阿里云天池牛年读书会

机器学习的数学读书分享

分享嘉宾：SIGAI 雷明

SIGAI

天池读书会

TIANCHI天池

异步社区 www.epubit.com

《机器学习的数学》

系统介绍机器学习中涉及的数学知识的入门图书

直播嘉宾：SIGAI CEO 雷明

直播时间：2月23日20:00～21:00

扫码领取更多学习资料

目录

TIANCHI天池

作者简介

TIANCHI天池

- 清华大学出版社《机器学习-原理、算法与应用》，人民邮电出版社《机器学习的数学》作者；
- 2009年毕业于清华大学计算机系，研究方向为计算机视觉、机器学习，发表论文数篇；
- 曾就职于百度，任高级软件工程师/项目经理；zmodo/meshare，任CTO（创业），技术合伙人；
- 2018年创立SIGAI，致力于研发机器视觉框架，用标准化的算法为各个行业赋能，目前已经应用于物流等领域。已完成pre A轮融资

图书概况

TIANCHI天池

- 为什么数学这么重要
- 需要哪些数学知识
- 微积分
- 线性代数与矩阵论
- 概率论
- 信息论
- 最优化方法
- 随机过程
- 图论（含谱图理论）

学习资料、直播回放交流

TIANCHI天池

大家可以使用**电脑**访问下方地址进入天池读书会页面，获取读书会相关学习资料以及预约其他读书会直播。

https://tianchi.aliyun.com/specials/promotion/activity/bookclub

为什么数学这么重要？

- 机器学习算法的实现离不开数学
- 算法的理论分析需要数学
- 机器学习的理论需要数学
- 做好模型的调参需要数学

机器学习算法的实现离不开数学

TIANCHI天池

- 算法模型的构造如假设函数（对于有监督学习）、策略函数（对于强化学习）离不开数学
- 目标函数的构造需要数学，对于有监督学习、无监督学习、半监督学习、强化学习都是如此
- 最优化问题的求解/训练算法的设计需要数学，如梯度下降法，牛顿法，拟牛顿法

机器学习算法的实现离不开数学

TIANCHI天池

4.3.2 预测算法

在预测时需要寻找具有最大条件概率的那个类，即最大化后验概率（Maximum A Posteriori，MAP），根据贝叶斯公式有

$$\arg\max_c(p(c \mid \boldsymbol{x})) = \arg\max_c\left(\frac{p(c)p(\boldsymbol{x} \mid c)}{p(\boldsymbol{x})}\right)$$

假设每个类的概率 $p(c)$相等，$p(\boldsymbol{x})$对于所有类都是相等的，因此，等价于求解该问题：

$$\arg\max_c(p(\boldsymbol{x} \mid c))$$

也就是计算每个类的 $p(\boldsymbol{x}|c)$值，然后取最大的那个。对 $p(\boldsymbol{x}|c)$取对数，有

$$\ln(p(\boldsymbol{x} \mid c)) = \ln\left(\frac{1}{(2\pi)^{\frac{n}{2}} \mid \boldsymbol{\Sigma} \mid^{\frac{1}{2}}}\right) - \frac{1}{2}((\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}))$$

进一步简化为

$$\ln(p(\boldsymbol{x} \mid c)) = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln(\mid \boldsymbol{\Sigma} \mid) - \frac{1}{2}((\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}))$$

其中，$-\frac{n}{2}\ln(2\pi)$是常数，对所有类都是相同的。求上式的最大值等价于求下式的最小值：

$$\ln(\mid \boldsymbol{\Sigma} \mid) + ((\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}))$$

其中，$\ln(|\boldsymbol{\Sigma}|)$可以根据每一类的训练样本预先计算好，与 $\boldsymbol{x}$ 无关，不用重复计算。预测时只需要根据样本 $\boldsymbol{x}$ 计算$(\boldsymbol{x}-\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}$ 的值，而 $\boldsymbol{\Sigma}^{-1}$也是在训练时计算好的，不用重复计算。

正态贝叶斯分类器的预测函数

机器学习算法的实现离不开数学

TIANCHI天池

In information theory, mutual information between X and Y, $I(X;Y)$, measures the "amount of information" learned from knowledge of random variable Y about the other random variable X. The mutual information can be expressed as the difference of two entropy terms:

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \tag{2}$$

This definition has an intuitive interpretation: $I(X;Y)$ is the reduction of uncertainty in X when Y is observed. If X and Y are independent, then $I(X;Y) = 0$, because knowing one variable reveals nothing about the other; by contrast, if X and Y are related by a deterministic, invertible function, then maximal mutual information is attained. This interpretation makes it easy to formulate a cost: given any $x \sim P_G(x)$, we want $P_G(c|x)$ to have a small entropy. In other words, the information in the latent code c should not be lost in the generation process. Similar mutual information inspired objectives have been considered before in the context of clustering [23–25]. Therefore, we propose to solve the following information-regularized minimax game:

$$\min_G \max_D V_I(D,G) = V(D,G) - \lambda I(c; G(z,c)) \tag{3}$$

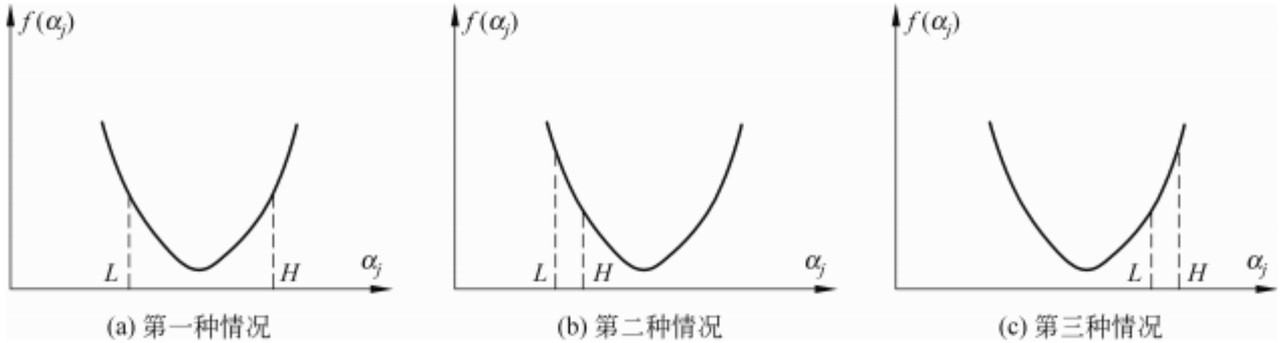
InfoGAN的目标函数

机器学习算法的实现离不开数学

利用这两个变量的等式约束条件，可以消掉 α_i，只剩下一个变量 α_j，目标函数简化为 α_j 的二次函数。可以直接求得这个二次函数的极值，假设不考虑约束条件得到的极值点为 $a_j^{\text{new,unclipped}}$，则最终的极值点为

$$\alpha_j^{\text{new}} = \begin{cases} H, & \alpha_j^{\text{new,unclipped}} > H \\ \alpha_j^{\text{new,unclipped}}, & L \leqslant \alpha_j^{\text{new,unclipped}} \leqslant H \\ L, & \alpha_j^{\text{new,unclipped}} < L \end{cases}$$

这三种情况如图 10.10 所示。



(a) 第一种情况　(b) 第二种情况　(c) 第三种情况

图 10.10　各种约束情况下的极小值

图 10.10(a)是抛物线的最小值点在$[L, H]$中；图 10.10(b)是抛物线的最小值点大于 H，被截断为 H；第三种情况是抛物线的最小值点小于 L，被截断为 L。

SVM的训练算法-SMO算法

算法的理论分析需要数学

TIANCHI天池

- 很多时候我们需要分析机器学习算法的特性和问题，并有针对性的给出一些解决方案。除实验之外，最有力的工具就是理论分析，这通常依赖于数学

- 万能逼近定理-证明多层神经网络能拟合闭区间上的任意连续函数

- 生成对抗网络（GAN）在训练过程中会出现令人讨厌的模式坍塌问题，为什么会出现这种问题？如何解决此问题？

- 监督学习模型如支持向量机、深度学习模型的泛化误差分析

算法的理论分析需要数学

tivation functions.

MATHEMATICAL APPENDIX

Because of the central role played by the Stone-Weierstrass theorem in obtaining our results, we state it here. Recall that a family **A** of real functions defined on a set E is an *algebra* if **A** is closed under addition, multiplication, and scalar multiplication. A family **A** *separates points* on E if for every x, y in E, $x \neq y$, there exists a function f in **A** such that $f(x) \neq f(y)$. The family A *vanishes at no point of* E if for each x in E there exists f in A such that $f(x) \neq 0$. (For further background, see Rudin, 1964, pp. 146–153.)

Stone-Weierstrass Theorem

Let **A** be an algebra of real continuous functions on a compact set K. If **A** separates points on K and if **A** vanishes at no point of K, then the uniform closure **B** of **A** consists of all real continuous functions on K (i.e., **A** is ρ_K-dense in the space of real continuous functions on K).

Proof of Theorem 2.1

Lemma A.2. Let F be a continuous squashing function and Ψ an arbitrary squashing function. For every $\varepsilon > 0$ there is an element H_ε of $\Sigma^1(\Psi)$ such that $\sup_{\lambda \in \mathbf{R}} |F(\lambda) - H_\varepsilon(\lambda)| < \varepsilon$.

Proof

Pick an arbitrary $\varepsilon > 0$. Without loss of generality, take $\varepsilon < 1$ also. We must find a finite collection of constants, β_j, and affine functions A_j, $j \in \{1,2, \ldots, Q-1\}$ such that $\sup_{\lambda \in \mathbf{R}} |F(\lambda) - \Sigma_{j=1}^{Q-1} \beta_j \Psi(A_j(\lambda))| < \varepsilon$.

Pick Q such that $1/Q < \varepsilon/2$. For $j \in \{1, \ldots, Q-1\}$ set $\beta_j = 1/Q$. Pick $M > 0$ such that $\Psi(-M) < \varepsilon/2Q$ and $\Psi(M) > 1 - \varepsilon/2Q$. Because Ψ is a sqashing function such an M can be found. For $j \in \{1, \ldots, Q-1\}$ set $r_j = \sup\{\lambda: F(\lambda) = j/Q\}$. Set $r_Q = \sup\{\lambda: F(\lambda) = 1 - 1/2Q\}$. Because F is a continuous squashing function such r_j's exist.

For any $r < s$ let $A_{r,s} \in A^1$ be the unique affine function satisfying $A_{r,s}(r) = M$ and $A_{r,s}(s) = -M$. The desired approximation is then $H_\varepsilon(\lambda) = \Sigma_{j=1}^{Q-1} \beta_j \Psi(A_{r_j, r_{j+1}}(\lambda))$. It is easy to check that on each of the intervals $(-\infty, r_1]$, $(r_1, r_2]$, $\ldots$, $(r_{Q-1}, r_Q]$, $(r_Q, +\infty)$ we have $|F(\lambda) - H_\varepsilon(\lambda)| < \varepsilon$. □

Proof of Theorem 2.3

万能逼近定理的证明

算法的理论分析需要数学

TIANCHI天池

2.1 SPECTRAL NORMALIZATION

Our spectral normalization controls the Lipschitz constant of the discriminator function f by literally constraining the spectral norm of each layer $g : \boldsymbol{h}_{in} \mapsto \boldsymbol{h}_{out}$. By definition, Lipschitz norm $\|g\|_{\text{Lip}}$ is equal to $\sup_{\boldsymbol{h}} \sigma(\nabla g(\boldsymbol{h}))$, where $\sigma(A)$ is the spectral norm of the matrix A (L_2 matrix norm of A)

$$\sigma(A) := \max_{\boldsymbol{h}:\boldsymbol{h}\neq \boldsymbol{0}} \frac{\|A\boldsymbol{h}\|_2}{\|\boldsymbol{h}\|_2} = \max_{\|\boldsymbol{h}\|_2\leq 1} \|A\boldsymbol{h}\|_2, \tag{6}$$

which is equivalent to the largest singular value of A. Therefore, for a linear layer $g(\boldsymbol{h}) = W\boldsymbol{h}$, the norm is given by $\|g\|_{\text{Lip}} = \sup_{\boldsymbol{h}} \sigma(\nabla g(\boldsymbol{h})) = \sup_{\boldsymbol{h}} \sigma(W) = \sigma(W)$. If the Lipschitz norm of the activation function $\|a_l\|_{\text{Lip}}$ is equal to 1 [1], we can use the inequality $\|g_1 \circ g_2\|_{\text{Lip}} \leq \|g_1\|_{\text{Lip}} \cdot \|g_2\|_{\text{Lip}}$ to observe the following bound on $\|f\|_{\text{Lip}}$:

$$\begin{aligned}\|f\|_{\text{Lip}} \leq & \|(\boldsymbol{h}_L \mapsto W^{L+1}\boldsymbol{h}_L)\|_{\text{Lip}} \cdot \|a_L\|_{\text{Lip}} \cdot \|(\boldsymbol{h}_{L-1} \mapsto W^L \boldsymbol{h}_{L-1})\|_{\text{Lip}} \\ & \cdots \|a_1\|_{\text{Lip}} \cdot \|(\boldsymbol{h}_0 \mapsto W^1\boldsymbol{h}_0)\|_{\text{Lip}} = \prod_{l=1}^{L+1} \|(\boldsymbol{h}_{l-1} \mapsto W^l \boldsymbol{h}_{l-1})\|_{\text{Lip}} = \prod_{l=1}^{L+1} \sigma(W^l).\end{aligned} \tag{7}$$

Our *spectral normalization* normalizes the spectral norm of the weight matrix W so that it satisfies the Lipschitz constraint $\sigma(W) = 1$:

$$\bar{W}_{\text{SN}}(W) := W/\sigma(W). \tag{8}$$

GAN的稳定性分析

算法的理论分析需要数学

Let $x \in \mathcal{X}$ be an input and $y \in \mathcal{Y}$ be a target. Let ℓ be a loss function. Let $R[f]$ be the expected risk of a function f, $R[f] = \mathbb{E}_{x,y\sim P}[\ell(f(x), y)]$, where P is the true distribution. Let $f_{\mathcal{A}(S_m)} : \mathcal{X} \to \mathcal{Y}$ be a model learned by a learning algorithm $\mathcal{A}$ (including random seeds for simplicity) using a training dataset $S_m = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ of size m. Let $\hat{R}_{S_m}[f]$ be the empirical risk of f, $\hat{R}_{S_m}[f] = \frac{1}{m}\sum_{i=1}^{m} \ell(f(x_i), y_i)$ with $\{(x_i, y_i)\}_{i=1}^{m} = S_m$. Let F be a set of functions or a *hypothesis space*. Let $\mathcal{L}_F$ be a family of loss functions associated with F, defined by $\mathcal{L}_F = \{g : f \in F, g(x, y) \triangleq \ell(f(x), y)\}$. All vectors are *column* vectors in this paper. For any given variable v, let d_v be the dimensionality of the variable v.

A goal in machine learning is typically framed as the minimization of the expected risk $R[f_{\mathcal{A}(S_m)}]$. We typically aim to minimize the non-computable expected risk $R[f_{\mathcal{A}(S_m)}]$ by minimizing the computable empirical risk $\hat{R}_{S_m}[f_{\mathcal{A}(S_m)}]$ (i.e., empirical risk minimization). One goal of the generalization theory is to explain and justify when and how minimizing $\hat{R}_{S_m}[f_{\mathcal{A}(S_m)}]$ is a sensible approach to minimizing $R[f_{\mathcal{A}(S_m)}]$ by analyzing

$$\text{the generalization gap} \triangleq R[f_{\mathcal{A}(S_m)}] - \hat{R}_{S_m}[f_{\mathcal{A}(S_m)}].$$

In this section only, we use the typical assumption that S_m is generated by i.i.d. draws according to the true distribution P; in general, this paper does not utilize this assumption. Under this assumption, a primary challenge of analyzing the generalization gap stems from the *dependence* of $f_{\mathcal{A}(S_m)}$ on the same dataset S_m used in the definition of $\hat{R}_{S_m}$. Several approaches in *statistical learning theory* have been developed to handle this dependence.

The *hypothesis-space complexity* approach handles this dependence by decoupling $f_{\mathcal{A}(S_m)}$ from the particular S_m by considering the worst-case gap for functions in the hypothesis space as

$$R[f_{\mathcal{A}(S_m)}] - \hat{R}_{S_m}[f_{\mathcal{A}(S_m)}] \leq \sup_{f\in F} R[f] - \hat{R}_{S_m}[f],$$

深度学习模型的泛化性能分析

机器学习的理论需要数学

机器学习理论对这个领域具有提纲挈领的作用。它可以回答下面一些基本问题

什么样的任务是可以学习的？

机器学习模型能够学习到何种程度，即在训练集，测试集上的误差上界是多少？

模型假设与泛化误差上界存在何种关系？

这些理论研究，更离不开数学这个工具

PAC可学习

任意充分小的正数

$$p\left(\mathrm{E}\left(h\right)\leq\varepsilon\right)\geq1-\delta$$

模型泛化误差

Third we should note that programs that compute concepts should be distinguished from those that compute merely the Boolean function. This distinction has little consequence for the three classes of expressions that we consider, since in these cases programs for the concepts follow directly from the specification of the expressions. For the sake of generality, our definitions do allow the value of a program to be undefined, since a nontotal vector will not, in general, determine the value of an expression as 0 or 1.

It would be interesting to obtain negative results, other than the informal cryptographic evidence mentioned in the introduction, indicating that the class of unrestricted Boolean circuits is not learnable. A recent result [6] in this direction does establish this, under the hypothesis that integer factorization is intractable. Their result applies to the learning protocol consisting of EXAMPLES and ORACLE, when the latter is restricted to total vectors. It may also be interesting to go in the opposite direction and determine, for example, whether the hypothesis that P equals NP would imply that all interesting classes are learnable.

4. A COMBINATORIAL BOUND

The probabilistic analysis needed for our later results can be abstracted into a single lemma. The proof of the

is essentially linear both in h and in S.

PROPOSITION: *For all integers $S \geq 1$ and all real $h > 1$.*

$$L(h,S) \leq 2h(S + \log_e h).$$

Proof: We use the following three well-known inequalities of which the last is due to Chernoff (see [6] page 18).

1. For all $x > 0$, $(1 + x^{-1})^x < e$.
2. For all $x > 0$, $(1 - x^{-1})^x < e^{-1}$.
3. In m independent trials, each with probability at least p of success, the probability that there are at most k success, where $k < mp$, is at most

$$\left(\frac{m - mp}{m - k}\right)^{m-k}\left(\frac{mp}{k}\right)^{k}.$$

The first factor in the expression in (3) above can be rewritten as

$$\left(1 - \frac{mp - k}{m - k}\right)^{((m-k)/(mp-k))(mp-k)}.$$

Using (2) with $x = (m - k)/(mp - k)$, we can upper bound the product by

$$e^{-mp+k}(mp/k)^k.$$

PAC理论

做好模型的调参需要数学

TIANCHI天池

梯度下降法中为什么需要学习率？

梯度下降法能保证在每次迭代时目标函数值一定下降吗？随机梯度下降法呢？

支持向量机中的惩罚因子C应该怎么设置？核函数该如何选择？

TIANCHI天池

需要哪些数学知识

现状分析

TIANCHI天池

数学是给机器学习、深度学习、强化学习的初学者和进阶者造成困难的主要原因之一，学生普遍对数学存在一种恐惧心理，数学自信的人只占少部分

国内本科数学教学方式、学生学习质量上存在的不足 - 过于抽象，偏重于计算，忽视了对数学思维、建模能力的培养 - 清华大学换用国外线性代数教材事件，如果结合一些具体的例子来讲解会好很多

某些数学知识超出了本科一般理工科专业的范畴 - 矩阵论/矩阵分析，信息论，最优化方法，随机过程，图论

通常情况下，高校、其他机构在教《机器学习》、《深度学习》、《强化学习》之前不会为学生把这些数学知识补齐

	算法	所用的数学知识
分类与回归	贝叶斯分类器	随机变量，贝叶斯公式，正态分布，最大似然估计
	决策树	熵，信息增益，Gini系数
	KNN算法	距离函数
	线性判别分析	散度矩阵，逆矩阵，瑞利商，拉格朗日乘数法，特征值与特征向量，标准正交基
	神经网络	矩阵运算，链式法则，交叉熵，欧氏距离，梯度下降法
	支持向量机	点到超平面的距离，拉格朗日对偶，强对偶，Slater条件，KKT条件，凸优化，核函数，Mercer条件，SMO算法
	logistic回归与softmax回归	条件概率，伯努利分布，多项分布，最大似然估计，凸优化，梯度下降法，牛顿法
	随机森林	抽样，方差
	Boosting算法	牛顿法，泰勒公式
	线性回归，岭回归，LASSO回归	均方误差，最小二乘法，向量范数，梯度下降法，凸优化
数据降维	主成分分析	均方误差，协方差矩阵，拉格朗日乘数法，特征值与特征向量，标准正交基
	核主成分分析	核函数
	流形学习	线性组合，均方误差，相似度图，拉普拉斯矩阵，特征值与特征向量，拉格朗日乘数法，KL散度，t分布，测地距离
距离度量学习	NCA	概率，梯度下降法
	ITML	KL散度，带约束的优化
	LMNN	线性变换，梯度下降法
	高斯混合模型与EM算法	正态分布，多项分布，边缘分布，条件分布，数学期望，Jensen不等式，凸函数，最大似然估计，拉格朗日乘数法
	高斯过程回归	正态分布，条件分布
概率图模型	HMM	马尔可夫过程，条件分布，边缘分布，最大似然估计，EM算法，拉格朗日乘数法
	CRF	图，条件概率，最大似然估计，拟牛顿法
	贝叶斯网络	图，条件概率，贝叶斯公式，最大似然估计
聚类	K均值算法	EM算法
	谱聚类	图，拉普拉斯矩阵，特征值与特征向量
	Mean Shift算法	核密度估计，梯度下降法
深度生成模型	GAN	概率分布变换，KL散度，JS散度，互信息，梯度下降法
	VAE	概率分布变换，KL散度，变分推断，梯度下降法
	变分推断	KL散度，变分法，贝叶斯公式
	MCMC采样	马尔可夫链，平稳分布，细致平衡条件，条件概率

究竟需要哪些数学知识？

TIANCHI天池

1.微积分 - 一元函数微积分，多元函数微积分，是整个高等数学的基石

2.线性代数与矩阵论 - 矩阵论本科一般不讲

3.概率论 - 内容基本已经覆盖机器学习的要求

4.信息论 - 一般专业不会讲，如果掌握了概率论，理解起来并不难

5.最优化方法 - 学了这门课的学生非常少，但对机器学习、深度学习非常重要，几乎所有算法归结为求解优化问题

6.随机过程 - 本科一般不学，但在机器学习中经常会使用，如马尔可夫过程，高斯过程，后者应用于贝叶斯优化

7.图论 - 计算机类专业本科通常会学，但没有讲谱图理论


信息论
随机过程
图论
最优化方法
概率论
多元函数微积分
一元函数微积分
线性代数与矩阵论

第1部分-微积分

为什么需要微积分？

TIANCHI天池

- 研究函数的性质 - 单调性，凹凸性
- 求解函数的极值 - 最优化方法的理论基础
- 概率论、信息论等课程的基础

一元函数微积分

TIANCHI天池

极限与连续　数列的极限，函数的极限，函数的连续性与间断点，上确界与下确界，Lipschitz连续性

导数　一阶导数，高阶导数，单调性判别法则，极值判别法则，凹凸性

微分中值定理 罗尔中值定理，拉格朗日中值定理，柯西中值定理

泰勒公式

不定积分 换元积分法，分部积分法

定积分 牛顿-莱布尼兹公式，换元积分法，分部积分法，变上限积分，广义积分

常微分方程

多元函数微积分

TIANCHI天池

- 偏导数 一阶偏导数，链式法则，高阶偏导数，全微分
- 梯度与方向导数
- 雅可比矩阵 链式法则的矩阵形式
- Hessian矩阵 多元函数的极值，凹凸性
- 向量与矩阵求导
- 微分算法 手动微分，符号微分，数值微分，自动微分
- 多元函数泰勒公式
- 多重积分 二重积分，三重积分，n重积分，多重积分的换元法

$$z=f(y)\qquad y=g(x)$$

$$(f(g(x)))'=f'(g(x))g'(x)$$

一元函数链式法则

推广到多元函数

$$h=h(x_1,x_2,...,x_n)$$

$$x_i=x_i(u_1,u_2,...,u_m)$$

$$\frac{\partial h}{\partial u_j}=\sum_{i=1}^{n}\frac{\partial h}{\partial x_i}\frac{\partial x_i}{\partial u_j}$$

多元函数链式法则

使用雅可比矩阵

$$z=f(y_1,...,y_m)$$

$$y_j=g_j(x_1,...,x_n),j=1,...,m$$

$$\nabla_{\mathbf{x}}z=\left(\frac{\partial \mathbf{y}}{\partial \mathbf{x}}\right)^{\mathrm{T}}\nabla_{\mathbf{y}}z$$

链式法则的矩阵形式

$$\nabla \mathbf{w}^{\mathrm{T}}\mathbf{x}=\mathbf{w}$$

$$\nabla \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}=\left(\mathbf{A}+\mathbf{A}^{\mathrm{T}}\right)\mathbf{x}$$

$$\nabla \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}=2\mathbf{A}\mathbf{x}$$

机器学习中的常用求导公式

$$g'(y)=\frac{1}{f'(g(y))}$$

一元反函数求导

推广到多元函数

$$\frac{\partial \mathbf{x}}{\partial \mathbf{y}}=\left(\frac{\partial \mathbf{y}}{\partial \mathbf{x}}\right)^{-1}\qquad \left|\frac{\partial \mathbf{x}}{\partial \mathbf{y}}\right|=\left|\frac{\partial \mathbf{y}}{\partial \mathbf{x}}\right|^{-1}$$

多元反函数求导

$$f(a+\Delta x)=\sum_{i=0}^{n}f^{(i)}(a)\frac{1}{i!}\Delta x^i+o(\Delta x^n)$$

$$f(a+\Delta x)=f(a)+\frac{f'(a)}{1!}\Delta x+\frac{1}{2}f''(a)\Delta x^2+o(\Delta x^2)$$

一元函数泰勒公式

单调性

费马定理，极值的必要条件。梯度下降法

极值的充分条件，判别法则。牛顿法，拟牛顿法

函数值的增长方向

推广到多元函数

$$f(a_1+\Delta x_1,...,a_m+\Delta x_m)=\sum_{p=0}^{n}\frac{1}{p!}\left(\Delta x_1\frac{\partial}{\partial x_1}+...+\Delta x_m\frac{\partial}{\partial x_m}\right)^p f(a_1,...,a_m)+o\left(\|\Delta \mathbf{x}\|^p\right)$$

$$f(\mathbf{x})=f(\mathbf{a})+(\nabla f(\mathbf{a}))^{\mathrm{T}}(\mathbf{x}-\mathbf{a})+\frac{1}{2}(\mathbf{x}-\mathbf{a})^{\mathrm{T}}\mathbf{H}(\mathbf{x}-\mathbf{a})+o\left(\|\mathbf{x}-\mathbf{a}\|^2\right)$$

多元函数泰勒公式

$$\int_a^b f(x)dx = F(b) - F(a)$$

牛顿-莱布尼兹公式

$$\int\ldots\int_D f(x_1,\ldots,x_n)dx_1\ldots dx_n = \int_{\varphi}^{\psi} dx_1 \int_{\varphi(x_1)}^{\psi(x_1)} dx_2 \ldots \int_{\varphi(x_1,\ldots,x_{n-1})}^{\psi(x_1,\ldots,x_{n-1})} f(x_1,\ldots x_n)dx_n$$

多重积分化为累次积分

$$\int_{\varphi(\alpha)}^{\varphi(\beta)} f(x)dx = \int_{\alpha}^{\beta} f(\varphi(t))\left|\varphi'(t)\right|dt$$

定积分换元法

推广到多元函数

$$\int\int\ldots\int_D f(\mathbf{x})d\mathbf{x} = \int\int\ldots\int_{D'} f(\varphi(\mathbf{y}))\left|\det\left(\frac{\partial \mathbf{x}}{\partial \mathbf{y}}\right)\right|d\mathbf{y}$$

定积分换元法

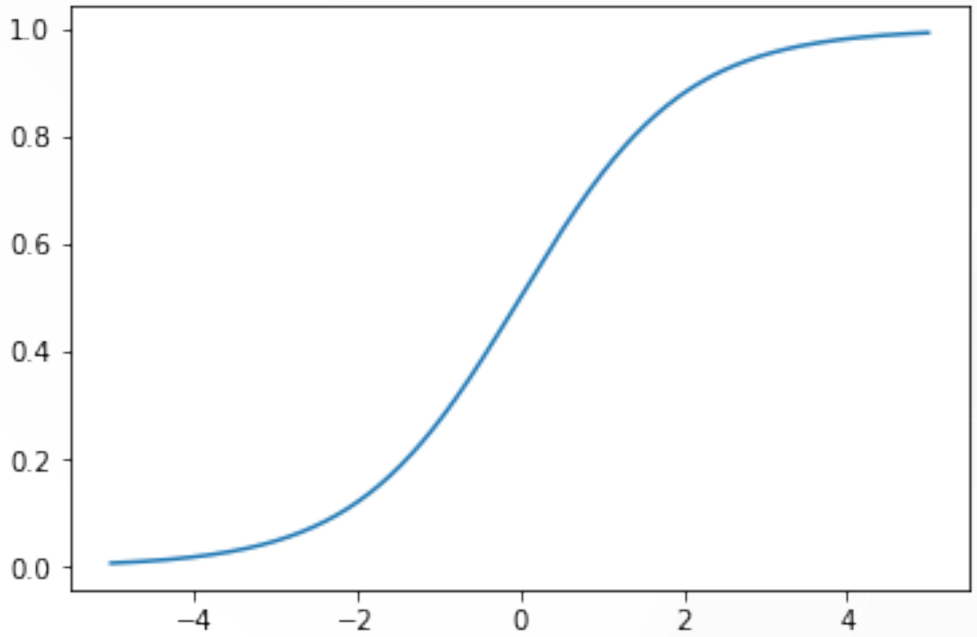
sigmoid激活函数 - 最基本的激活函数

$$f(x)=\frac{1}{1+e^{-x}}$$

它的导数为

$$f'(x)=-\frac{1}{\left(1+e^{-x}\right)^2}\left(1+e^{-x}\right)'=\frac{e^{-x}}{\left(1+e^{-x}\right)^2}$$

在整个定义域内单调递增



反向传播算法 - 链式法则的经典应用

如果使用欧氏距离损失函数，神经网络训练时的目标函数为

$$L(\mathbf{w})=\frac{1}{2}\left\|f\left(\mathbf{W}^{(n_l)}f\left(\mathbf{W}^{(n_l-1)}f\left(\ldots f\left(\mathbf{W}^{(1)}\mathbf{x}+\mathbf{b}^{(1)}\right)\ldots\right)+\mathbf{b}^{(n_l-1)}\right)+\mathbf{b}^{(n_l)}\right)-\mathbf{y}\right\|^2$$

可以用梯度下降法求解目标函数的极小值，为此需要计算每一层权重和偏置的梯度值

根据链式法则，可以推导出梯度计算公式。误差项可以递推计算

$$\boldsymbol{\delta}^{(l)}=\nabla_{\mathbf{u}^{(l)}}L=\begin{cases}\left(\mathbf{x}^{(l)}-\mathbf{y}\right)\odot f'\left(\mathbf{u}^{(l)}\right) & l=n_l\\ \left(\mathbf{W}^{(l+1)}\right)^{\mathrm{T}}\left(\boldsymbol{\delta}^{(l+1)}\right)\odot f'\left(\mathbf{u}^{(l)}\right) & l\neq n_l\end{cases}$$

根据误差项可以计算出权重与偏置的梯度

$$\nabla_{\mathbf{W}^{(l)}}L=\left(\nabla_{\mathbf{u}^{(l)}}L\right)\left(\mathbf{x}^{(l-1)}\right)^{\mathrm{T}}$$

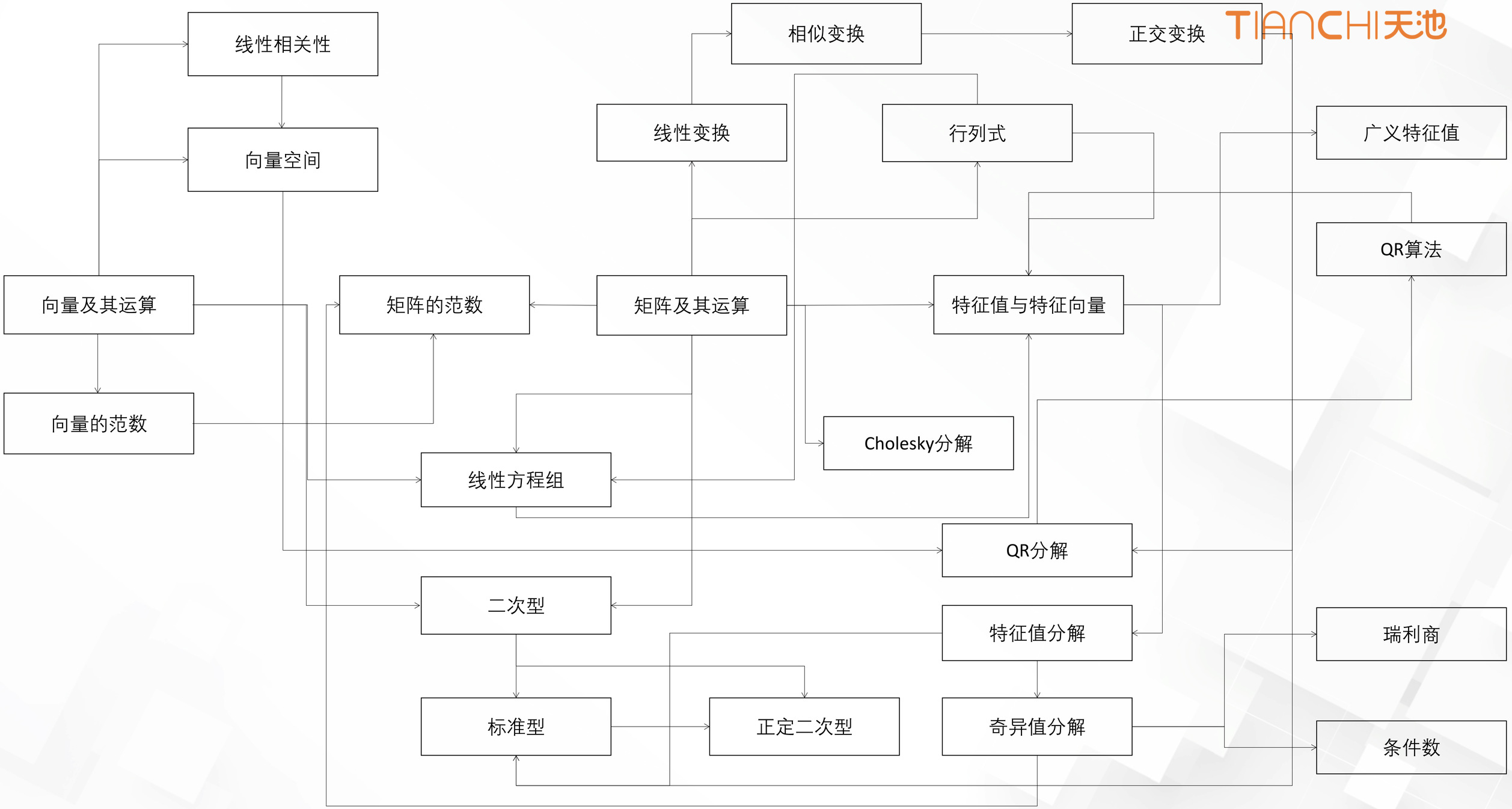
第2部分-线性代数与矩阵论

为什么需要线性代数？

TIANCHI天池

- 机器学习算法的输入、输出、中间结果，通常为向量，矩阵，张量
- 简化问题的表达
- 与微积分结合，研究多元函数的性质，也是概率论中随机向量的基础
- 在图论中亦有应用 - 图的拉普拉斯矩阵
- 在随机过程中同样有应用 - 状态转移矩阵

- 向量及其运算　基本运算，向量的范数，解析几何，线性相关性，线性空间
- 矩阵及其运算 基本运算，逆矩阵，矩阵的范数，线性变换
- 行列式的定义与计算
- 线性方程组　齐次线性方程组，非齐次线性方程组
- 特征值与特征值向量 定义与计算，相似变换，正交变换，QR算法，广义特征值，瑞利商，谱范数，条件数
- 二次型 正定二次型与正定矩阵，标准型
- 矩阵分解　　Cholesky分解，QR分解，特征值分解，奇异值分解


线性相关性
向量空间
向量及其运算
向量的范数
矩阵的范数
矩阵及其运算
相似变换
正交变换
线性变换
行列式
特征值与特征向量
广义特征值
QR算法
Cholesky分解
线性方程组
QR分解
二次型
特征值分解
瑞利商
标准型
正定二次型
奇异值分解
条件数

谱正则化

TIANCHI天池

正则化是机器学习中减轻过拟合的一种技术，它迫使模型的参数值很小，使模型变得更简单，一般情况下简单的模型有更好的泛化性能。正则化可以通过在目标函数中增加正则化项实现，正则化项通常为参数向量的L1或L2范数的平方。谱正则化（Spectral Regularization）用谱范数构造正则化项

$$\frac{1}{N}\sum_{i=1}^{N}L\left(\mathbf{x}_i,\mathbf{y}_i\right)+\frac{\lambda}{2}\sum_{i=1}^{l}\sigma\left(\mathbf{W}^{(i)}\right)^2$$

谱正则化项由神经网络所有层权重矩阵的谱范数平方之和构成，可以防止神经网络的权重矩阵出现大的谱范数，从而保证神经网络的映射有较小的李普希茨常数。谱范数是矩阵的最大奇异值

第3部分-概率论

为什么需要概率论？

TIANCHI天池

- 将机器学习算法的输入、输出看作随机变量/向量，用概率论的观点进行建模
- 对不确定性进行建模
- 挖掘变量之间的概率依赖关系
- 实现因果推理
- 随机算法 - 蒙特卡洛算法，遗传算法
- 数据生成问题 - 基本随机数生成，采样算法

随机事件与概率　条件概率，全概率公式，贝叶斯公式，条件独立

随机变量 离散型随机变量，连续型随机变量，数学期望与方差，标准差，Jesen不等式

常用概率分布 均匀分布，伯努利分布，二项分布，多项分布，几何分布，正态分布，t分布

概率分布变换 随机变量函数，逆变换采样

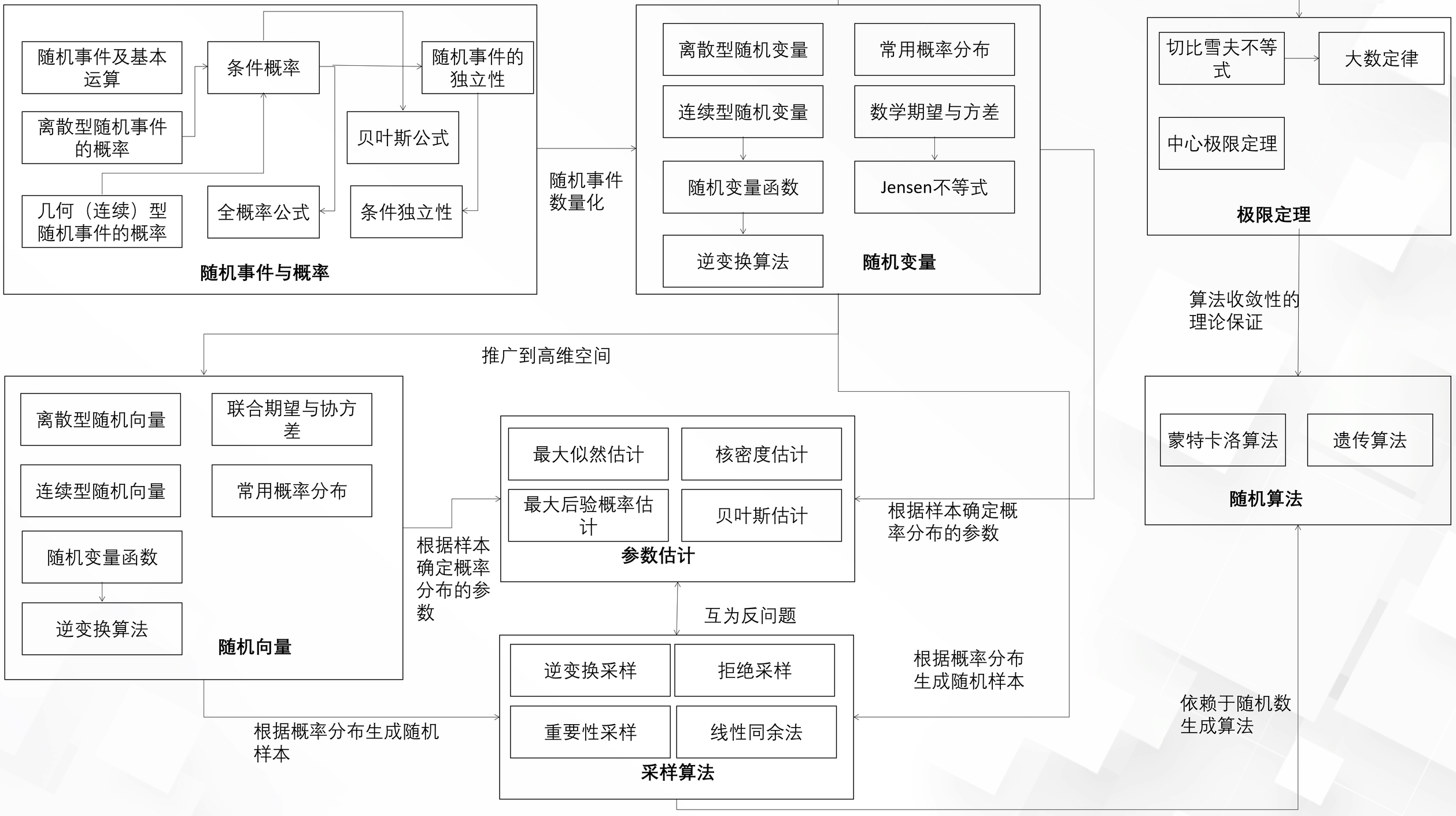
随机向量 离散型随机向量，连续型随机向量，联合期望，协方差，多维正态分布，概率分布变换

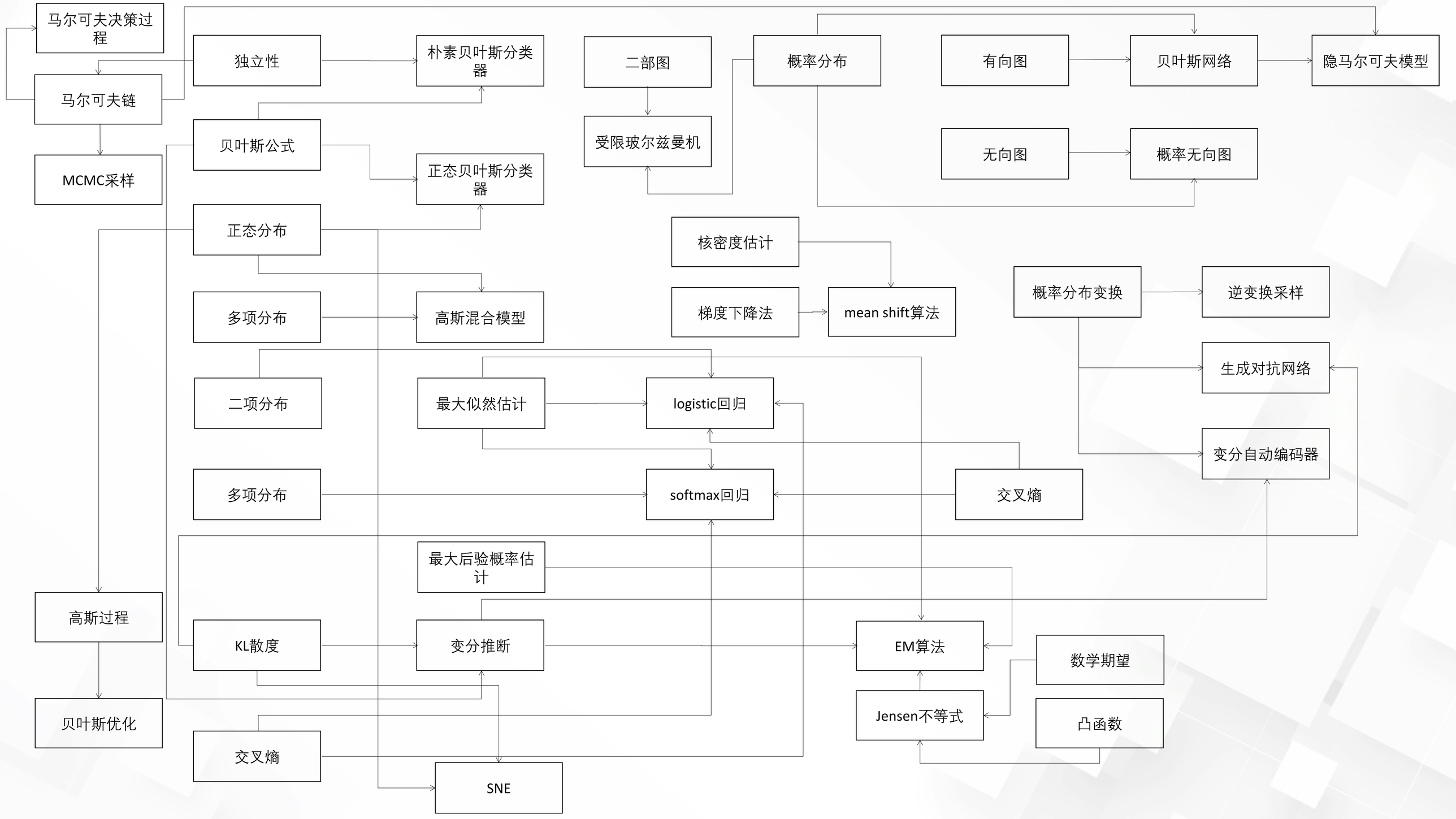
极限定理 切比雪夫不等式，大数定律，中心极限定理

参数估计 最大似然估计，最大后验概率估计，贝叶斯估计，核密度估计

随机算法 基本随机数生成，遗传算法，蒙特卡洛算法

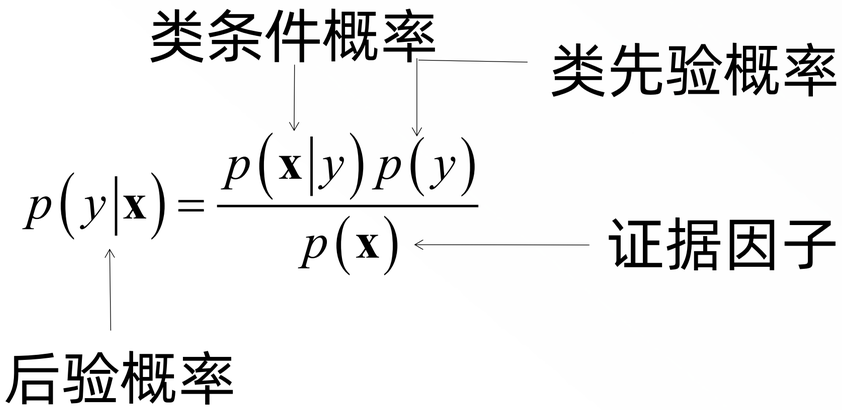
采样算法 拒绝采样，重要性采样


随机事件及基本运算
条件概率
随机事件的独立性
离散型随机事件的概率
贝叶斯公式
几何（连续）型随机事件的概率
全概率公式
条件独立性
随机事件与概率
随机事件数量化
离散型随机变量
常用概率分布
连续型随机变量
数学期望与方差
随机变量函数
Jensen不等式
逆变换算法
随机变量
切比雪夫不等式
大数定律
中心极限定理
极限定理
算法收敛性的理论保证
推广到高维空间
离散型随机向量
联合期望与协方差
连续型随机向量
常用概率分布
随机变量函数
逆变换算法
随机向量
根据样本确定概率分布的参数
最大似然估计
核密度估计
最大后验概率估计
贝叶斯估计
参数估计
根据样本确定概率分布的参数
互为反问题
逆变换采样
拒绝采样
重要性采样
线性同余法
采样算法
根据概率分布生成随机样本
根据概率分布生成随机样本
蒙特卡洛算法
遗传算法
随机算法
依赖于随机数生成算法


马尔可夫决策过程
独立性
朴素贝叶斯分类器
二部图
概率分布
有向图
贝叶斯网络
隐马尔可夫模型
马尔可夫链
受限玻尔兹曼机
无向图
概率无向图
贝叶斯公式
正态贝叶斯分类器
MCMC采样
正态分布
核密度估计
概率分布变换
逆变换采样
多项分布
高斯混合模型
梯度下降法
mean shift算法
生成对抗网络
二项分布
最大似然估计
logistic回归
变分自动编码器
多项分布
softmax回归
交叉熵
最大后验概率估计
高斯过程
KL散度
变分推断
EM算法
数学期望
贝叶斯优化
Jensen不等式
交叉熵
凸函数
SNE

贝叶斯分类器-利用贝叶斯公式解决分类问题

利用贝叶斯公式，计算样本属于每个类的概率



将样本判定为后验概率最大的那个类

$$\arg\max_y p(\mathbf{x}|y)p(y)$$

机器翻译问题 - 找到条件概率最大的那个目标句子

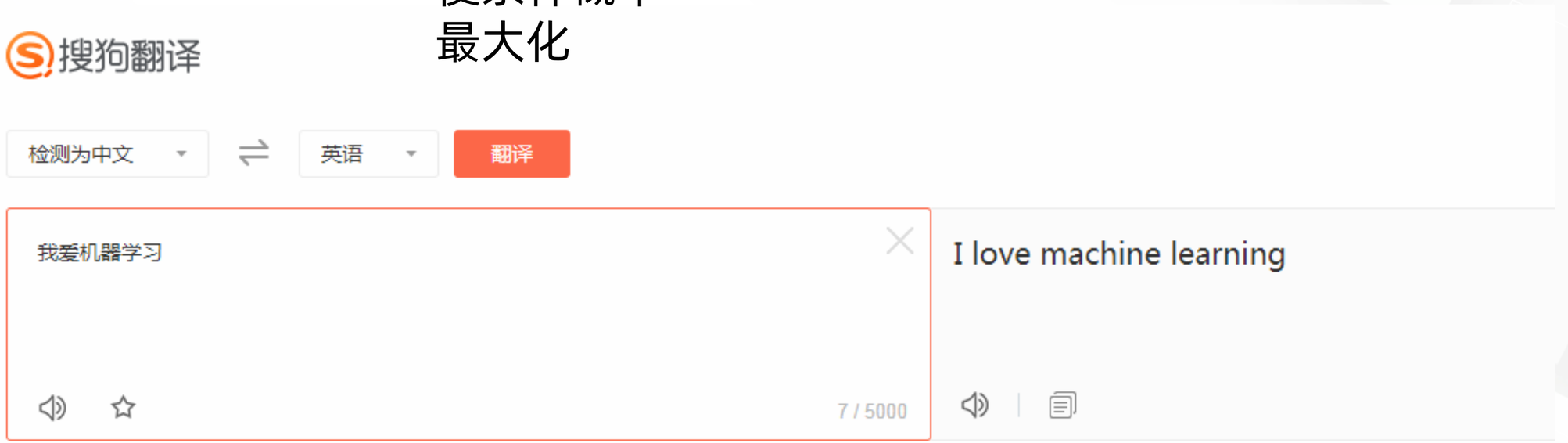
机器翻译的目标是给定一种语言的句子，找出另一种语言对应的句子的最优方案，二者有相同的语义

源语言的句子

$$\arg\max_{s_{english}} p\left(s_{english} \middle| \text{"我爱机器学习"}\right)$$

目标语言的句子

使条件概率最大化

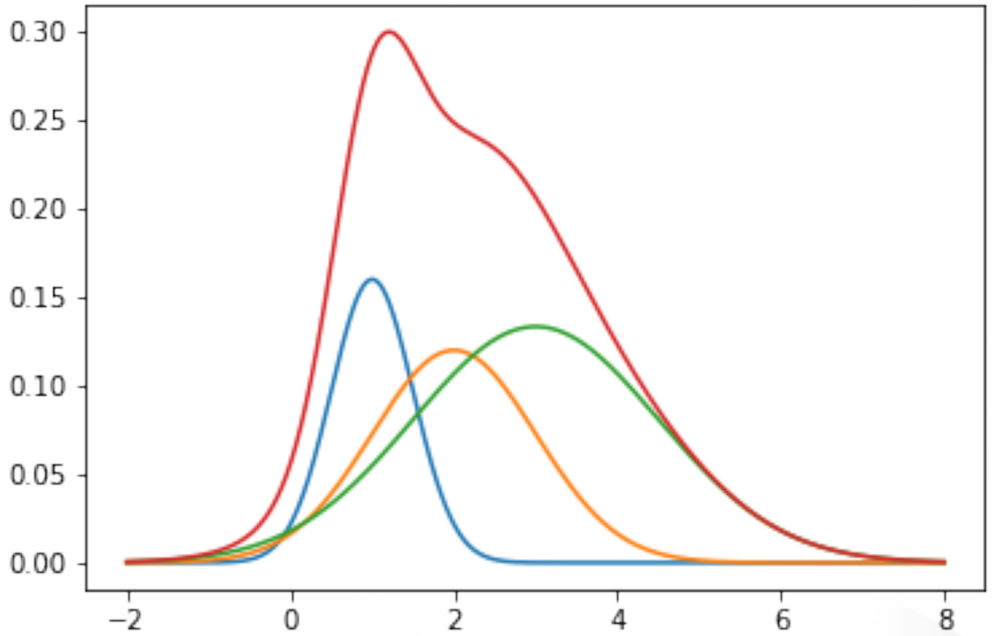


高斯混合模型

用一组高斯分布的加权和来表示概率密度函数

$$p(\mathbf{x}) = \sum_{i=1}^{k} w_i N(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$$

是多项分布与正态分布的结合



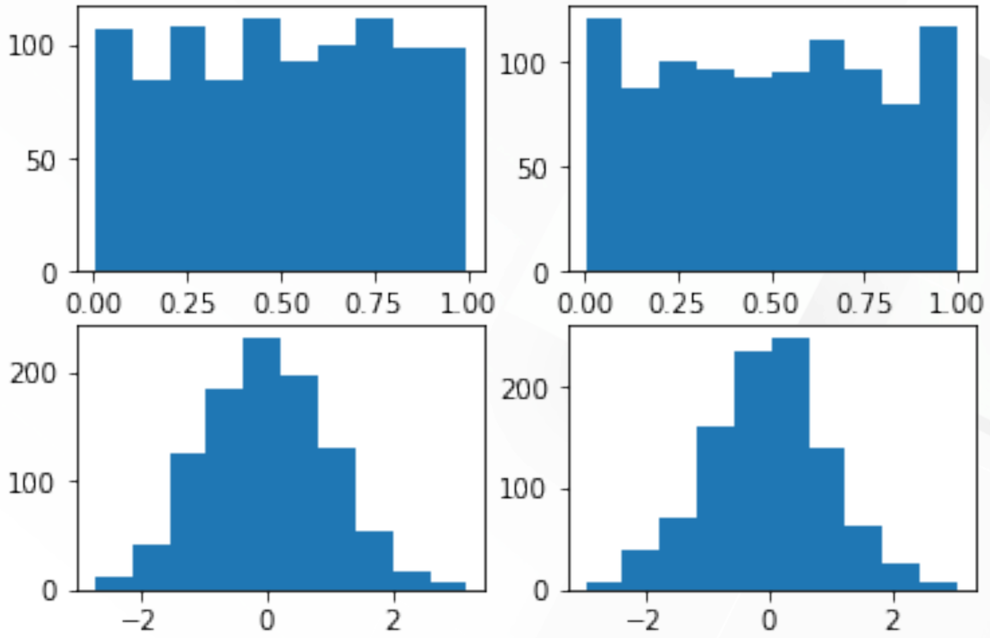
均匀分布的随机数生成 - 线性同余法

$$x_{i+1} = (a \cdot x_i + b) \bmod m$$

$$a = 7^5 = 16807$$

$$b = 0$$

$$m = 2^{31} - 1 = 2147483647$$



正态分布的随机数生成 - Box-Muller算法

$$z_1 = \sqrt{-2 \ln u_1} \cos(2\pi u_2)$$

$$z_2 = \sqrt{-2 \ln u_1} \sin(2\pi u_2)$$

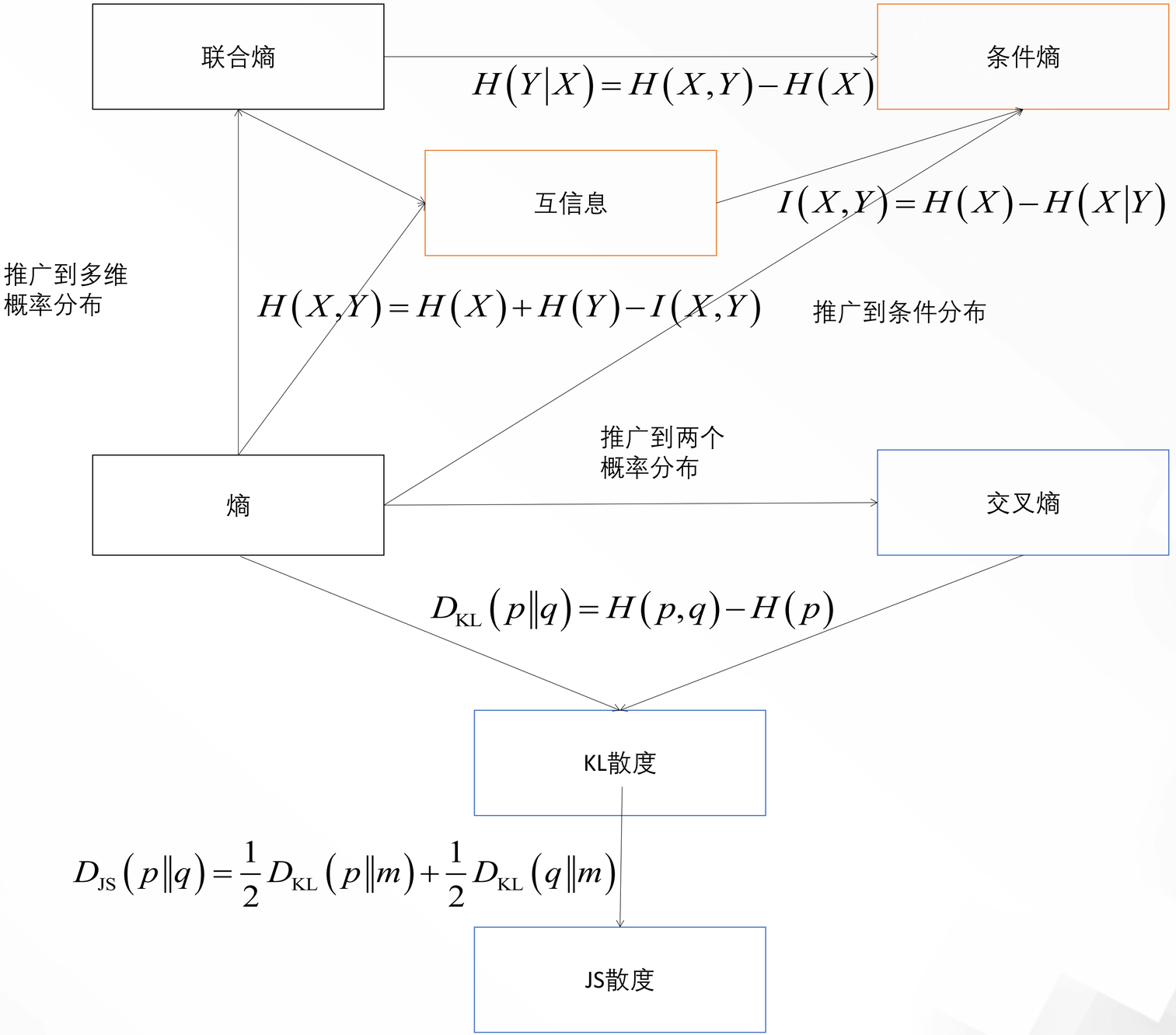
第4部分-信息论

为什么需要信息论？

TIANCHI天池

- 构造机器学习模型的目标函数
- 对机器学习模型进行理论分析
- 挖掘变量之间的概率关系，筛选变量

- 熵与联合熵　熵，联合熵，熵的性质，某些特殊概率分布的熵，在决策树训练中的应用
- 交叉熵 定义与性质，logistic回归，softmax回归
- KL散度 定义与性质，与交叉熵的关系，变分推断，SNE降维，t-SNE降维
- JS散度 定义与性质，在生成对抗网络中的应用
- 互信息 定义与性质，与熵的关系
- 条件熵 定义与性质，与熵以及互信息的关系


联合熵
条件熵
$H(Y|X)=H(X,Y)-H(X)$
互信息
$I(X,Y)=H(X)-H(X|Y)$
推广到多维
概率分布
$H(X,Y)=H(X)+H(Y)-I(X,Y)$
推广到条件分布
推广到两个
概率分布
熵
交叉熵
$D_{KL}(p\|q)=H(p,q)-H(p)$
KL散度
$D_{JS}(p\|q)=\frac{1}{2}D_{KL}(p\|m)+\frac{1}{2}D_{KL}(q\|m)$
JS散度

softmax回归的训练问题

最大似然估计，对数似然函数为

$$\sum_{i=1}^{l}\sum_{j=1}^{k}\left(y_{ij}\ln\frac{\exp\left(\boldsymbol{\theta}_j^{\mathrm{T}}\mathbf{x}_i\right)}{\sum_{t=1}^{k}\exp\left(\boldsymbol{\theta}_t^{\mathrm{T}}\mathbf{x}_i\right)}\right)$$

$$\mathbf{y}^{*}==\frac{1}{\sum_{i=1}^{k}e^{\boldsymbol{\theta}_i^{\mathrm{T}}\mathbf{x}}}\begin{bmatrix}e^{\boldsymbol{\theta}_1^{\mathrm{T}}\mathbf{x}}\\ \cdots\\ e^{\boldsymbol{\theta}_k^{\mathrm{T}}\mathbf{x}}\end{bmatrix}$$

让对数似然函数取极大值等价于让下面的损失函数取极小值

$$-\sum_{i=1}^{l}\sum_{j=1}^{k}\left(y_{ij}\ln\frac{\exp\left(\boldsymbol{\theta}_j^{\mathrm{T}}\mathbf{x}_i\right)}{\sum_{t=1}^{k}\exp\left(\boldsymbol{\theta}_t^{\mathrm{T}}\mathbf{x}_i\right)}\right)$$

交叉熵反映了预测出的概率分布与真实标签值概率分布得差异，二者均为多项分布

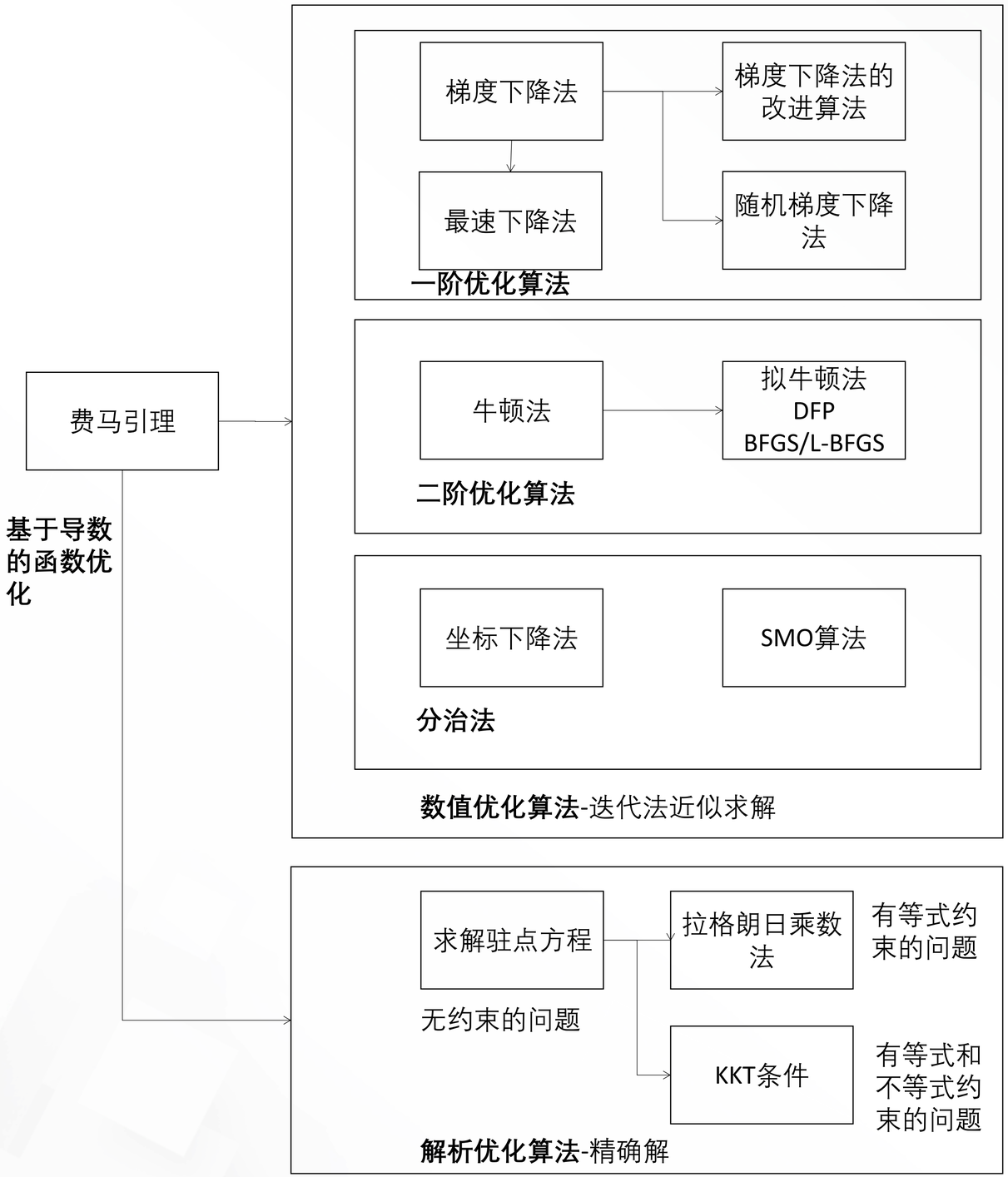
第5部分-最优化方法

为什么需要最优化方法？

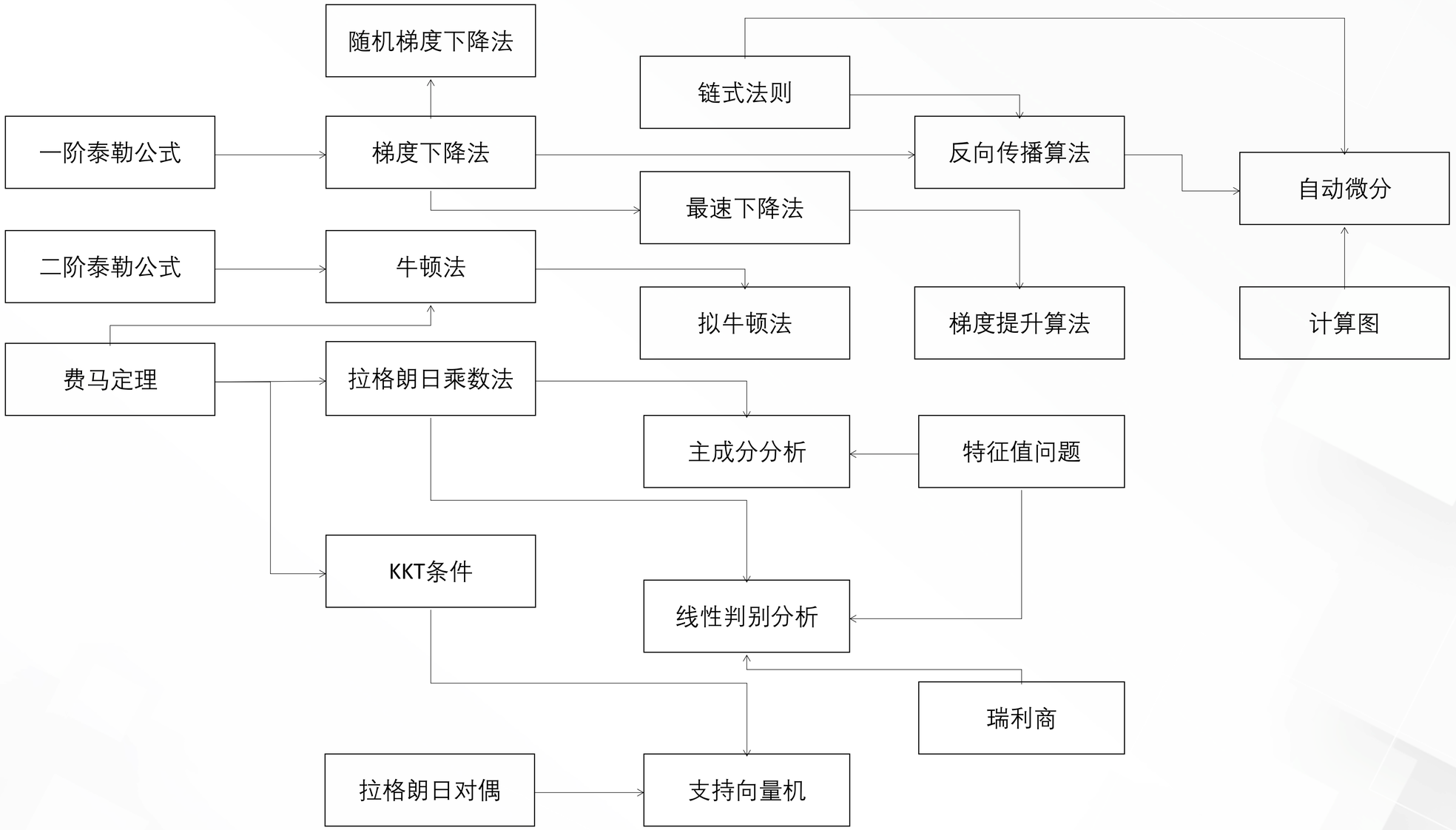
TIANCHI天池

- 确定机器学习模型的参数/函数
- 确定机器学习模型的预测值

- 基本概念 问题定义，迭代法的基本思想
- 一阶优化算法 梯度下降法，最速下降法，梯度下降法的各种改进，随机梯度下降法
- 二阶优化算法 牛顿法，拟牛顿法
- 分治法 坐标下降法，分阶段优化
- 凸优化
- 带约束的优化问题 拉格朗日乘数法，拉格朗日对偶，KKT条件
- 多目标优化
- 泛函与变分法
- 目标函数的构造 有监督学习，无监督学习，强化学习


梯度下降法
梯度下降法的改进算法
最速下降法
随机梯度下降法
一阶优化算法
费马引理
牛顿法
拟牛顿法
DFP
BFGS/L-BFGS
二阶优化算法
基于导数的函数优化
坐标下降法
SMO算法
分治法
数值优化算法-迭代法近似求解
求解驻点方程
拉格朗日乘数法
有等式约束的问题
无约束的问题
KKT条件
有等式和不等式约束的问题
解析优化算法-精确解


启发式搜索
遗传算法
模拟退火
蚁群优化
基于随机数的函数优化
ε
线性标量化
约束法
多目标函数优化
欧拉-拉格朗日方程
泛函优化


随机梯度下降法
链式法则
一阶泰勒公式
梯度下降法
反向传播算法
自动微分
最速下降法
二阶泰勒公式
牛顿法
拟牛顿法
梯度提升算法
计算图
费马定理
拉格朗日乘数法
主成分分析
特征值问题
KKT条件
线性判别分析
瑞利商
拉格朗日对偶
支持向量机

支持向量机的训练算法

在所有训练样本被正确分类的前提下，最大化分类间隔

$$\min \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$$

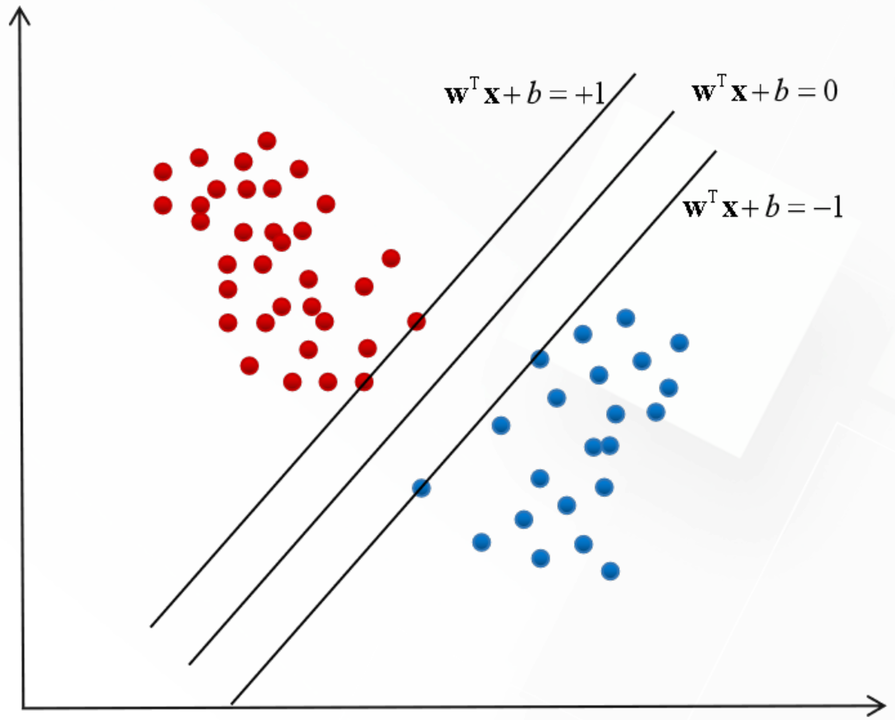
$$y_i\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i+b\right)\geq 1$$

使用拉格朗日对偶，转化为对偶问题求解

$$\min_{\boldsymbol{\alpha}} \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j-\sum_{i=1}^{l}\alpha_i$$

$$\alpha_i\geq 0, i=1,\ldots,l$$

$$\sum_{i=1}^{l}\alpha_i y_i=0$$



logistic回归的训练问题

最大似然估计，求解下面的优化问题

$$\ln L(\mathbf{w}) = \sum_{i=1}^{l} \left(y_i \ln h(\mathbf{x}_i) + (1 - y_i) \ln \left(1 - h(\mathbf{x}_i) \right) \right)$$

$$h(\mathbf{x}) = \frac{1}{1 + \exp\left(-\mathbf{w}^{\mathrm{T}} \mathbf{x} \right)}$$

等价于求解下面的极小值问题

$$f(\mathbf{w}) = -\sum_{i=1}^{l} \left(y_i \ln h(\mathbf{x}_i) + (1 - y_i) \ln \left(1 - h(\mathbf{x}_i) \right) \right)$$

梯度计算公式为

$$\nabla_{\mathbf{w}} f(\mathbf{w}) = \sum_{i=1}^{l} \left(h(\mathbf{x}_i) - y_i \right) \mathbf{x}_i$$

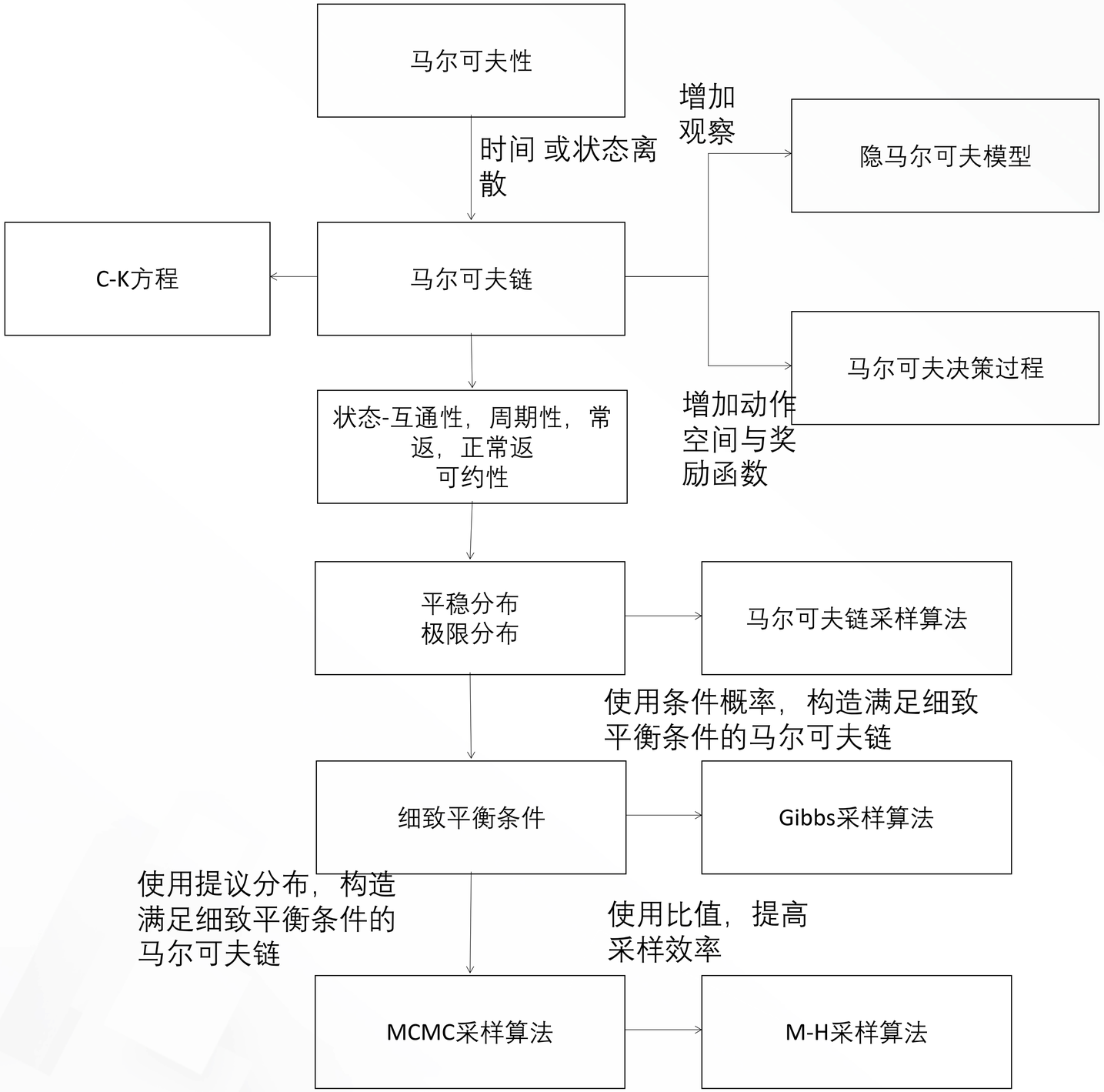
第6部分-随机过程

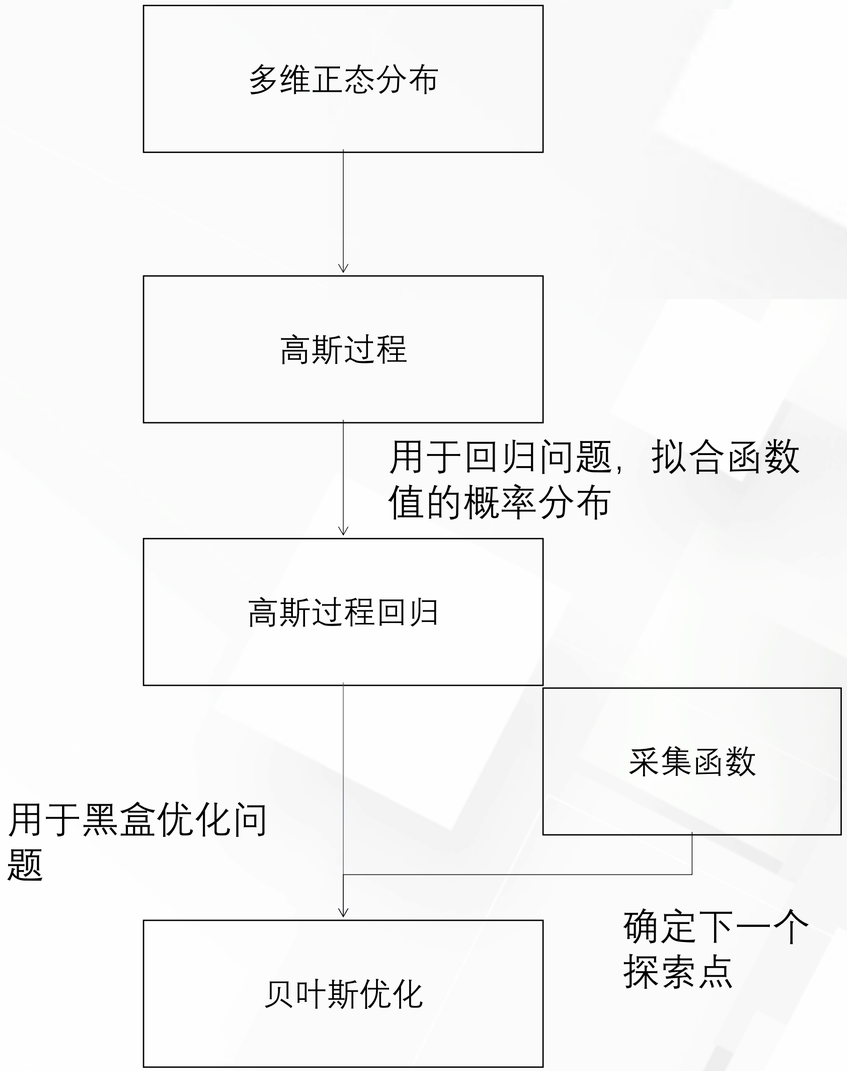
为什么需要随机过程？

TIANCHI天池

- 对随机变量序列进行建模
- 对不确定性进行建模

- 马尔可夫过程　马尔可夫性，马尔可夫链，平稳分布与极限分布，细致平稳条件
- 马尔可夫链采样算法　Metropolis-Hastings算法，Gibbs采样
- 高斯过程　高斯过程，高斯过程回归，贝叶斯优化

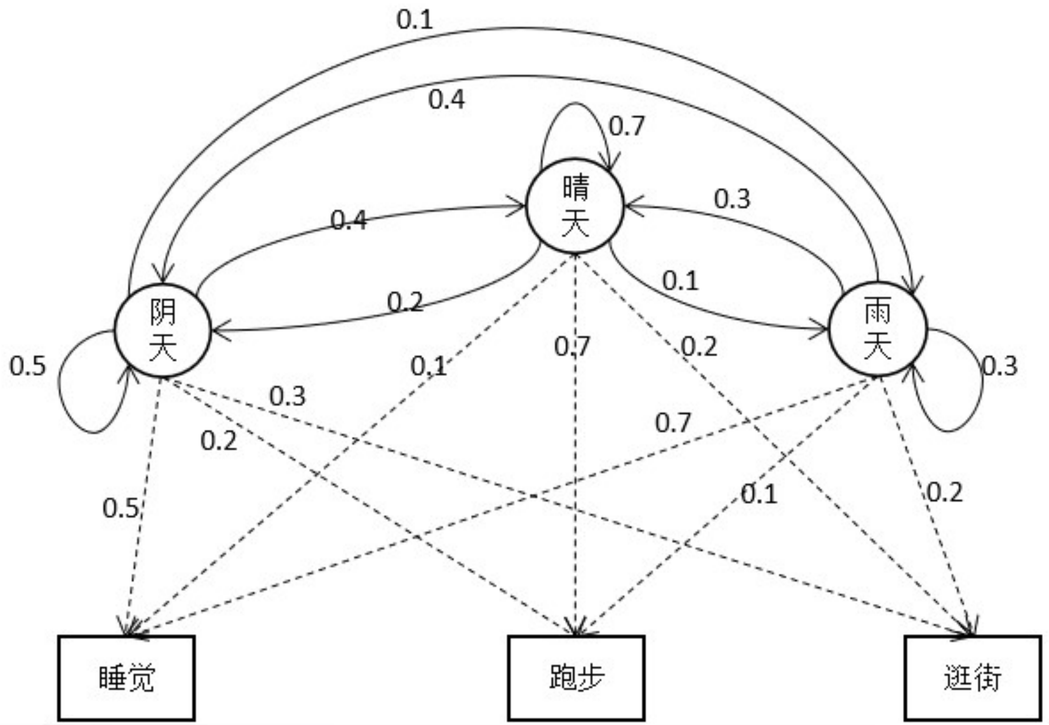

马尔可夫性
时间 或状态离散
增加观察
隐马尔可夫模型
C-K方程
马尔可夫链
马尔可夫决策过程
状态-互通性，周期性，常返，正常返
可约性
增加动作空间与奖励函数
平稳分布
极限分布
马尔可夫链采样算法
使用条件概率，构造满足细致平衡条件的马尔可夫链
细致平衡条件
Gibbs采样算法
使用提议分布，构造满足细致平衡条件的马尔可夫链
使用比值，提高采样效率
MCMC采样算法
M-H采样算法


多维正态分布
高斯过程
用于回归问题，拟合函数值的概率分布
高斯过程回归
采集函数
用于黑盒优化问题
确定下一个探索点
贝叶斯优化

隐马尔可夫模型

系统由状态值和观测值构成，状态值形成马尔可夫链，状态值不可直接得到；观测值可直接得到

根据观测值推理出最有可能的状态值



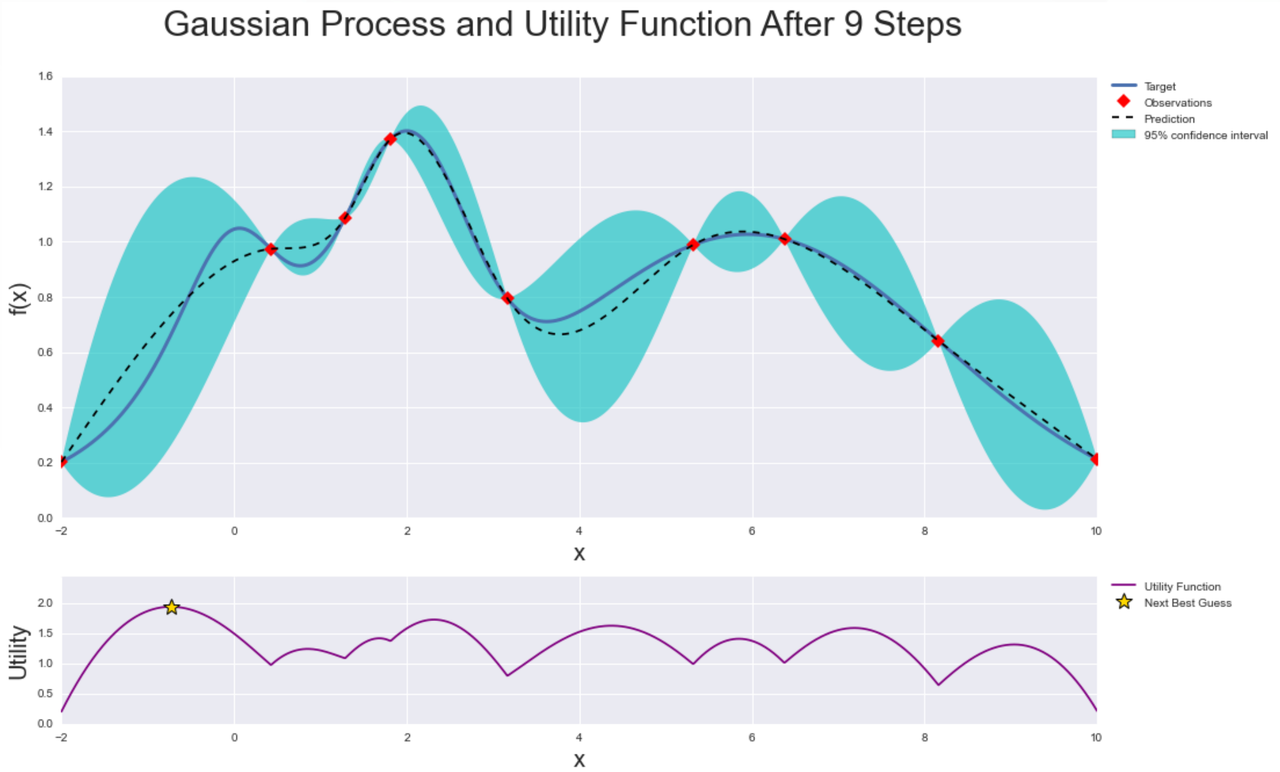
$$\mathbf{B}=\begin{bmatrix}0.5 & 0.2 & 0.3\\ 0.1 & 0.7 & 0.2\\ 0.7 & 0.1 & 0.2\end{bmatrix}$$

$$\mathbf{A}=\begin{bmatrix}0.7 & 0.2 & 0.1\\ 0.4 & 0.5 & 0.1\\ 0.3 & 0.4 & 0.3\end{bmatrix}$$

贝叶斯优化 - 解决黑盒优化问题的有力工具

根据一组样本点预测函数在每个点处的函数值的概率分布，由高斯过程回归实现，是一个一维正态分布

根据正态分布的数学期望和方差构造采集函数，代表每个点是极值点的可能性，求解该函数的极值，作为下一个探索点



TIANCHI天池

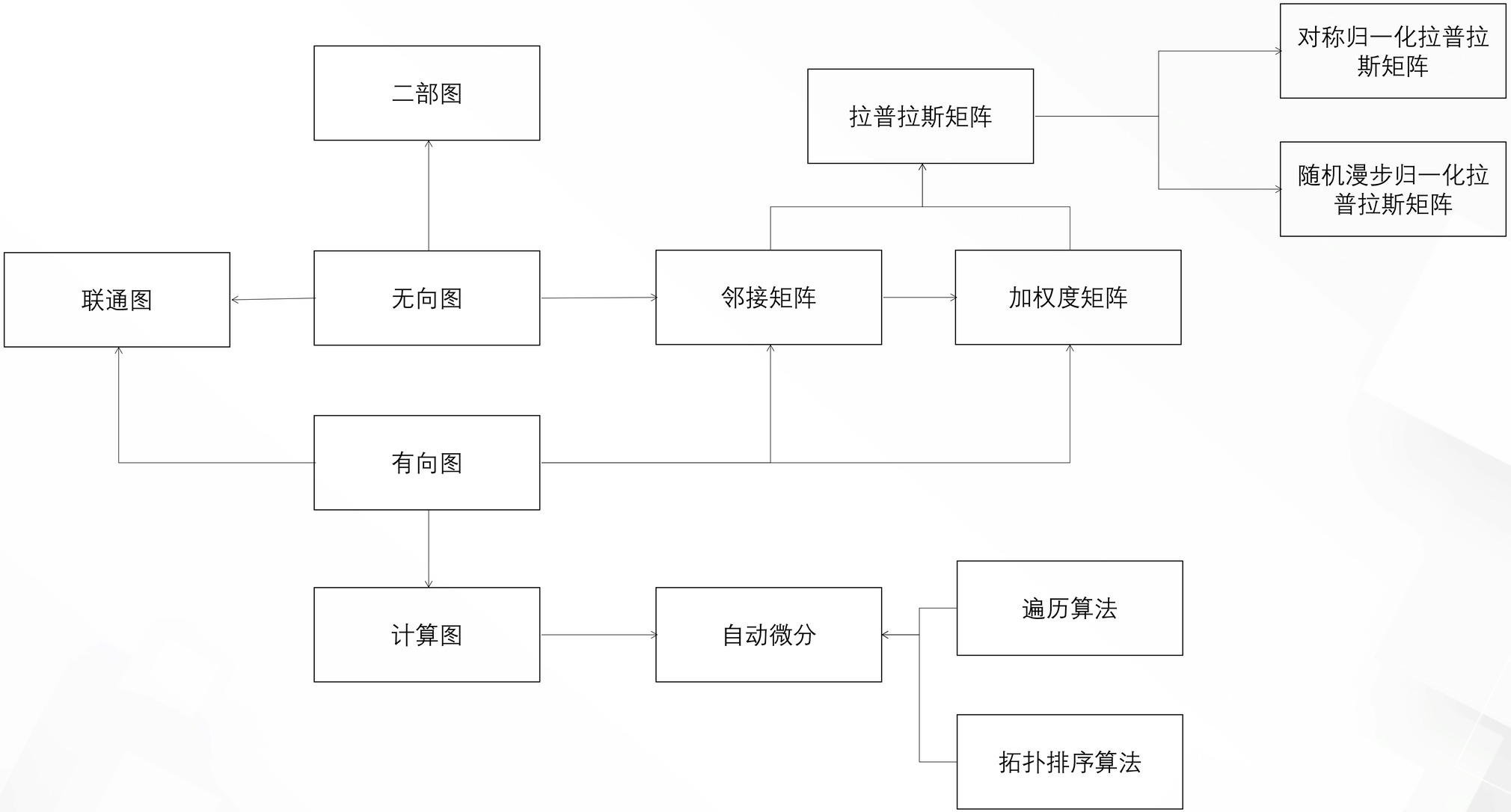
第7部分-图论

为什么需要图论？

TIANCHI天池

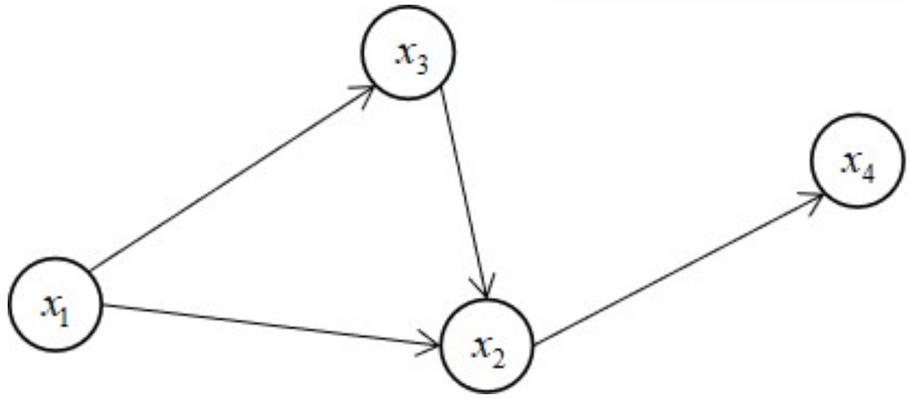
- 概率图模型
- 计算图
- 用谱图理论解决机器学习问题　流形降维，谱聚类
- 图神经网络

- 基本概念　基本定义，图的矩阵表示
- 特殊的图　联通图，二部图，有向无环图
- 图的算法　遍历算法，Dijkstra算法，拓扑排序
- 谱图理论　拉普拉斯矩阵，归一化拉普拉斯矩阵


二部图
拉普拉斯矩阵
对称归一化拉普拉斯矩阵
随机漫步归一化拉普拉斯矩阵
联通图
无向图
邻接矩阵
加权度矩阵
有向图
计算图
自动微分
遍历算法
拓扑排序算法

概率图模型

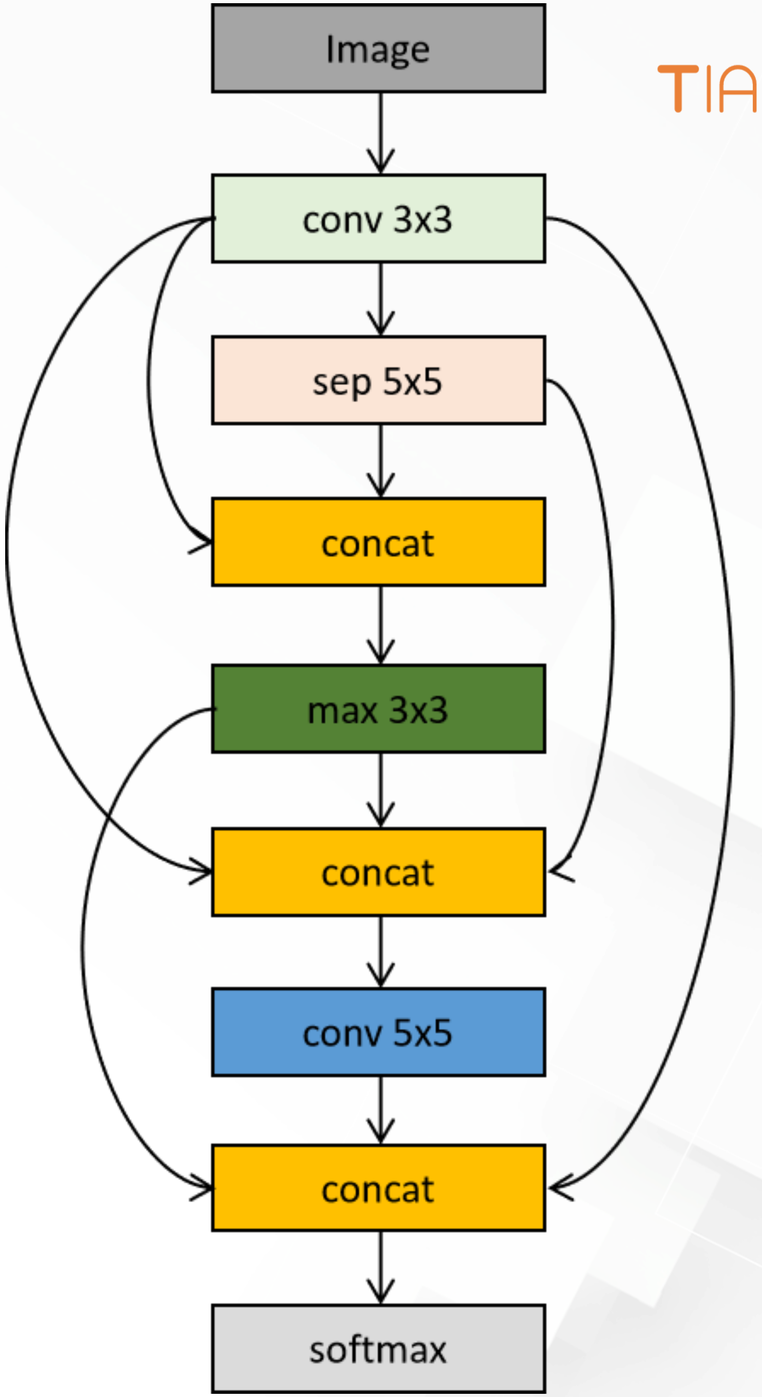
概率图模型是概率论与图论相结合的产物。它用图表示随机变量之间的概率关系，对联合概率或条件概率建模。在这种图中，顶点是随机变量，边为变量之间的依赖关系



$$p(x_1, x_2, x_3, x_4) = p(x_1)\,p(x_2 | x_1, x_3)\,p(x_3 | x_1)\,p(x_4 | x_2)$$

神经网络的拓扑结构图

神经网络由多个层组成，各个层之间存在连接关系，因此可以将网络的拓扑结构表示为一个图



基于图的降维算法-拉普拉斯特征映射

为样本集构造相似度图，然后对图的拉普拉斯矩阵进行特征值分解，得到投影结果

原本要优化的目标为

$$\min_{\mathrm{y}} \sum_{i=1}^{n}\sum_{j=1}^{n}\left(y_i - y_j\right)^2 w_{ij}$$

用拉普拉斯矩阵表示为

$$\min_{\mathrm{y}} \mathbf{y}^{\mathrm{T}}\mathbf{L}\mathbf{y}$$

$$\mathbf{y}^{\mathrm{T}}\mathbf{D}\mathbf{y} = 1$$

最后归结于矩阵的特征值问题

$$\mathbf{D}^{-1}\mathbf{L}y = \lambda\mathbf{y}$$

TIANCHI天池

Q&A

天池读书会

TIANCHI天池

异步社区 www.epubit.com

《机器学习的数学》

系统介绍机器学习中涉及的数学知识的入门图书

直播嘉宾：SIGAI CEO 雷明

直播时间：2月23日20:00～21:00

扫码领取更多学习资料