准），则可认为该程序从经验 E 中进行了学习。

P2——设计者不必定义具体的流程细节，只需告诉机器一些通用知识，定义一些足够灵活的通用结构（如图模型定义的概率关系，神经网络定义的拓扑结构）

机器学习流派



原文摘录

- P10——在贝叶斯学派看来，只需将表达事件的两个随机变量之间的条件概率关系确定，所有时间将会组成一个相互连接的网络
- P12——连接学派和贝叶斯学派都依赖一个结点网络，不同的是在贝叶斯学派中，网络中的每个结点都有清晰的定义，而连接学派网络中的结点模仿神经元，是同质的，不代

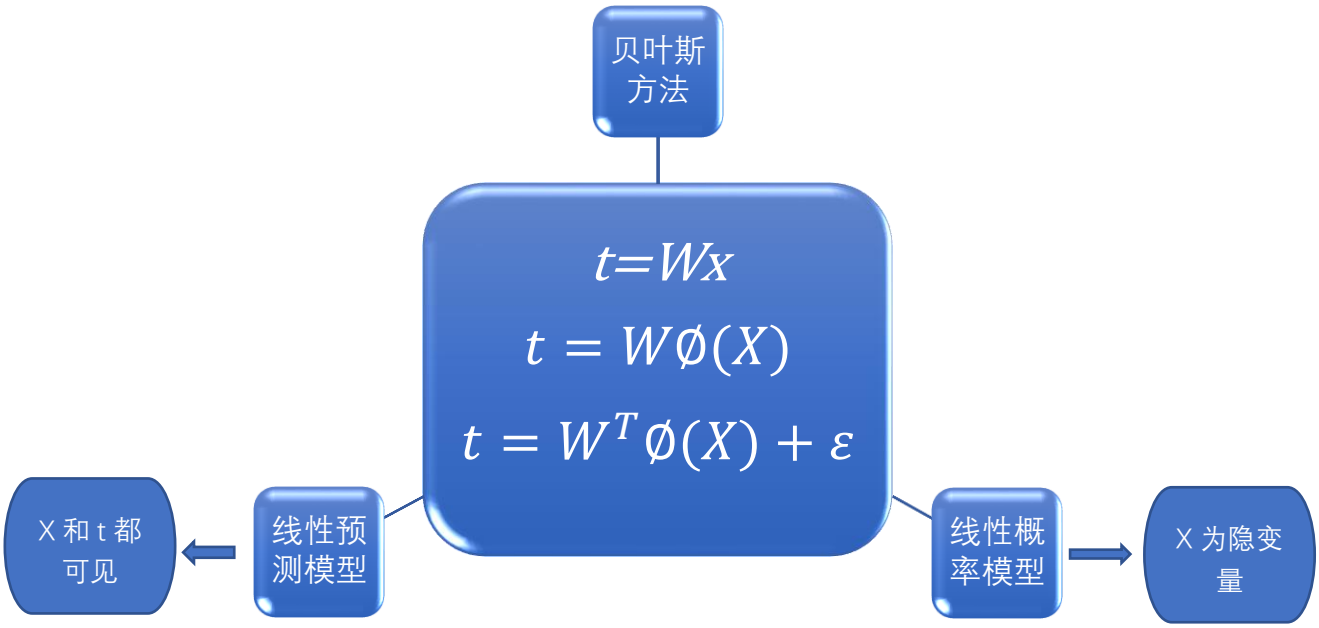
表具体事件；另一方面，贝叶斯学派中的节点都是随机变量，有概率意义，而连接学派中的节点更像计算结点，较少具有概率意义。

- P12——它将复杂事件之间的关系统一到概率框架中，将演绎过程归结为边际概率计算，将推理过程归结为后验概率计算。

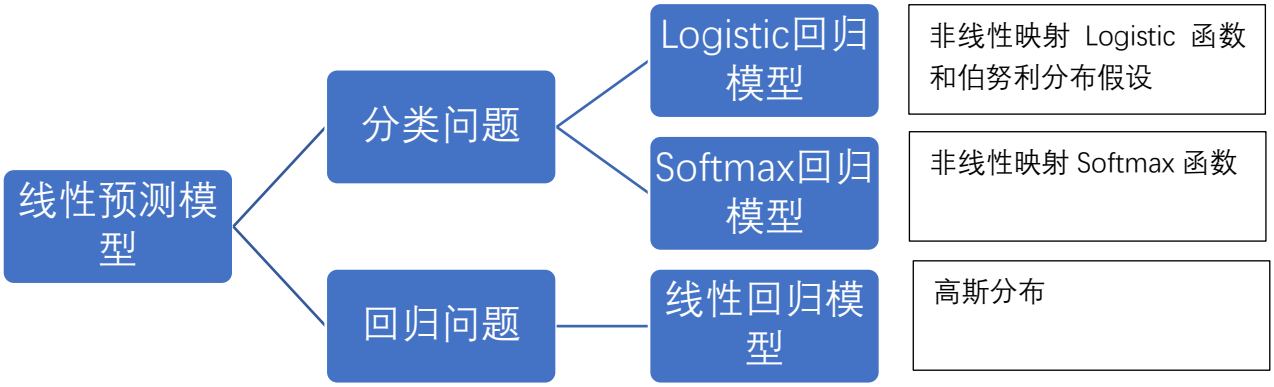
问题总结

P9——符号学派是什么内容没看懂
P35——生成模型和区分性模型的逻辑关系没理顺

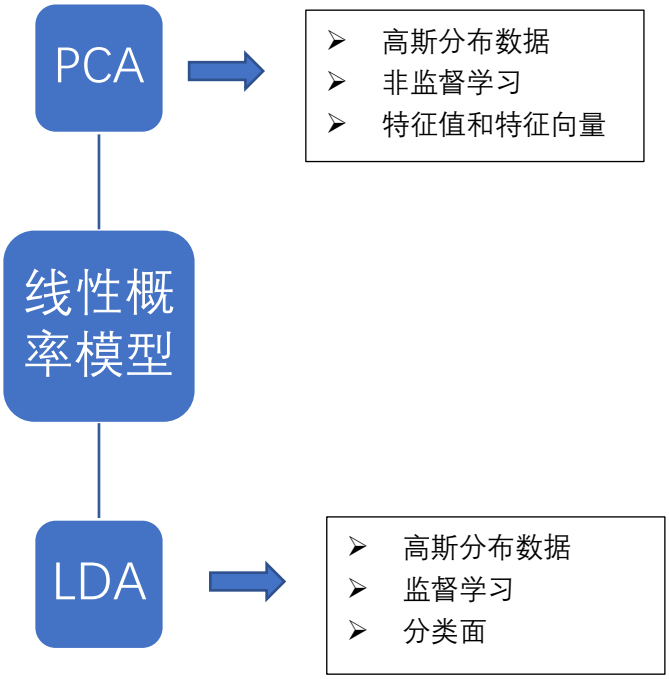
Chapter2 线性模型



线性预测模型



线性概率模型



贝叶斯方法

核心公式$\mathrm{p(w|D)} = \frac{p(D|w)p(w)}{p(D)}$

原文摘录

P51——最大似然估计事实上等价于线性拟合。

线性拟合中的平方误差事实上假设了目标观察值中的噪声符合高斯分布。

P53——当数据具有很强的长尾特性时，考虑 Student-t 分布和拉普拉斯分布。一般地，高斯分布。（例如语音识别里的 HMM-GMM）

问题总结

- P50——Fig 2.2 的表示方法没看懂
- 2.1.3 没看懂
- 公式……

Chapter 3 神经模型

新奇事物

- 感知器收敛定理

  感知器模型：$\mathrm{y} = \mathrm{g}\left(\sum_{i=0}^{D} w_i x_i\right) = \mathrm{g}(W^T \mathrm{X})$。其中 g()为阶跃函数。感知器目标函数包含阶跃函数，不能直接用梯度下降法进行优化。解决方法是不考虑阶跃函数，仅考虑线性预测部分，得到误差函数$\mathrm{L(w)} = -\sum_{n\in\mu} w^t x_n t_n$。

  参数更新后μ可能会发生变化，因此误差函数可能会发生阶跃性变化。但人们研究发现，上述迭代过程在任何一个线性可分的数据集上经过有限步迭代后，都可确保收敛到一个对该数据集完美可分的分类器，成为感知器收敛定理。
- 近似定理

  如果激发函数选择适当（如 Sigmoid，Tanh，Relu 等），当隐藏层的神经元个数足够多，包含一个隐藏层的 MLP 可以有效模拟一切连续函数。这一结论成为一般近似定理。
-

原文摘录

P76——对人工神经网络的研究可分为两个方向。一部分研究者集中研究如何描述人类大脑的实际运作方式，如激励方式、抑制机理、传导模型等；另一部分研究者更关注神经网络的表达能力，关注通过神经网络可实现的功能，至于该网络是否对应真实神经系统则不是核心

内容。机器学习中对神经网络的研究主要采用第二种思路，设计各种结构来提高网络对数据的建模能力和推理能力。
P79——线性回归模型的最小平方误差估计等价于假设目标变量的观察值 t 为以 y 为中心的高斯分布时的最大似然估计。
假设观察值（分类标记）是以 y 为参数的伯努利分布，则该模型的最大似然估计等价于一个以交叉熵为准则的优化问题。

有感而发

- P76 的那段话超级喜欢，人类大脑神经元的同质性只是灵感，而人类的研究应该主要面向结果，就是神经元如何连接如何产生效果都可以自行设计，完全还原人类大脑结构是不现实也是无意义的。
-

问题总结

- P78——提到的符号方法还是比较模糊
- P83——在第二章介绍线性模型时，提到对 x 进行非线性变换，但该变换是固定的，不需要学习，因此模型依然是线性的。第三章讨论的非线性变换需要对变换函数进行学习，因而模型是非线性的。
- P84——在回归任务中，激发函数一般取线性 g(x)=x；分类任务中一般取 Logistic 或 Softmax 函数？回归任务取线性？
- P85 ——SGD 相较于 GD 的好处？
- DNN 相较于 MLP 是否就是层数上的加深。MLP 就是一层隐层
- P86——对梯度反向传播算法的解释不太清楚。
-

Chapter 4 深度学习

新奇事物

Fine tuning

原文摘录

- P141——图 4.10 所示的 DNN 模型分为特征提取模型和 Softmax 分类模型两部分，前者通过非监督学习得到，后者通过监督学习进行优化。由此可见，虽然 DNN 在形式上和传统多层神经网络没有太大区别，但在概念上已经发生根本变化：其基本原则已经不

仅是以任务最优化为目标的分类/回归函数的近似，而是对原始、粗糙特征的层次化学习。

- P143——更深刻的变化是，当我们对图 4.10 所示的 DNN 网络进行整体优化时（即不仅对 softmax regression 分类模型进行优化，而且将误差信息向特征提取层反向传递，从而对包括特征提取和分类模型在内的整个网络进行优化），事实上是将特征和模型视为一个整体进行共同学习，这一过程成为 fine tuning。
- P143——深度学习是一种将特征提取和分类模型有机结合在一起的学习方式，是一种将非监督学习和监督学习统一在一起的学习方式。这种新的理解推翻了传统学习框架中将特征提取和统计建模独立对待的二分思路，而是将两者统一到多层神经网络的层次学习结构中：越是网络底层，越重视特征提取，越是网络高层，越重视目标任务建模。这种将特征和模型统一学习的方式有时也成为“端到端学习”。
- P144——深度学习极大改变了这一传统认识，人们不再追求过多的人为干涉，转而相信数据的自我代表性，让机器自动从数据中发现知识和规律，实现更大胆的数据驱动。这一观念的转变显然源于神经模型的普适性：将一些同质神经元通过简单的网络结构连接在一起，只要数据足够充分，即可超过人们精心设计的各种特征提取方法和复杂的概率模型，这事实上表明数据中蕴含的知识已经超过人为设计所能覆盖的范围。因此，数据超过人为设计作为主要知识源已经成为共识。

有感而发

- 感觉图像识别模型好做端到端是因为图像信息相较于语音信号更好做数字化存储。图像的每个像素数字就可定位一张图片，而语音信号是时序信号需要分帧提取特征，数字储存上更难操作。
- 导数&偏导数&方向导数&梯度
  1. 导数
     一元函数：$y = f(x)$。导数$f'(x_0)$就是函数在某点的切线斜率。曲线在某点的切线有且仅有一条。所有以一元函数举例的梯度下降都是没有意义的!!! 因为变量只有一个，根本没有方向的问题!!! 终于找到症结所在了!!!
  2. 偏导数
     多元函数：$z = f(x, y)$。曲面的一点切线有无数条。偏导数是多元函数沿着坐标轴的变换率，但是很多时候我们要考虑多元函数沿任意方向的变化率。
  3. 梯度
     梯度方向是方向导数最大的方向。

问题总结

- P143——既然图像识别的端到端学习已经非常完善了，但是语音识别的端到端学习还是不够完善，感觉可以有很大发挥空间。终于知道自己思路混乱的原因了！图像识别的特征提取和分类是杂糅的，但是早期的 HMM-GMM 模型和特征提取阶段是分开的，因为图像识别和语音识别的建模发展处在不同的阶段!
- P145——梯度下降最大的方向的意思是？？？
- P146 关于梯度下降和随机梯度下降和 Mini-Batch 优缺点的论述不明所以。

- P147——SGD 的一个显著缺陷是在优化过程中对不同参数的学习率相同？？？
- P149——BFGS 算法
- P186——迁移学习是不是就是别的训练好的神经网络直接拿来用
- P188——P（t）和 P（t | x）的分布形式
- P199——梯度公式仅包括乘法和加法，这一关系并不是因为图 4.43 表示的嵌套函数具有乘加性质，而是因为 chain rule 在计算梯度时，不同层次嵌套函数的导数间存在乘法关系，而同一函数不同变量的梯度间存在加法关系。
- P206——关于 GPU 和 TPU 使用的原因，不太懂

Chapter5 核方法

新奇事物

Gram 矩阵
原始问题&对偶问题
核函数&核方法
再生核希尔伯特空间（RKHS）

原文摘录

- P232——首先该方法只关注数据之间的关系，而不是数据本身，因此特别适合数据样本难以用向量明确表示的任务。其次，由$\mathrm{k}(\mathrm{x},\mathrm{x}')$引导出来的特征空间$\emptyset$可能具有非常高的维度，甚至是无限维，因此可以满足对复杂数据分布的线性化要求。最后，特征空间中的模型是线性的，因此模型训练是一个凸优化问题，可保证得到全局最优解。
- P235——在这种学习中，用训练数据集合代替参数模型，用数据间的关系代替数据本身，前者使得该方法是一种非参数方法，后者使得该方法具有相似性学习的特征。
- P236——任何一个对称半正定的二元函数对应唯一一个 RKHS。
- P237——可能有多个$\emptyset(x)$对应同一个核函数。
- P237——Mercer 定理
- P238
  1. **线性核**：对应等值映射，不能提高表示性
  2. **多项式核**：等价于对原始数据进行特征扩展，缺点是当阶数 d 比较大时，在取值时容易出现数值上的不稳定，可能出现过大或过小值。
  3. **高斯核**：高斯核对应无限维特征空间。
  4. **概率核**
- P253——可得到全局最优解是 SVM 相对神经网络和很多其他模型的重要优势。

问题总结

- P232——研究表明，$\mathrm{k}(\mathrm{x},\mathrm{x}') = \emptyset(x)^T\emptyset(x')$，任何一个对称半正定的函数都是核函数？
- P236——映射$\emptyset(x)$将原始空间 x 映射到了一个函数空间 H，而 H 中的函数通常可看作是无限维向量，这相当于通过核函数将原始数据映射到了一个无限维空间中？

Chapter6 有向图模型

新奇事物

信任网络/贝叶斯网络 & 马尔科夫随机场
Explain Away
D-Separation
Clique
势函数
Moralization
Moralized Graph
马尔可夫性质
边缘化
加和-乘积算法
联合树算法

有感而发

- 关于鉴别性模型和生成性模型的区分终于有了大概的理解，就是那句“鉴别性模型对分类面建模，生成性模型对数据分布建模。“而这种区别是和任务紧密相关的。可以理解为回归和分类问题的不同吗？也就是说任何模型都可以是鉴别性也可以是生成性？
- 神经网络和混合高斯模型在数据拟合方式和能力是没有太大区别：GMM 相当于层峦叠起的山丘地带，有很多鼓包，代表很多高斯组成成分，模型设计上先验成分更多；而神经网络所代表的数据空间也是很多局部最优解，也相当于很多沟壑，这样就好理解了。
- EM 算法和 BW 算法的关系
  EM 算法在每一步优化时设计一个具有闭式解的下界函数，用迭代法求该下界函数的最优参数，从而实现对原似然函数的优化；然而优化下界函数需要对大量可能的状态路径进行边缘化，计算效率低，BW 算法基于动态规划原则对路径进行整理，避免重复计算。

原文摘录

- **强化学习的特征**

1) 强化学习是一种弱标记学习
2) 强化学习是一种主动学习方法，通过主动和环境进行交互产生学习样本。
3) 与其关注每个具体状态的具体动作，不如关注动作组合产生的总体效果，这正是强化学习的优势。
4) 强化学习最能发挥作用的地方是系统复杂、数据难以标注、时间和空间相关性强的复杂任务。

- P266——图模型本身更关注变量间的拓扑结构，而不是变量间概率关系的具体形式。不同拓扑结构的推理和参数估计方法之间有很大区别，图模型关注同一拓扑结构下的通用算法。
- （1）通常用灰色节点代表观测变量，用白色节点代表隐藏变量（2）通常用圆圈代表连续变量，用方框代表离散变量（3）有时会对某些节点加框，表示一组独立同分布的变量（4）模型的参数通常用一些实心的黑点表示。
- P268——图模型打破了监督学习和非监督学习的界限，将这两种方法统一到一个概率学习和推理框架之中。
- P285——和 GMM 模型类似，HMM 模型也可以表示成一个有向概率图。
- P285——一般离散观测值的条件概率是个 Multinomial 分布，连续观测值的条件概率多采用 GMM 模型。以 GMM 为输出概率的 HMM 可以认为是 GMM 模型的多状态扩展，而 GMM 可认为是只有一个状态的 HMM。

问题总结

- P266——“图模型本身更关注变量之间的拓扑结构，而不是变量间概率关系的具体形式。不同拓扑结构的推理和参数估计方法有很大区别，图模型关注同一拓扑结构下的通用算法。“
- P266——概率独计性？
- P277——最后一段没看懂
- P281——最后一段没看懂
- P283——对 GMM 参数的估计方法没有完全看明白
- P283——贝叶斯公式推导过程可以看懂，但是不能感性理解
- P286—— $\mathrm{L}(\theta)=\sum_{n=1}^{N} lnp(x_n;\theta)$ 显然这一似然函数不是一个凸函数；$\tilde{L}(\theta)=\sum_{n=1}^{N}\sum_{q} p(q|x_n;\theta')\ln p(x_n,q;\theta)$是一个凸函数？？？
- 6.4.3
- P294——EM 算法的讲解

Chapter7 无监督学习

新奇事物

流行（Manifold）
重心聚类法

聚类粒度

有感而发

- P326 对 GMM 模型做了另外一个角度的解释: 第六章是图模型角度, 这里是聚类角度; 换句话说图模型中的隐变量就是一个类的话，那就可以看作是聚类也就是无监督学习。感觉很多小节的内容都是你中有我我中有你。

原文摘录

- P323——无监督学习的本质并不是有没有数据标注，而是是否对数据做了因子分析。
- P324——与图模型中的因子分析方法不同，神经模型中的因子分析对因子的物理意义没有明确定义，更多是对因子的分解，即将显著因子无差别地表示在由隐藏结点确定的抽象空间中。
- P327——Minkowski 距离一般适用于无约束数据。对有约束数据，则需对距离做特殊定义。例如当数据分布在一个球面上时，一般采用 Cosine 距离。Cosine 距离在 NLP 领域里有广泛应用。
- P333——直观上，谱聚类是将数据通过连接矩阵映射到某一特征空间，之后在该特征空间完成聚类。这种通过样本间距离对样本进行特征映射的思路正是核方法的主要特征。
- P356——流行学习有两个主要目标：一是数据降维，二是数据可视化。PCA、KPCA 一般被认为是降维工具，MDS、ISOMap、SOM、谱嵌入、LLE、t-SNE 一般被用做可视化工具。
- P357——PCA，SOM 需要高维数据的具体坐标；MDS，ISOMap，谱嵌入，LLE，t-SNE，KPCA 都可基于样本间的距离度量，不需要数据的具体坐标，因此都属于嵌入方法。

问题总结

- P333——不同维度的相对值代表不同点成为类中心点的先验概率，整体对角元素的大小决定聚类算法对聚类数目的倾向性，值越大，产生的聚类数越多?
- P343——如果要取低维的x_i，只需要将…中的最小特征值置为 0，使分解误差最小。
- MDS 没看懂
- 最终的策略是要确定在某个状态下采取什么动作, 感觉动作值函数直接相关而状态值函数并无较大意义啊。。。
-

Chapter8 非参数模型

新奇事物

贝叶斯非参数模型方法
狄利克雷过程

原文摘录

- P367——参数 P 决定了模型 M，对模型进行优化的过程即是对 P 的选择过程。
- P368——高斯过程和迪利克雷过程都基于概率图模型，通过设计无限维空间上的先验概率，实现依训练数据对模型复杂度进行调整，同时保证这种调整不因过度依赖数据而产生太大偏差。

- P370——高斯过程定义了另一种先验，这一先验不是基于某种模型形式的参数上的先验，而是某类映射函数的先验。
- P371——高斯过程本质上是一个随机函数，描述函数的不确定性。然而，这一随机函数的变量有无限维。
- P372——协方差越大，则不同样本间的相关性越强，函数取值越倾向于一致，对应的函数曲线越平坦。
- P380——不论是 GMM 还是 LDA，都需要设定模型中高斯成分或主题的个数，实际应用中，我们很难判断数据应分成多少类或多少个主题；另外，GMM 和 LDA 中的高斯成分或主题数一旦确定后即很难改变，不易随数据规模增大而增加类别或主题。
  一种思路是对数据集的不同聚类方式赋予一个先验概率，从而可以利用概率图模型的推理方法得到聚类方式的后验概率。

问题总结

- P371——由于 W 是高斯的，因此 Y 也是一个高斯分布？嗯？
- P376 图 8.3 &P377 图 8.4
- 8.3.2——中国餐馆问题中第 N 个顾客以概率…坐在第 K 张桌子上，概率加起来等于 1 嘛？
- P384——狄利克雷分布的解释不是很清晰
- Fig 8.8 Dir(4,4,2)？

Chapter9 遗传算法

新奇事物

- 适应函数——在优化任务中，这一函数代表某一解法的优劣，一般取任务的目标函数作为适应函数。

原文摘录

- P404——GA 算法通常用于常规的优化方法难以解决的复杂问题。
- P409——GA 算法一般仅适用于适应函数比较简单的任务，如果这一条件得不到满足，可能需要考虑用其它优化方法或建模方式。
- P409——一些研究表明，交叉策略可能并不必要，在一些任务中仅有变异策略也可以得到较好的效果。
- P412——可以证明，GA 中的个体选择过程事实上是以一定概率对所有超平面进行同步采样的过程，其中每个超平面获得样本点的概率正比于该超平面的适应函数。

Chapter10 强化学习

新奇事物

长期收益
加合收益
折扣收益
平均收益
多轮任务&连续任务
贪心策略
通用策略迭代算法（GPI）
回溯
Bootstrapping 方法

原文摘录

- P448——$\pi(a|s) = P[A_t = a|S_t = s]$，该式意味着该策略是时不变的，即在任何时刻，只要系统处在 s 状态，其采取的动作符合同一分布$\pi(a|s)$。

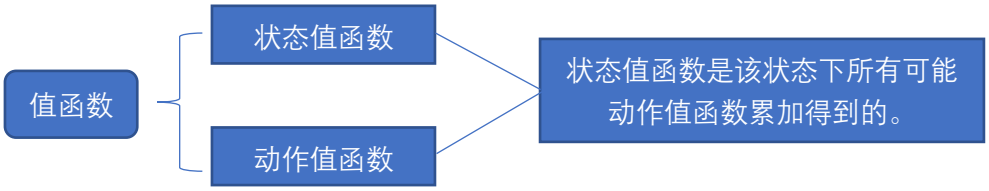
- P450——强化学习中一个重要任务是给定某个策略π，计算V_π和Q_π，这一任务通常称为“策略评价”。
- P450——如果能发现一个优化策略，则该策略对应的值函数必然是优化函数。反过来如果某一策略对应的值函数是最优值函数，则该策略必然是最优策略。在一定条件下（如马尔可夫决策过程），可以证明对一个强化学习任务，这一策略是存在的，并可以通过

  最优值函数构造出来（V(S)和 Q（s,a））
- P451——大多数强化学习算法都是基于值函数的。
- P451——相比强化学习，演化学习仅通过策略所产生的结果来判断策略的优劣，这种选择型优化不考虑交互过程中每个状态的估值，效率要低很多。
- P453——强化学习可以基于一个已知模型进行学习，也可以基于实际经验学习，前者一般称为“规划任务”，后者一般称为“学习任务”。
- P459——回溯是强化学习的基本概念，回溯方法上的不同是各种学习方法的主要区别。
- P463——动态规划（DP）方法是一种全回溯值函数计算方法，这一方法之所以可以进行全回溯，是因为环境动态模型已知。
- P464——**蒙特卡洛方法：**依概率采样所有可能的交互序列，并依$V_\pi(s)$求所有路径长期收益的均值。

有感而发

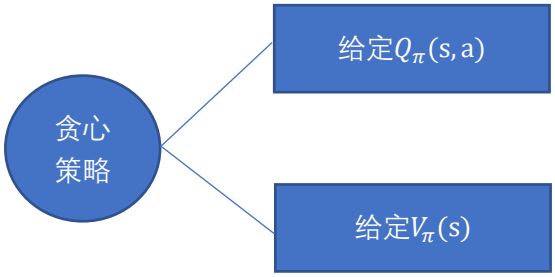
- 强化学习要做的就是定义一种策略，这种策略的具体表现就是在某种状态下采取某种行为的概率，这一策略使最后的获益最大（和损失函数相对应）。定义了一种策略，针对系统某一状态采取什么动作是随机的，而采取动作后，系统会反馈什么状态也是随机的。
- 在某种意义上是一种分类问题？根据当前输入状态得到输出动作的概率，动作有几个是确定的，输出得到每个工作的概率，只是输出的动作会对当前的输入产生影响，环环相扣。
- **马尔可夫决策过程和马尔可夫过程** 马尔可夫决策过程和马尔可夫过程原理是一样的，不同的是前者加入了动作和收益，下一个状态是动作的结果，而后者的状态转移概率有特定的分布，前者的动作以另一条主线影响着状态链。
- **全回溯和全路径回溯**是不同的。全回溯是一步考虑所有可能序列，而全路径回溯是回溯到初始时刻状态
- 一些关系

  a) 状态值函数和动作值函数



这两个值函数都是策略π的函数。强化学习中一个重要任务是给定某个策略π，计算V_π和Q_π，这一任务称为策略评价。
大多数强化学习算法都是基于值函数的。

  b) 通用策略迭代算法（GPI）



动作值函数和状态值函数优化策略的不同本质上源于状态和动作这两个元素不同性质。动作是策略的执行行为，因而动作值函数的优化表现为最优动作的选择；而状态是策略执行的结果，因而对状态值函数的优化应考虑到执行过程的细节。

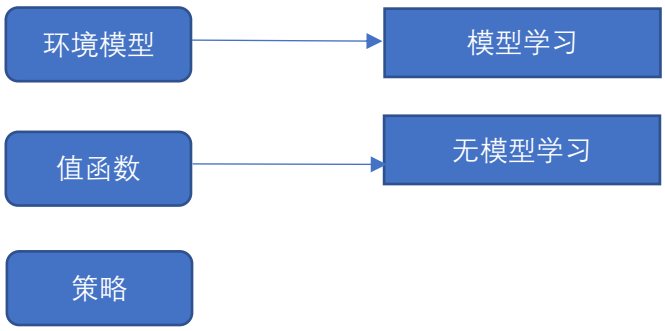
GPI：

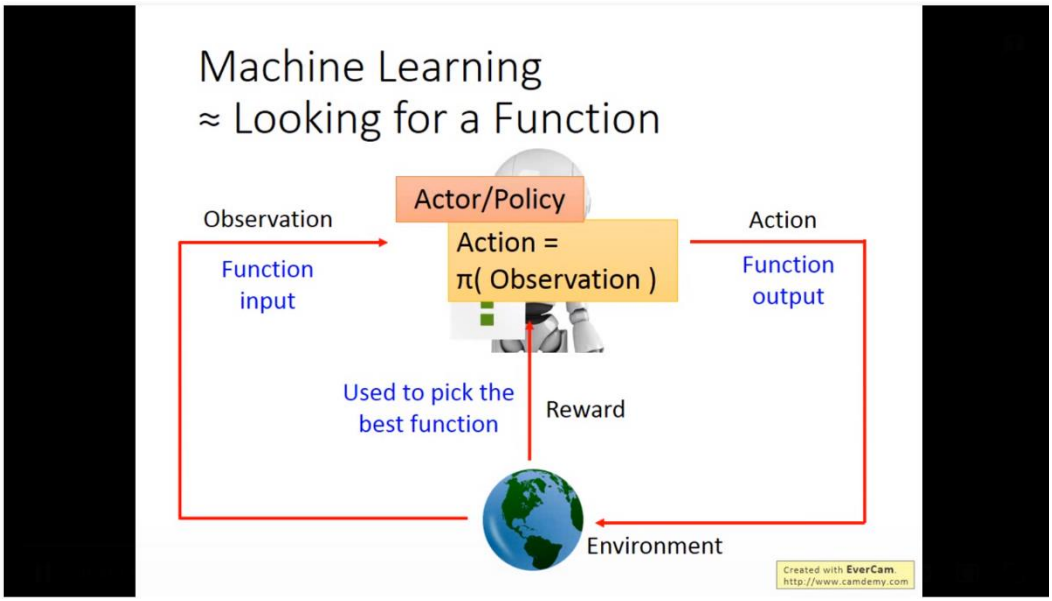
1. 设定随机策略π
2. 确定值函数 V 或 Q
3. 贪心原则改进策略

值函数和策略交叉优化并迭代改进π_0，V_1，π_1，V_2 …

- 强化学习分类

规划任务基于模型，学习任务基于采样





问题总结

- P448——为此，我们需要定义长期收益，并依此确定策略间的偏序关系。
- P455——模型规划和模型学习的图
- P456——为保证重复进入某一状态的值函数相等，一般选择折扣收益作为长期收益。
- P463——最上面不懂

Chapter11

新奇事物

- 泰勒定理

  一元函数：$\mathrm{f(x)} = \mathrm{f}(x_k) + (\mathrm{x} - x_k)f'(x_k) + \frac{1}{2!}(x - x_k)\,f''(x_k)$

多元函数：$\mathrm{f(X)} = \mathrm{f}(X_k) + \nabla f(X_k)^T + \frac{1}{2!}(X - X_k)^T\ \mathrm{H}(x_k)(X - X_k)$

- 线性搜索：$x_{k+1} = x_k + \alpha_k p_k$
  1. **GD**：p_k方向为梯度方向时函数值变化最大
     证明：$\mathrm{f}(x_{k+1}) - \mathrm{f}(x_k) = \mathrm{f}(x_k + \alpha_k p_k) - \mathrm{f}(x_k) \approx \alpha_k {p_k}^T \nabla f_k$
  2. **牛顿法**：利用目标函数的二阶曲率信息自动设置搜索步长
     公式：$x_{k+1} = x_k - (\nabla^2 f_k)^{-1} \nabla f_k$
  3. **拟牛顿法**：Hessian 矩阵的计算过于复杂

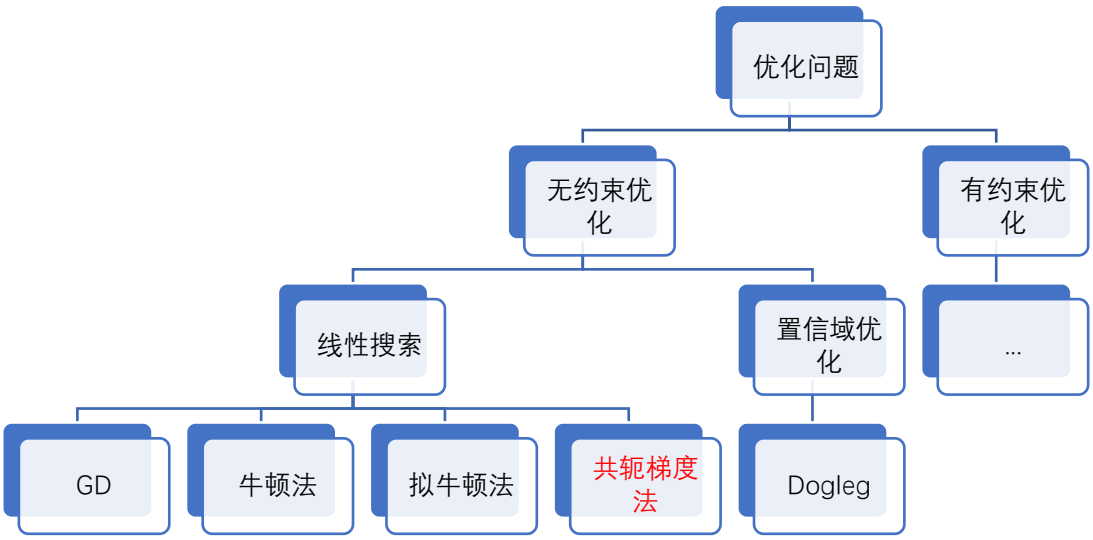
     SR1 更新法： $B_{k+1} = \boldsymbol{B_k} + -$

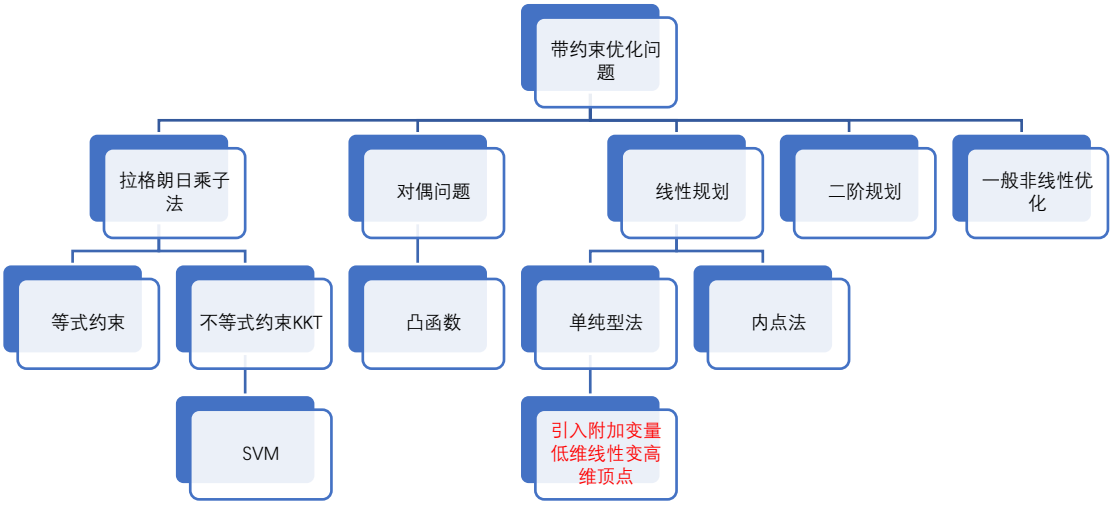
- 拉格朗日乘子法：
  等式约束：$\mathrm{minf(x)}\ \ s.t.\ g(x) = 0\ \ h(x) \geq 0$
  等价问题：$\mathrm{L(x, \lambda)} = \mathrm{f(x)} - \sum_i \lambda_i g_i(x) - \sum_j \mu_i h_i(x)$
  KKT:

原文摘录

- P500——离散优化中的主要问题是当 x 的维度过大时会产生组合爆炸，因此一般采用采样法求解。
- P501——然而，我们总可以在当前解的邻域内找到一个凸函数 m(x)来近似该点处的 f(x)，从而利用凸优化方法求解。如 SGD，Newton，Quasi-Newton，SQP 等。
- P505——对曲率较高的方向，更新步长会自动减小，从而避免在高曲率方向因步长过大导致的震荡；反之，如果在某一方向曲率比较小，则在该方向的步长会自动调大，从而避免在低曲率反向因步长过小导致的更新缓慢。

有感而发




带约束优化问题
拉格朗日乘子法
对偶问题
线性规划
二阶规划
一般非线性优化
等式约束
不等式约束KKT
凸函数
单纯型法
内点法
SVM
引入附加变量
低维线性变高维顶点

问题总结

1. 共轭梯度法