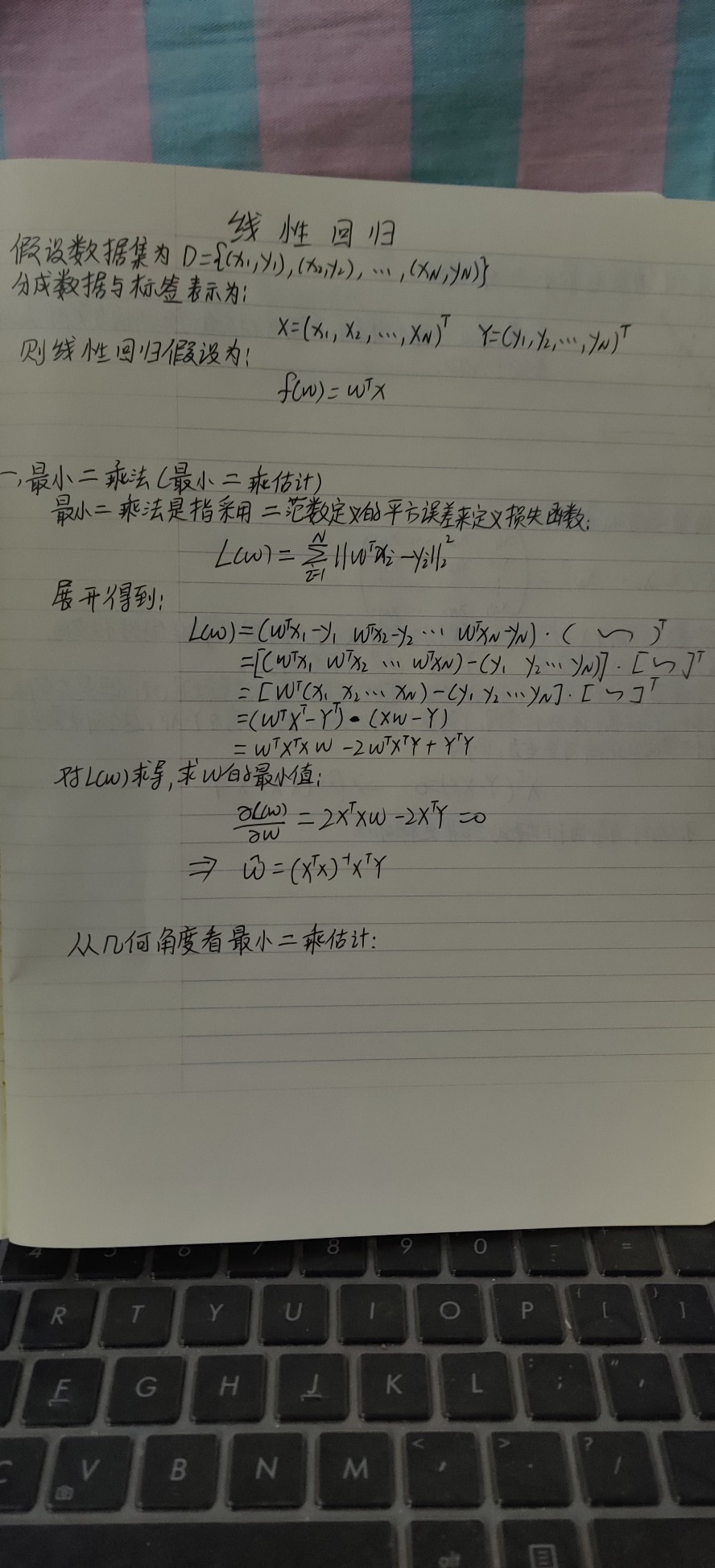
白板推导

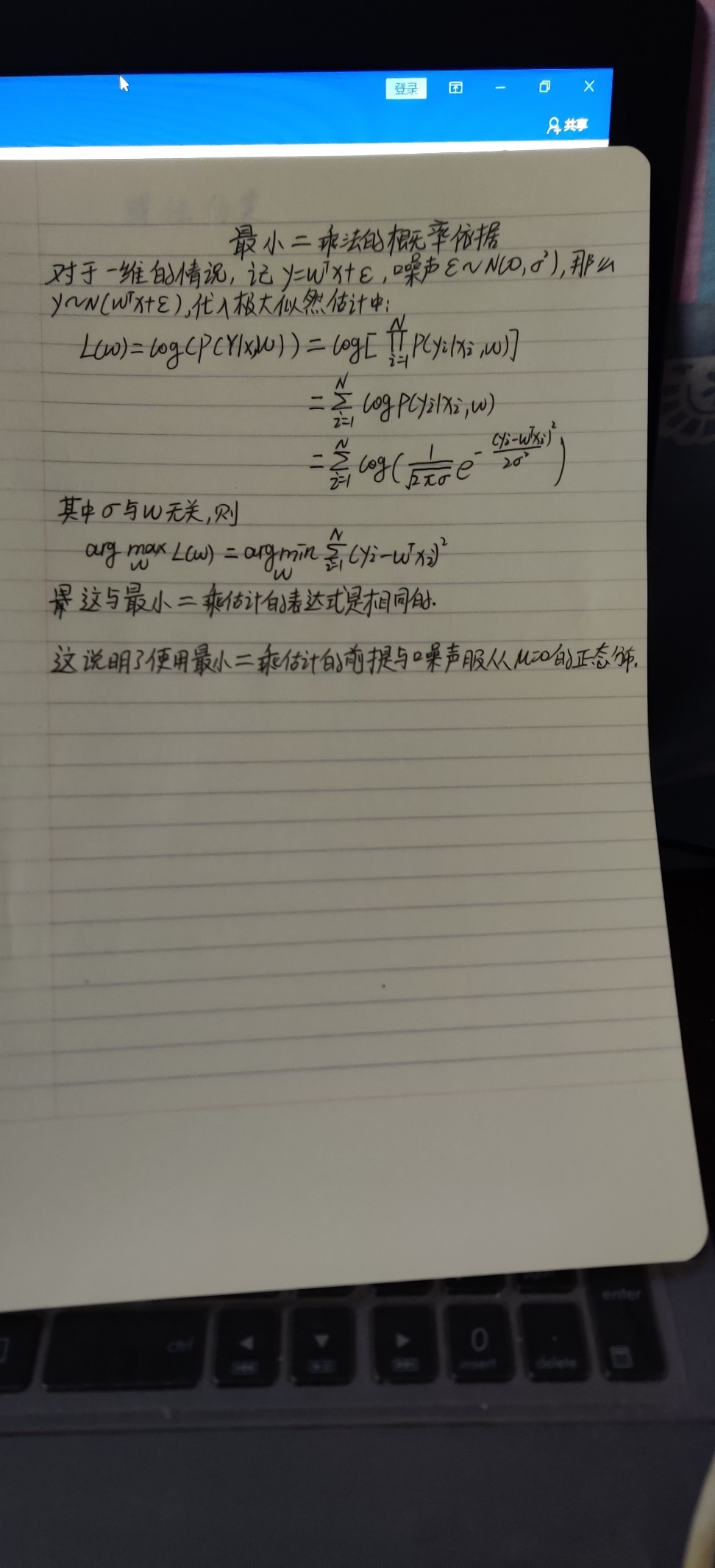
# 1 线性回归

## 1.1 最小二乘法

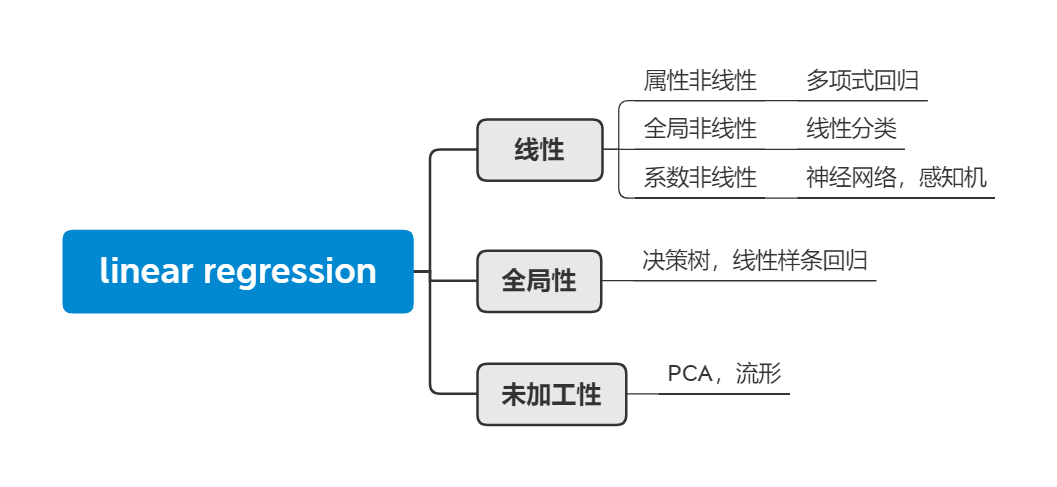


## 1.2从几何角度与向量空间角度理解最小二乘法

## 1.3 最小二乘法的概率依据



## 1.4 由LR引出机器学习框架



线性回归具有三个性质：线性，全局性，全局性。

线性中属性线性指的是变量x都是一次的，而不包含等；全局线性是指WTX计算出来多少就输出多少，而不会加上激活函数计算结果进行转换。

全局性指的是对整个样本空间进行拟合，学习，而不是讲样本空间划分成许多部分，再逐段进行拟合。

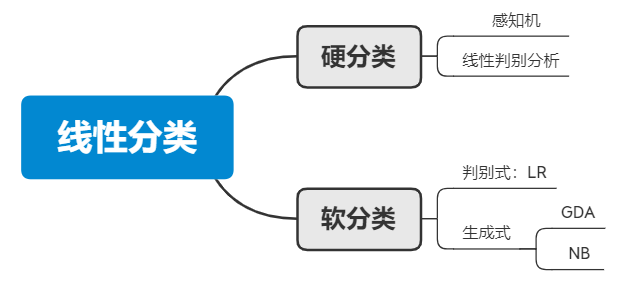
未加工性是指用原始数据进行拟合，没有对数据进行处理。

机器学习中的其他方法就是通过对线性回归中的一项或者是数项进行突破而得来得。

统计学习派是对LR进行突破，贝叶斯派也是么？等以后再补上回答

# 2 线性分类

## 2.1 线性分类结构



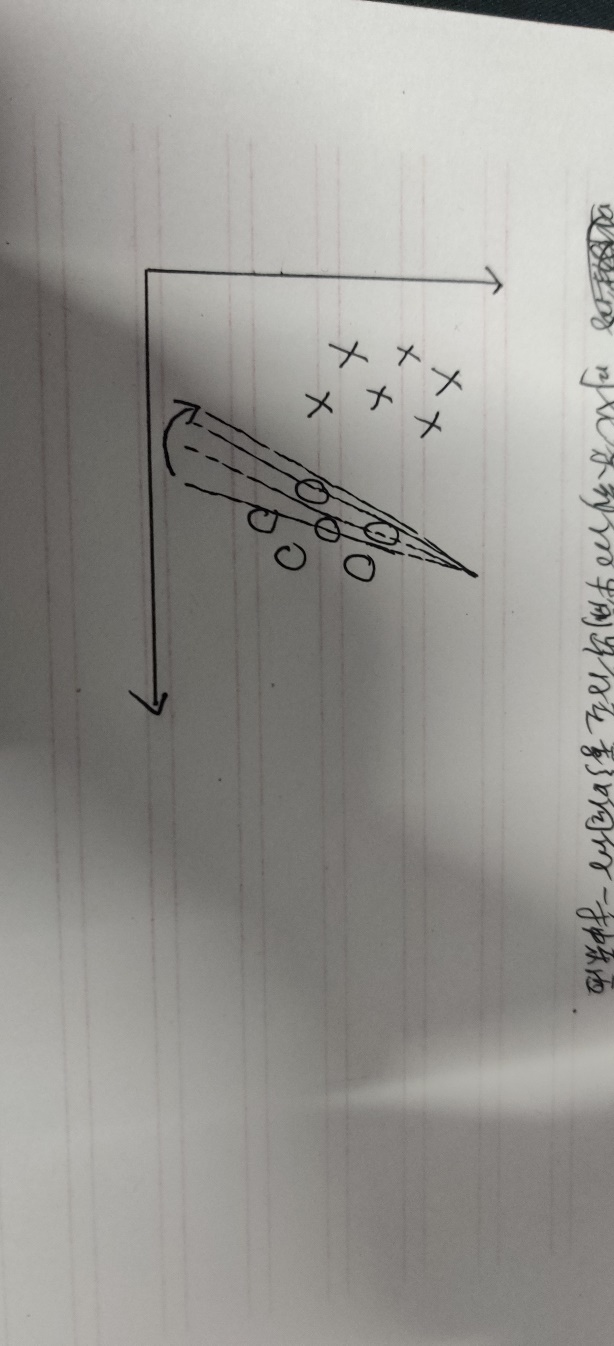
硬分类是指输出结果为整数，代表相应的类别；软分类是指输出的是概率。

在软分类中，判别式通过MLE直接计算模型的参数，最后输出的是样本属于某一类的概率；生成式通过MAP来计算，但是不会直接给出样本属于某个类别的概率，而是告知属于哪个类别的概率最大。

## 2.2 感知机

样本集：

### 主要思想：错误驱动，即通过改变模型参数来减少错误分类得样本数。



模型：，

### 代价函数

（1）首先想到得是用错误分类点的个数当作loss，定义loss function为：



上式中的I称为指示函数，但是指示函数是不可导的，因为I函数的输出值只有0或1，都不连续，所以难以进行优化。

为什么只是函数能表示错误分类点的个数？

因为符号函数的作用，预测值的具体数值并不重要，只有正负号才有用。那么实际上可以看作，只要大于0，那么就表示分类正确。

（2）由于wT的变化会引起的变化，且是连续的，所以直接使用作为loss，因为为负值，故加个负号：

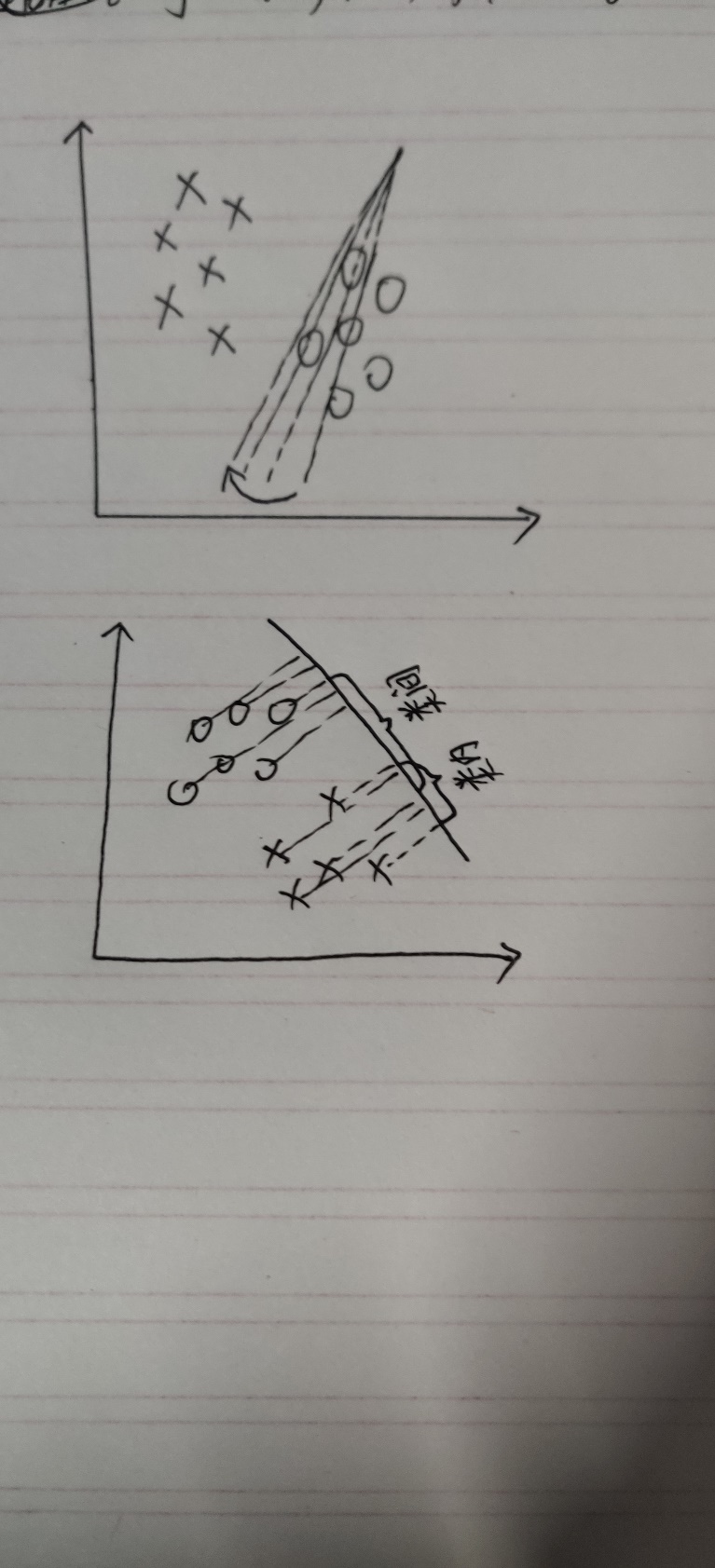


## 2.3 线性判别分析

主要思想

线性判别分析的思想是找到一个超平面，将原本的高维数据投影到该超平面上，使之线性可分。

**目标：类内方差小，类间距离大。**



## 2.4 logistic regression

### LR与perceptron的区别

两者的区别就在于激活函数的不同。perceptron的激活函数为符号函数，符号函数使得最后输出结果为类别，进而称为一个硬分类的线性分类方法；LR的激活函数为sigmoid，输出结果使介于0~1之间的概率，使得LR成为一个软分类的线性分类方法。

### 模型

数据集：

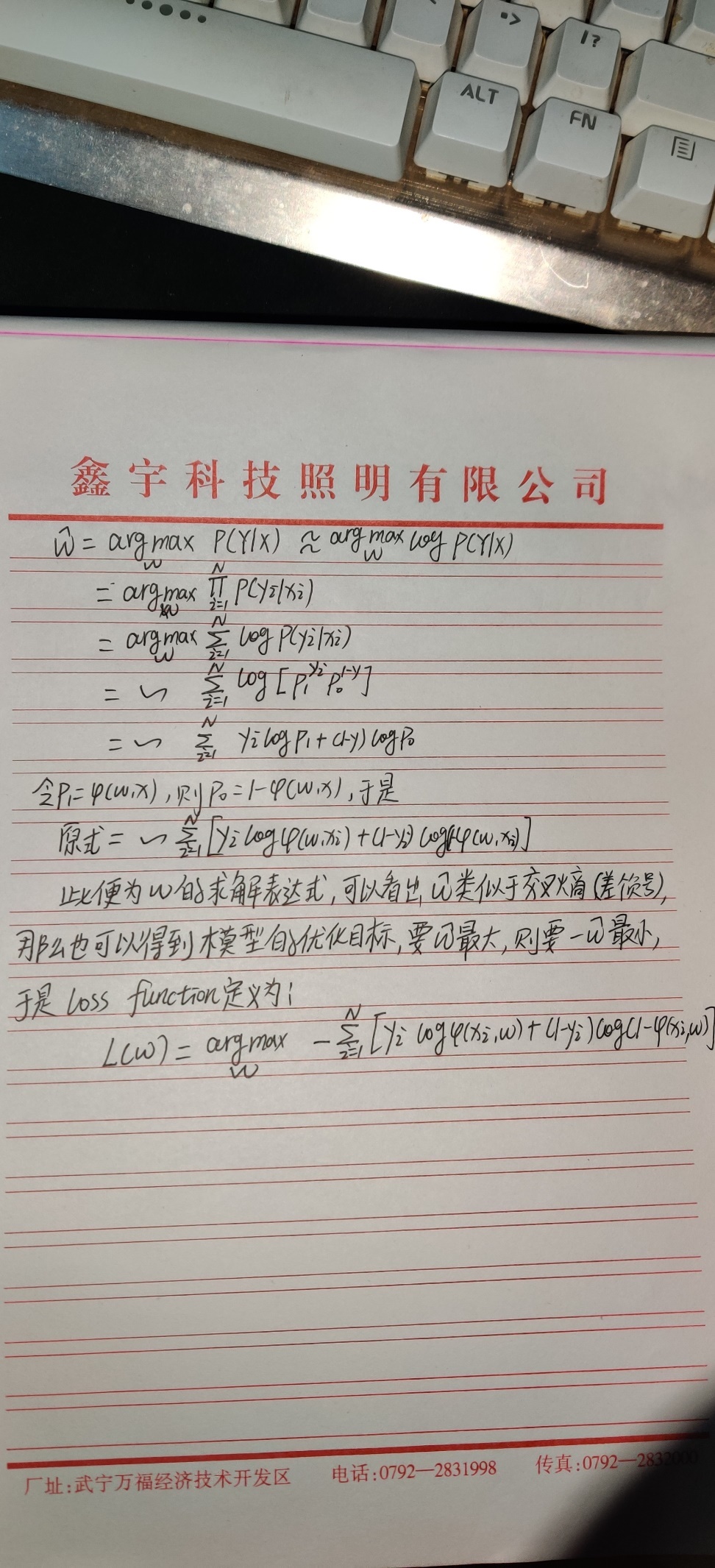
，或。

这两个式子其实是等价的前者表示样本属于1类的概率，后者表示样本属于0类的概率，前者与后者之和为1。

记前者为p1,后者为p0，将上面两个式子整合：



### 优化策略



## 2.5 Gaussian Discriminant Analysis

数据：







假设： 





**模型**：

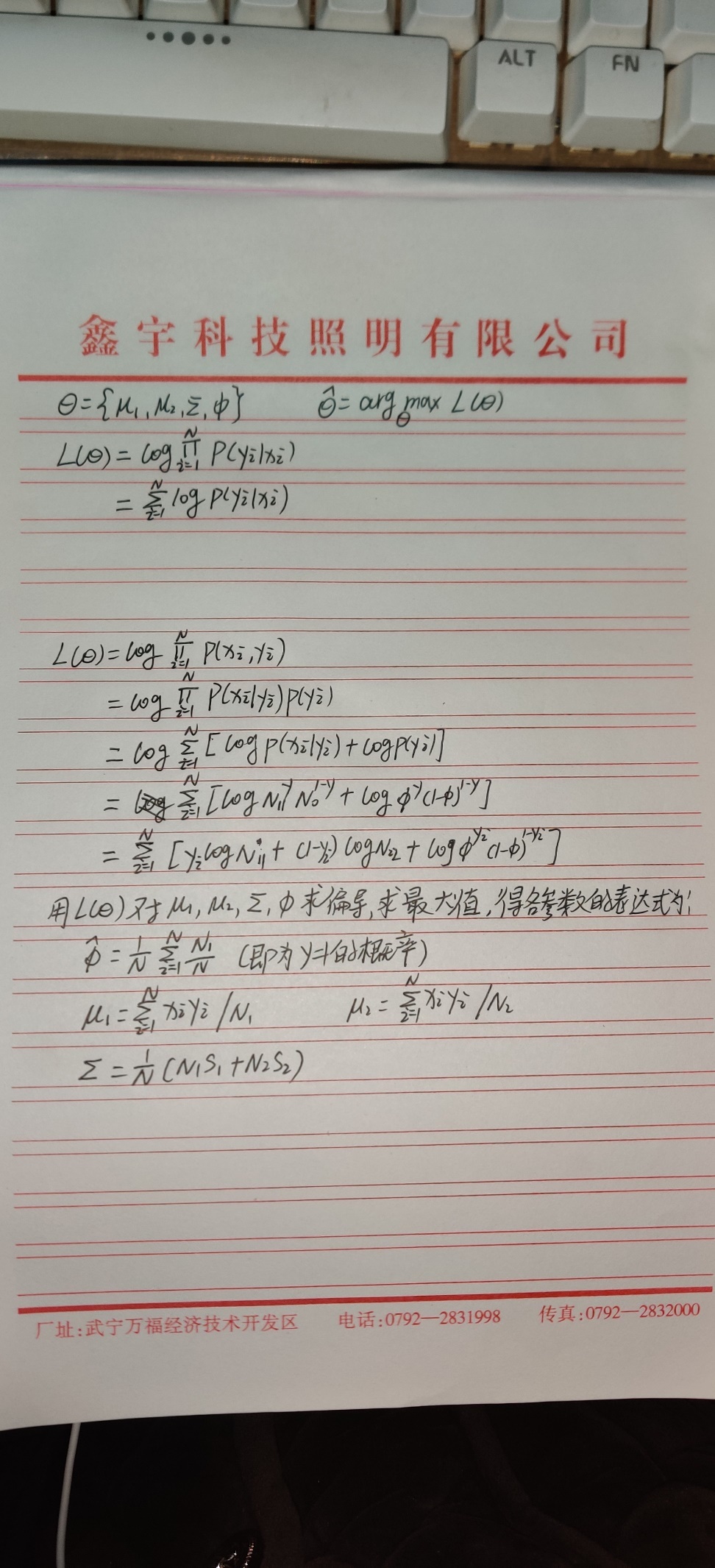
对于新输入的数据，通过计算并比较的大小，大的那个便是输出的预测值。

**学习策略**





**算法：**



## 2.5 Naïve Bayes

如果输入x是连续分布，可以使用GDA来进行分类；如果输入x是离散数据，则可以使用朴素贝叶斯来进行分类。

NB是基于贝叶斯原理与条件独立性假设的分类方法，由于这个假设太强了，所以‘朴素’。

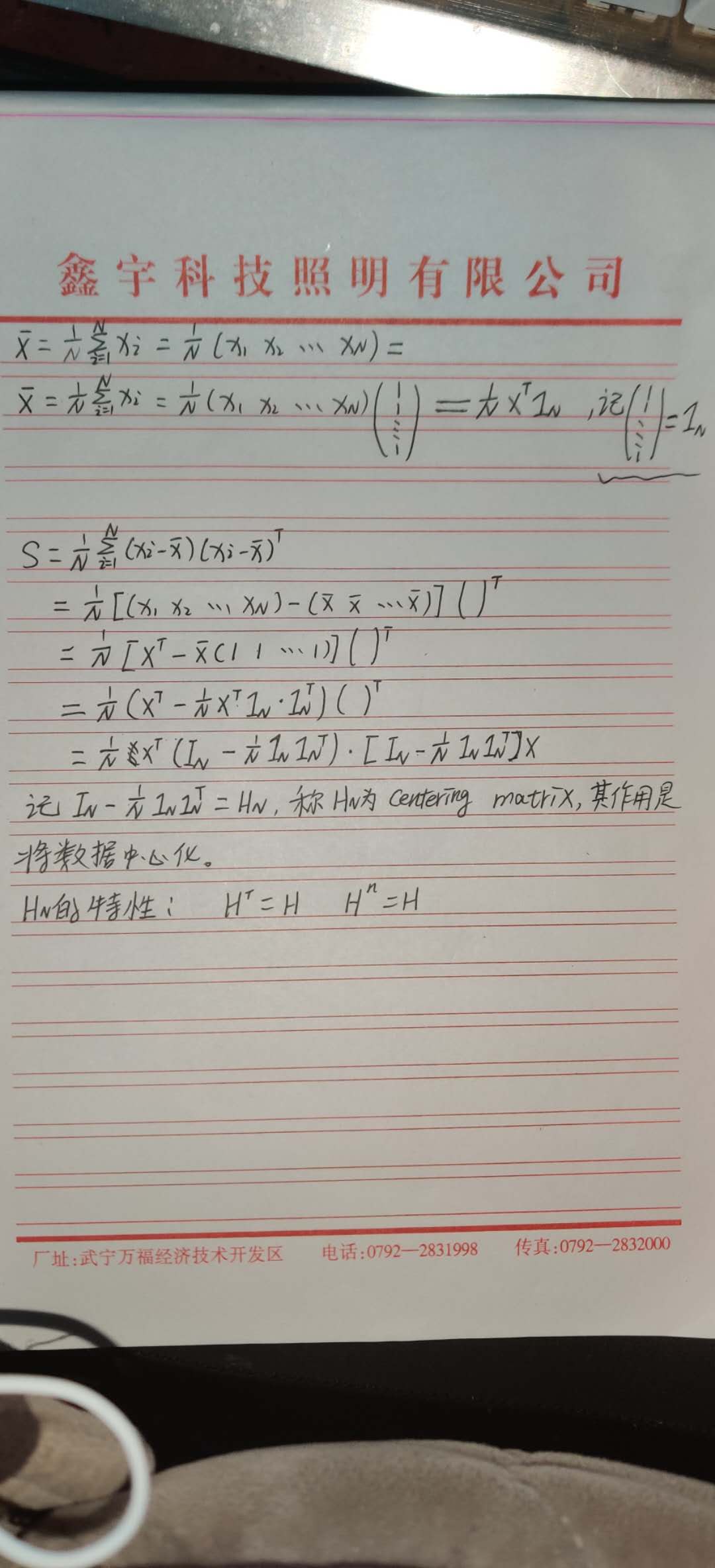
# 3 PCA

## 3.1 符号规定

DATA：

总体均值：

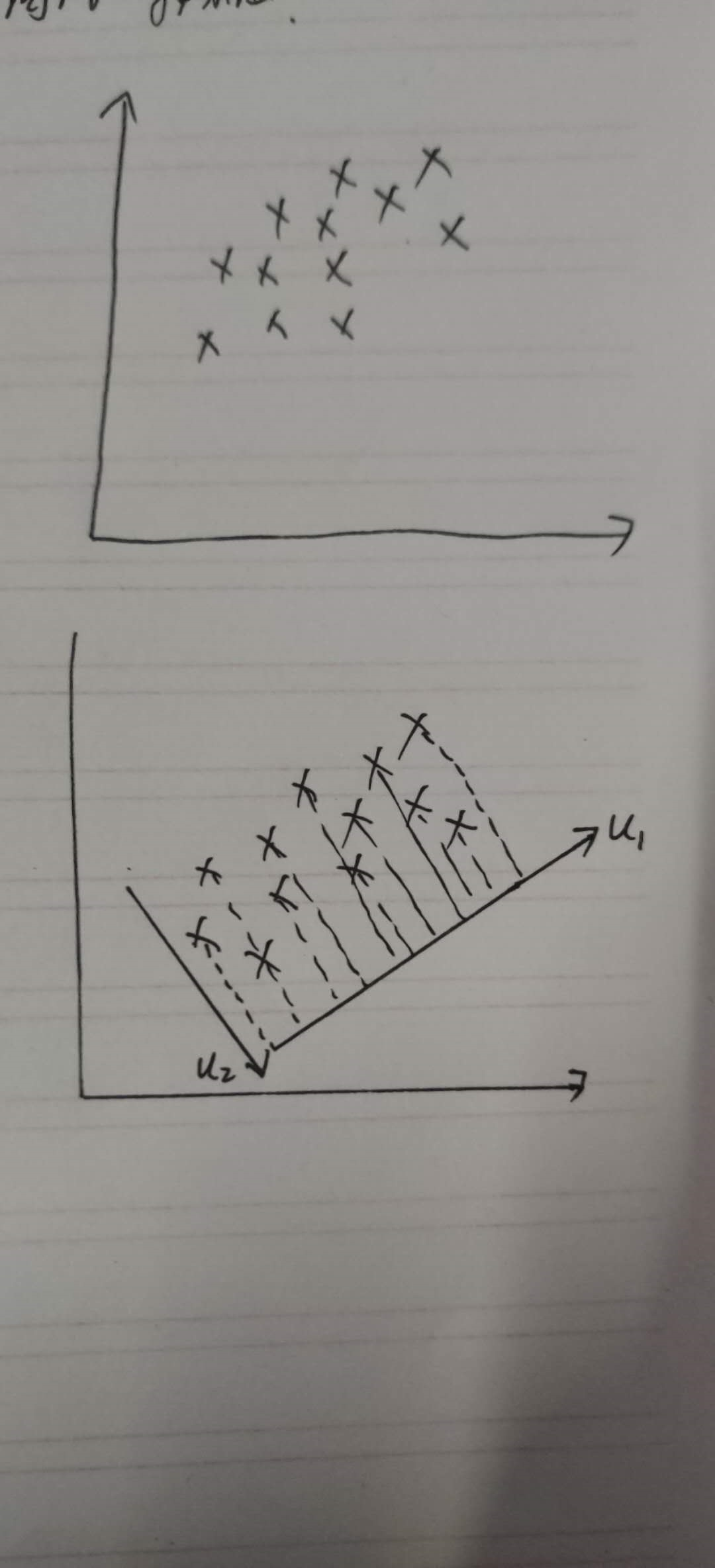
总体方差：



## 3.2 PCA主要思想

一个中心：将一组可能线性相关的变量，通过正交变换，变换成一组线性无关的变量。简而言之，对原始特征空间的重构。

两个基本点：最大投影误差，最小重构距离（代价）。



假设样本空间是z维的，PCA就是要找到一个新的投影方向u1，使得样本空间中的数据在u1的投影最为分散，跨度最长，也即方差最大；对应的，要将u1上的点恢复到原始样本空间也越简单，因为u1轴上的点彼此更为分散。

相反的，如果将样本点投影到u2轴上，那么样本点的投影跨度最小，也即方差小，点很密，那么重构能很难。

## 3.3 最大投影方差 定量化

设找到的投影方向维u1，且u1为单位向量，即。

定义函数J表示样本在u轴上投影的方差大小。



问题最后转化成了，在约束条件下，求解出使方差最大的投影方向u1，即：



通过拉格朗日乘数法，引入算子λ，建立函数



L对u1求偏导：



则λ为方差矩阵S的特征值矩阵，u1为S的特征向量矩阵。

对特征进行降维其实就是求出方差矩阵的特征值及其对应的特征向量。需要降至多少维，就由大到小取前面的特征值及其对应的特征向量。

**具体做法是：对S进行特征分解，取前前n个特征值及特征向量。**

## 3.4 最小重构距离 定量化

假设xi为中心化后的坐标，则x­i在新基中的表达式为



选取q为作为降维后的新基，则xi在q维新基中表示为：



定义重构代价函数Q：



于是问题转化为子啊约束条件下求解使Q取最小值的uk，即：



## SVD角度看PCA（PCoA）







T与S具有相同的征值，不同的使S是p\*p维的（数据的维度），T是N\*N维的（样本的数量），如果样本的数量小于数据的维度，那么对T操作会减少计算量。这种对T操作的算法，称为PCoA。

# 4 SVM

SVM不仅要分类，还要是的分类线两边离样本点的距离之和最大。

问题转换为：使分类线离最近样本点的距离最大。

硬间隔：考虑所有样本点。

软间隔：忽视部分离群点。

通过对核函数的应用，SVM能解决部分非线性问题。

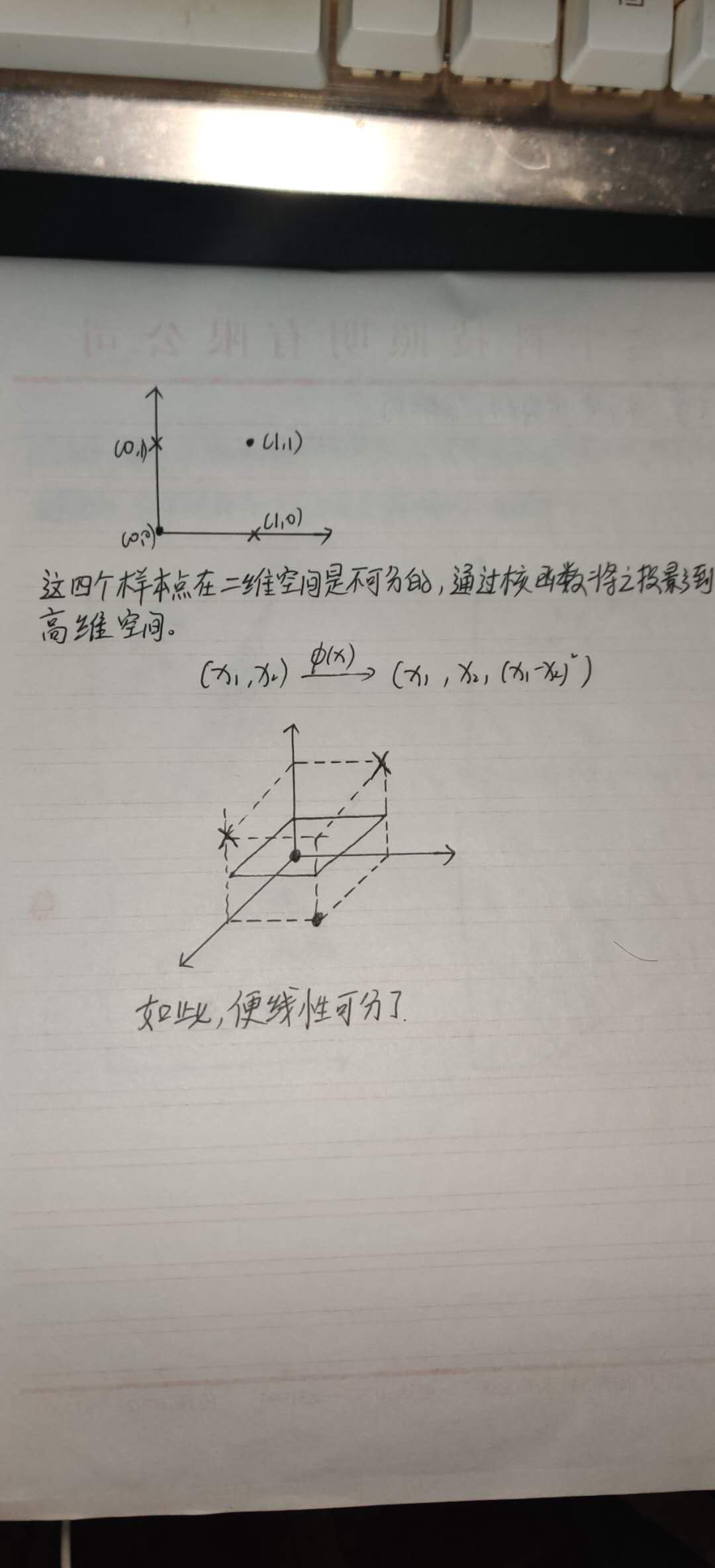
# 5 （kernel function）核函数

kernel function包括两层含义：

kernel method（思想方面）：将低维样本投影到高维空间，使原本线性不可分的问题变得线性可分。

kernel trick：和函数应该能直接计算内积。

例子：



# 6 EM算法(Expectation Maximization algorithm)

## 6.1为什么需要EM算法？

对于一般的简单的模型，我们可以使用MLE通过数据直接估计模型的参数。但是对于含隐变量的模型，MLE就不奏效了，于是EM算法闪亮登场了。换句话讲，**EM算法就是用来估计含隐变量模型的模型参数的。**

## 6.2 EM算法的总体步骤

假设被观察数据为， X与隐变量存在条件关系。

（X，Z）称为完备数据（完全数据），仅X称为不完备数据。

（1）给定一个参数值θ

（2）通过θ与X推导出Z的分布q(z)

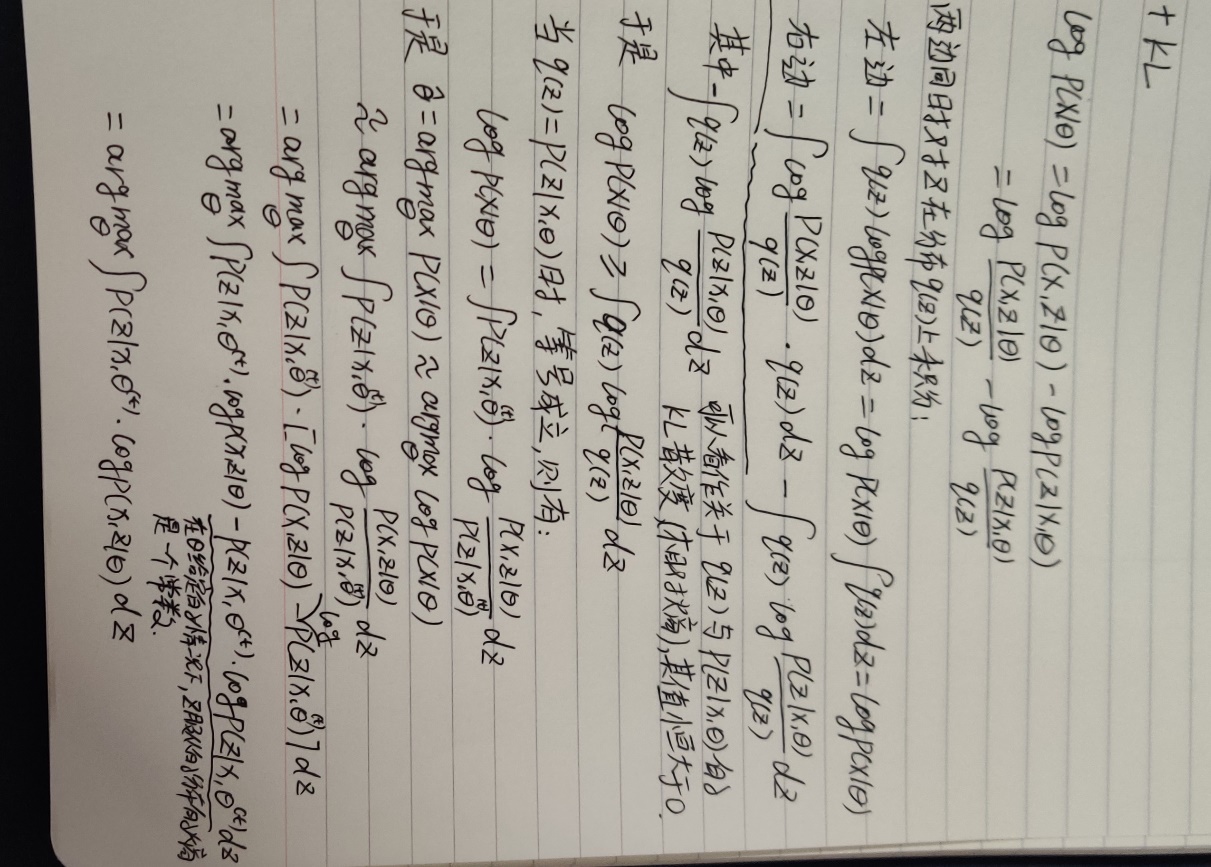
（3）写出X，Z的联合概率密度在Z的分布q(z)下的期望的表达式(E-step)

（4）求出使表达式取最大值的θ(t+1) (M-step)

重复以上步骤，直至θ收敛。

## 6.3 EM算法的推导

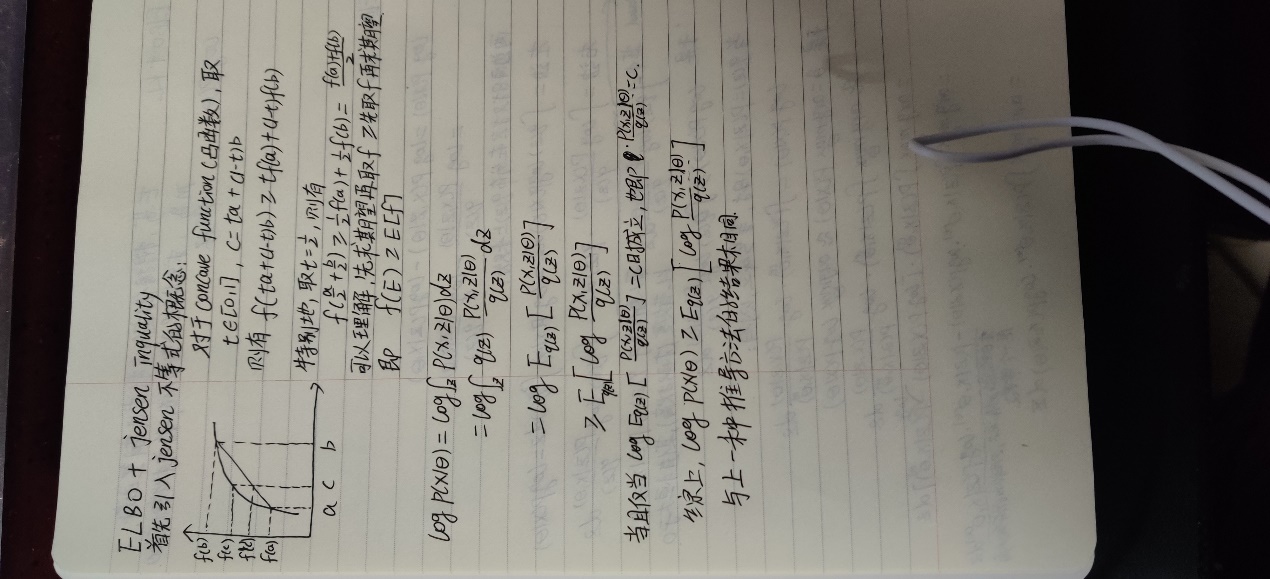
### 6.3.1 ELBO+KL



要最大，其实就是要下界ELBO最大。

其中，最大化的期望称为函数基于条件分布的期望。

### 6.3.2 ELBO+Jensen



## 6.4 广义EM算法

EM算法需要通过预先设定θ推导出z的后验分布，进而推测更合理的θ。但是实际情况下，后验概率可能无法计算，但是我们可以通过使KL(p||q)最小化，用q(z)来逼近p(z|x,θ)，进而更新θ。



# 7 变分推断（variational inference）

## 7.1 VI与EM的联系与区别

VI与EM都基于，都是用q(z)去逼近Z的后验概率。

EM的目的是通过推测隐变量Z的分布求解X服从的分布的参数，所以EM算法需要求解Z的分布与X的参数θ。但是VI算法只关注于求解Z的的分布，而不关心θ（将之作为一个未知的常数）。

## 7.2 基于平均场理论的VI

采用坐标上升法的思想，将q(z)分为数个部分，先固定其他部分求解q（zi）；进而更行其他部分。如此迭代。

## 7.3 基于梯度上升法的VI（SGVI，SGVB）

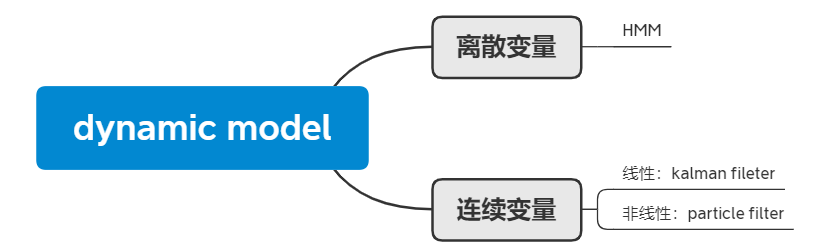
设分布q(z)由参数表达，则问题转换为求解q(z)的参数。通过最大似然估计求解ELBO的最大值，可以得到的表达式。由于表达式是一个期望，所以可以通过蒙特卡洛采样法来表示。

但是由于的表达式的方差很大，在实际中很难实现，用MC的方法也难以表示，故采用充参数优化的方法对的表达式进行优化。

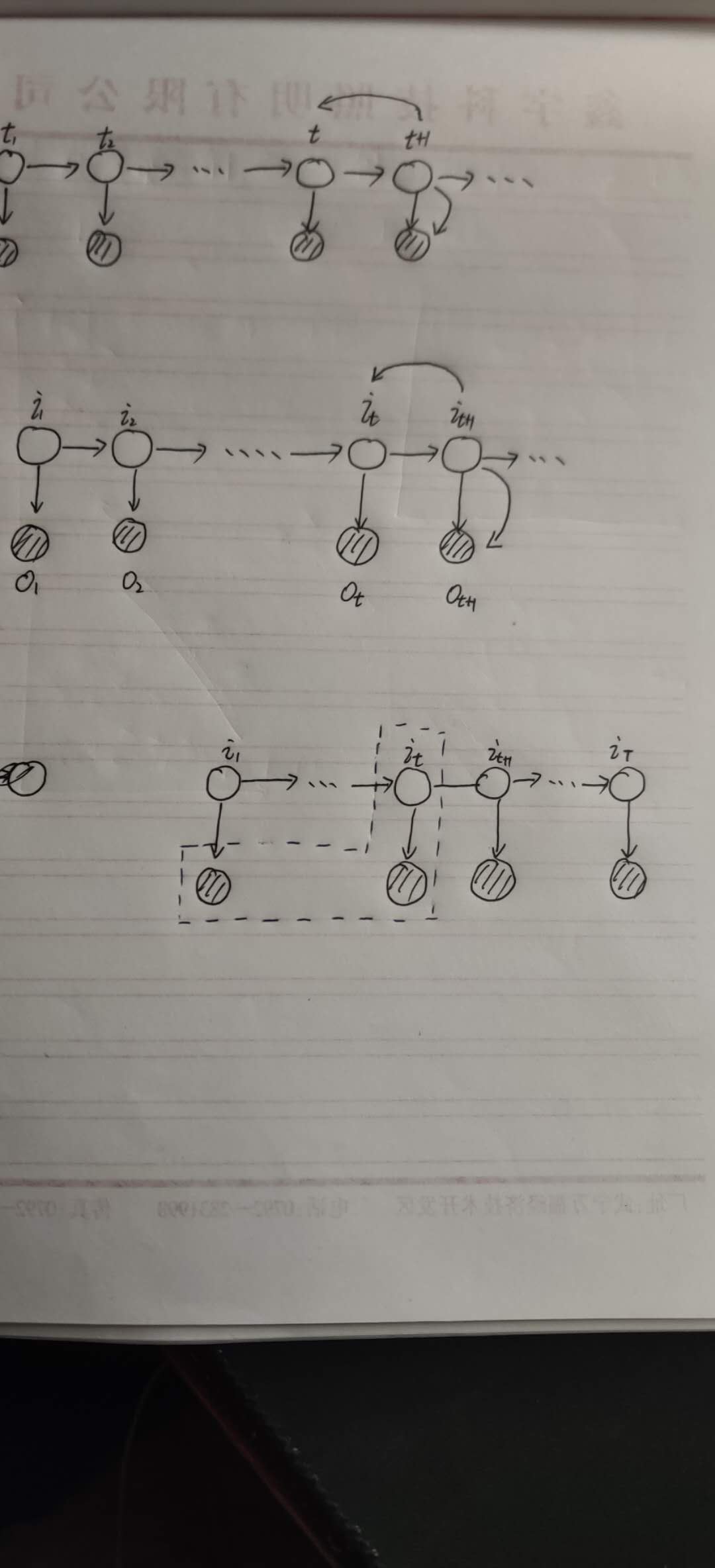
# 8 HMM

在概率图模型的基础上加一个时间维度就成了动态模型（注意动态模型的随机变量不是IID的），如果随机变量是离散的，那么就适用于HMM。

**HMM模型用一句话总结就是：一个模型，两个假设，三个问题。**



## 8.1 一个模型（标记解释）



模型：

其中为隐状态的初始概率分布。

A称为状态转移矩阵，B为发射矩阵。





观测变量机器取值域：



状态变量机器取值域：



## 8.2 两个假设





## 8.3 evaluation问题

evaluation问题总的来说就是：

### 8.2.1 暴力计算法



其中，

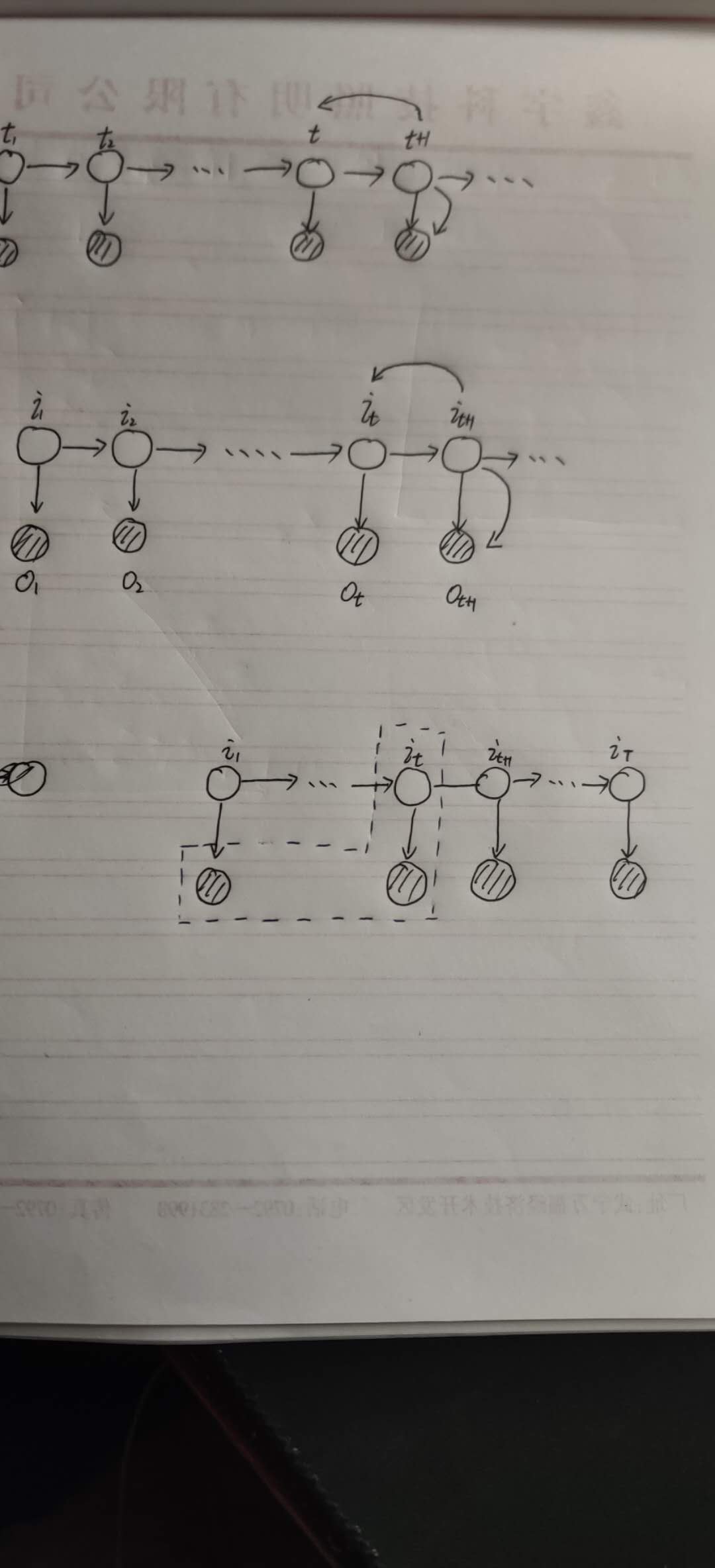




故：

这里对I的连加表示的是对可能的状态序列的连加。因为产生观测序列的状态序列有很多，状态序列的组合数a与观测序列元素数量b和状态序列取值域元素总数c有关系，a=bc。因为计算量是指数级别的，计算量太大，故这个方法不实用。

### 8.2.2 forward algorithm(前向算法)

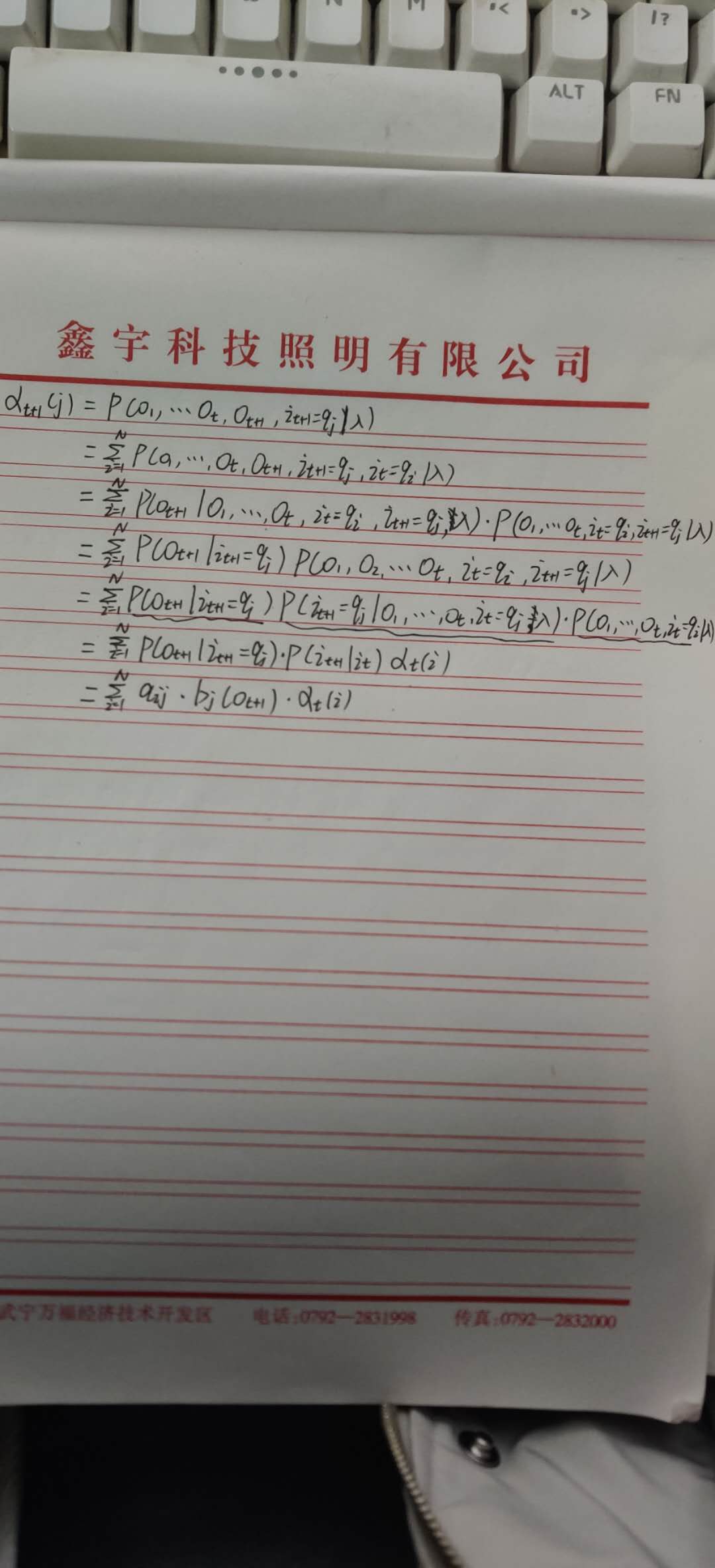




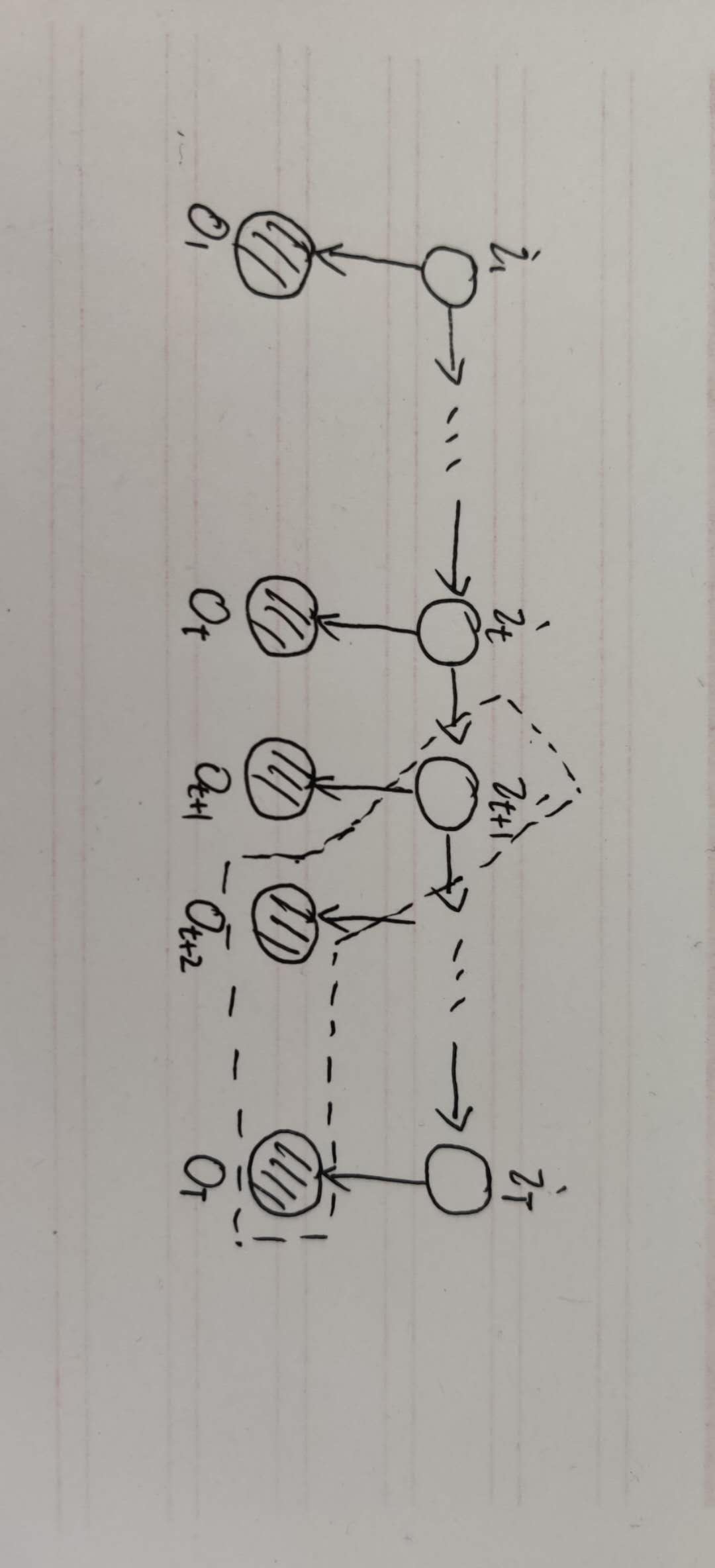


记为在给定λ的情况下，的联合条件概率密度。则有：





### 8.2.3 后向算法









即：



所以：



## 8.4 learning问题

通过baum weltch 算法（EM）求解模型参数。

## 8.5 decoding问题

decoding问题即求解使得观测序列出现概率最大的隐状态序列，即求解。viterbi 算法会考虑整体的最大值情况。viterbi算法的思想与动态固化类似，所以就要找到局部迭代公式。

<https://www.cnblogs.com/pinard/p/6991852.html>



其中保存的是从前一时刻各个可能的隐状态转移到当前各个可能的应状态的最大概率值，比如当前有m个可能的状态，前一时刻有n个可能的状态，那么保存的便是由前一时刻的n个状态值分别转移到当前时刻t状态1的n个概率值的中的最大值。当前由m个可能的状态，那么要保存m个概率——。

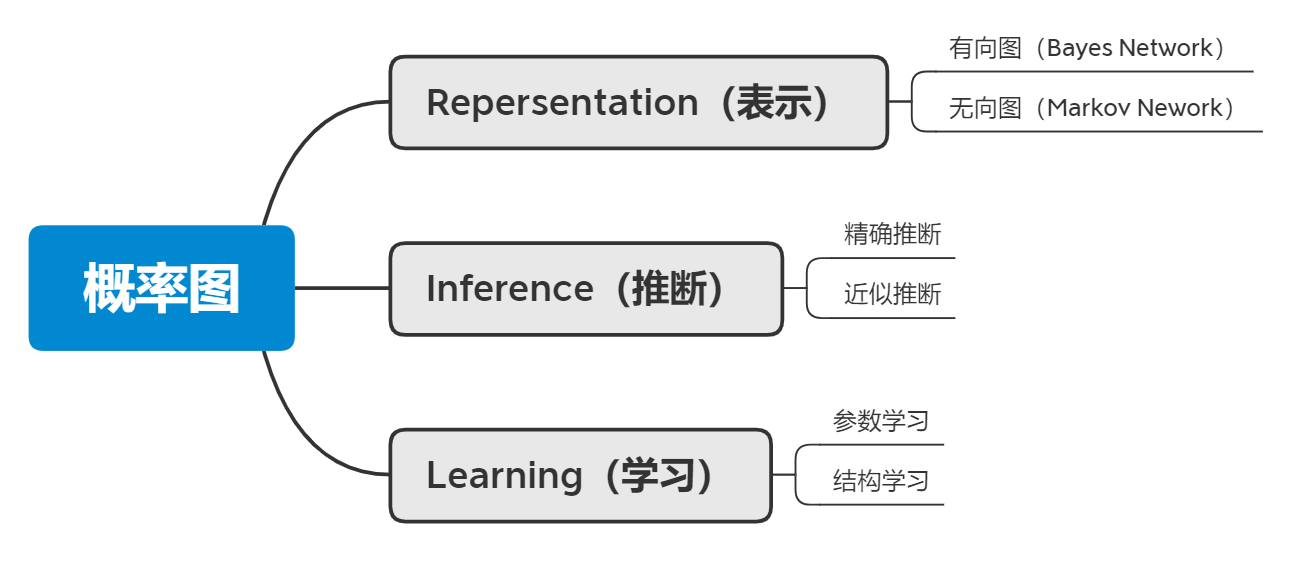


保存的是转移t时刻状态i的最大概率是由前一时刻的那个状态转移过来的。比如表示的使从t-1时刻的状态2转移到当前时刻t状态1的概率最大。

那么顺着时间轴一个时刻一个时刻的计算可以求得最大的，然后通过可以一步步向前推，最终确定最可能的隐状态序列。

# 9 概率图模型

概率图模型（Probability Graph Model, PGM），简称图模型，是指用图结构来表示多元随机变量之间的条件独立关系的概率模型。条件独立性在概率图中是能够直接观测出来的。

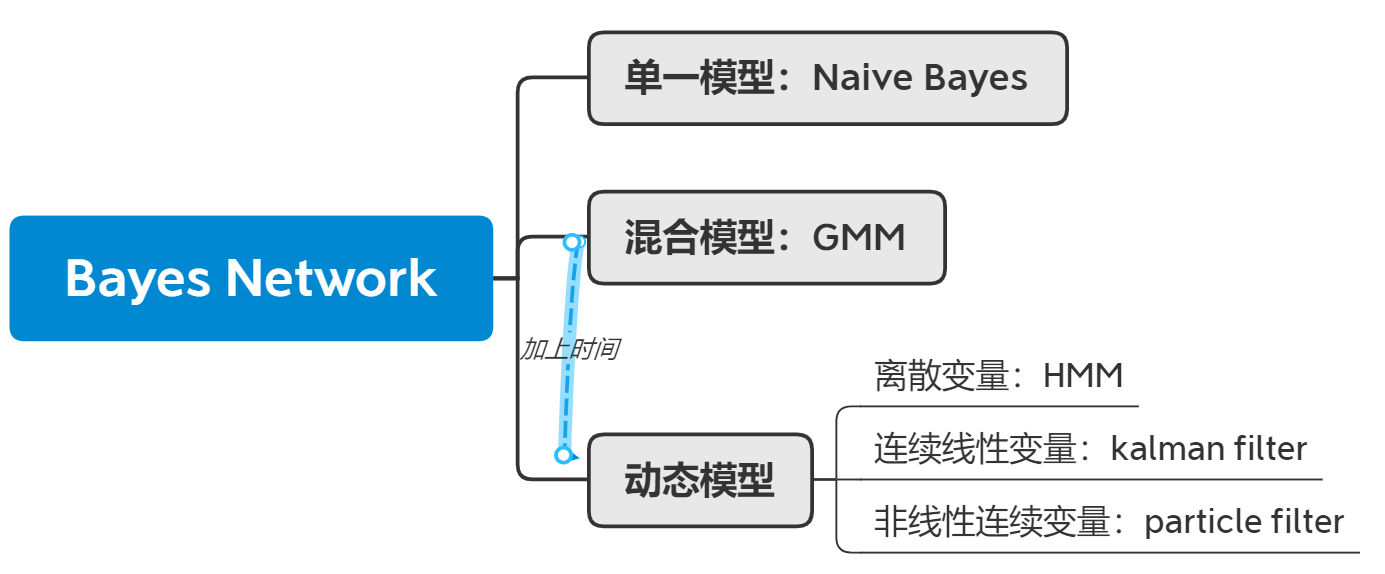


## 9.1 Representation

Representation解决的是概率模型如何用图的结构表示。

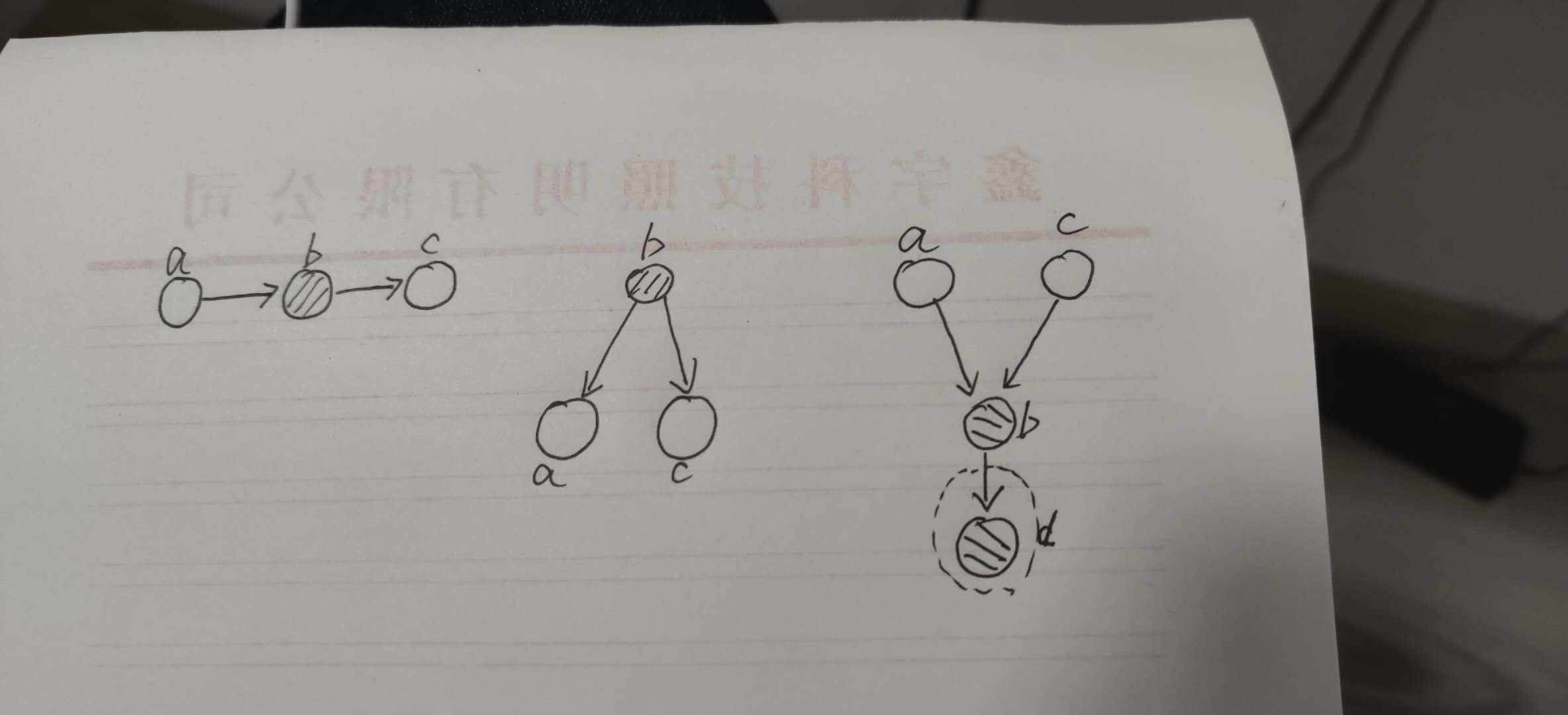
### 9.1.1 有向图

有向图是指节点与节点之间用带箭头的线连接的图模型（无循环）。在有向图中，如果两个节点相连接，那么他们一定服从条件关系，父节点是“因”，子节点是“果”。



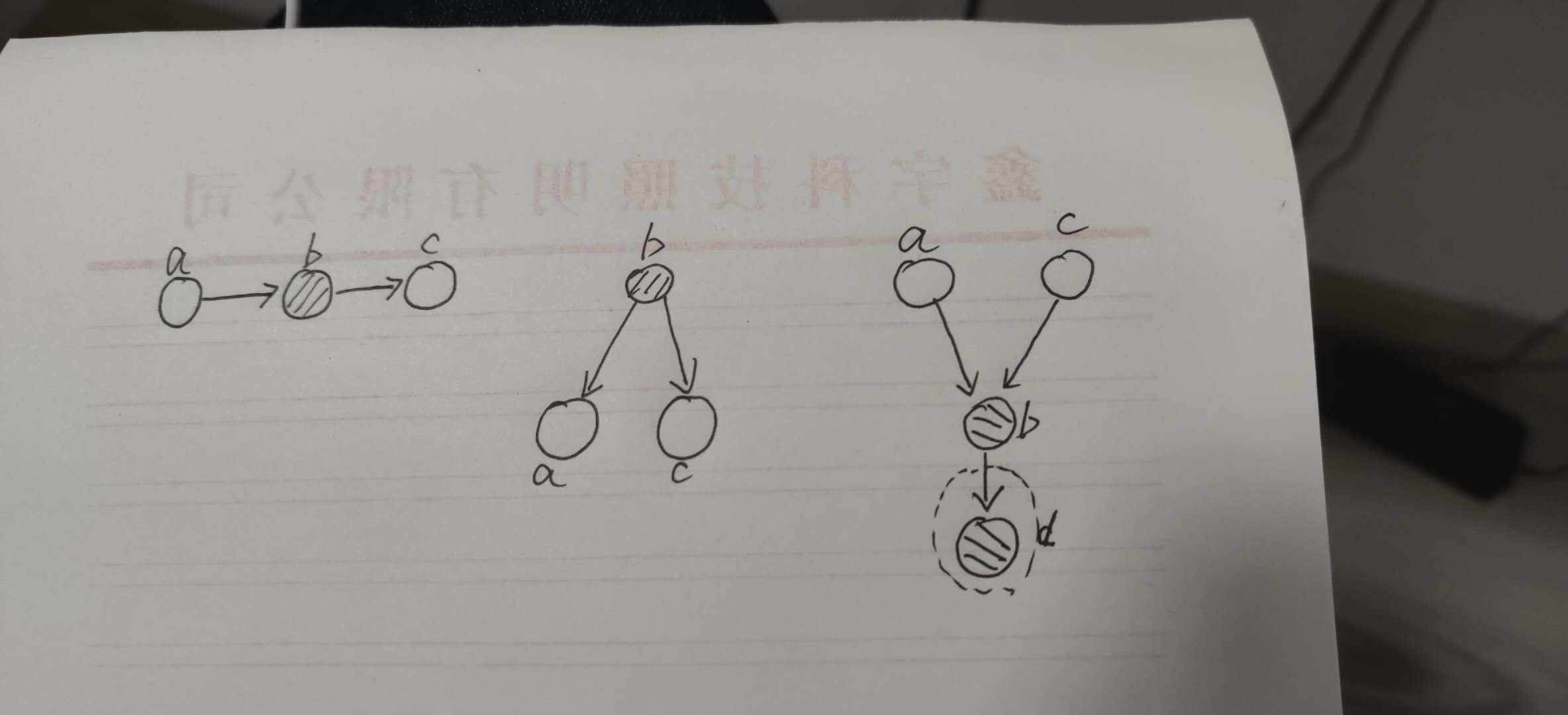
有向图中有三种结构

(1)head to tail



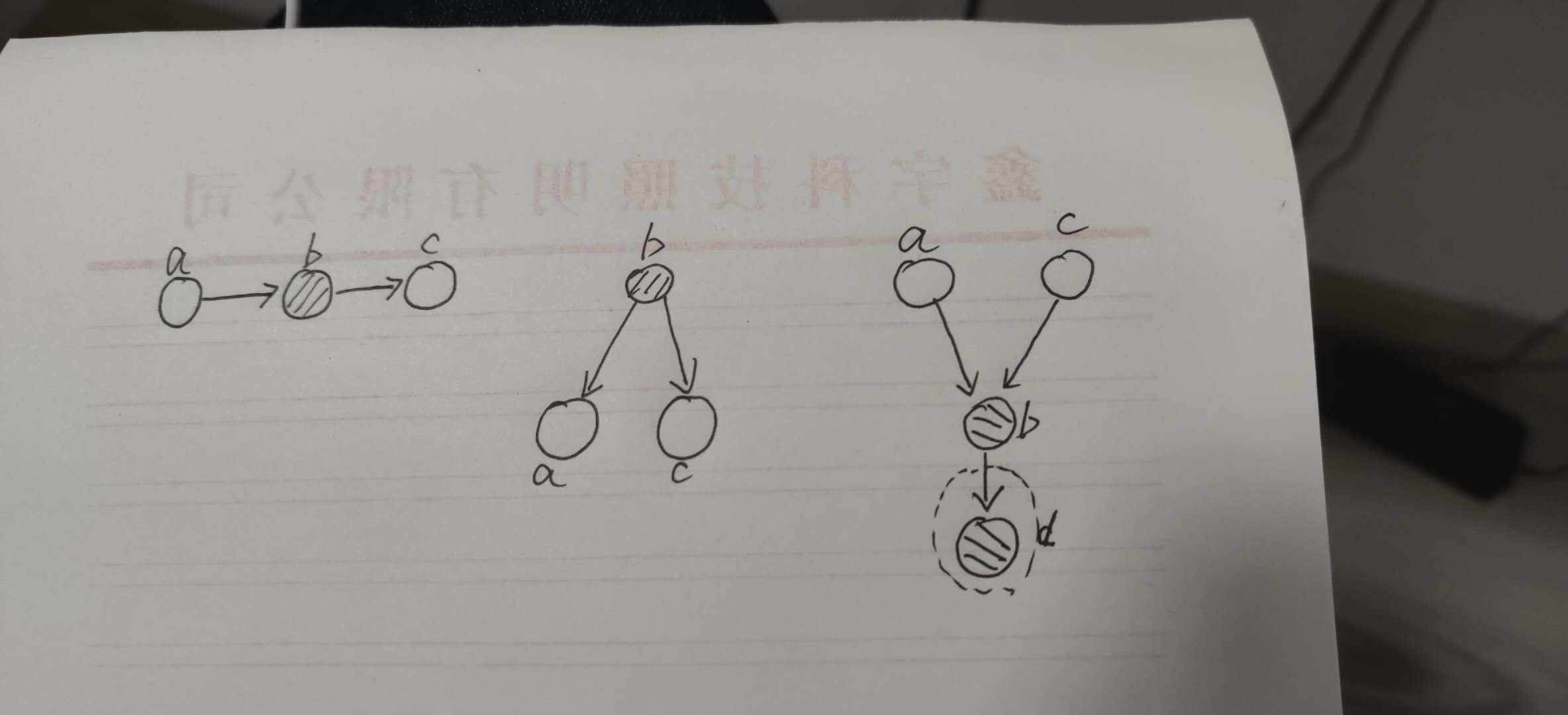
在b没有倍观测（确定）之前，a到c的路径是连通的，即a于c不是独立的。在b倍观测之后，a到c的路径堵塞，则a与c相独立。

(2)tail to tail



同上。

(3)head to head



在b(d)未被观测之前，a与c独立；在b(d)被观测之后，a与c不相互独立。

将上述三种结构进行归纳，就得到了**D-separation。**

### 9.1.2 无向图

无向图模型，也称为马尔可夫随机场（符合Markov property的随机过程）或马尔科夫网络，是一类用无向图来描述一组具有局部马尔可夫性质的随机向量的联合概率分布的模型。

**无向图中判断条件独立性：**

（1）Global Markov

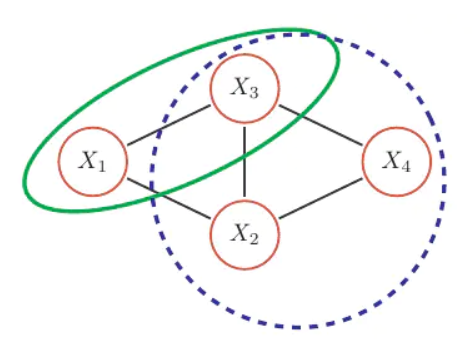


（2）Local Markov

只要不直接相连的两个随机变量就是相互独立的。

（3）成对Markov

**无向图模型的概率分解**



团：由于无向图模型并不提供一个变量的拓扑顺序，因此无法用链式法则对进行逐一分解。无向图模型的联合概率一般以全连通子图为单位进行分解。无向图中的一个全连通子图，称为团（Clique），即团内的所有节点之间都连边。在所有团中，如果一个团不能被其它的团包含，这个团就是一个最大团（Maximal Clique）。

### 9.1.3 Markov Blanket

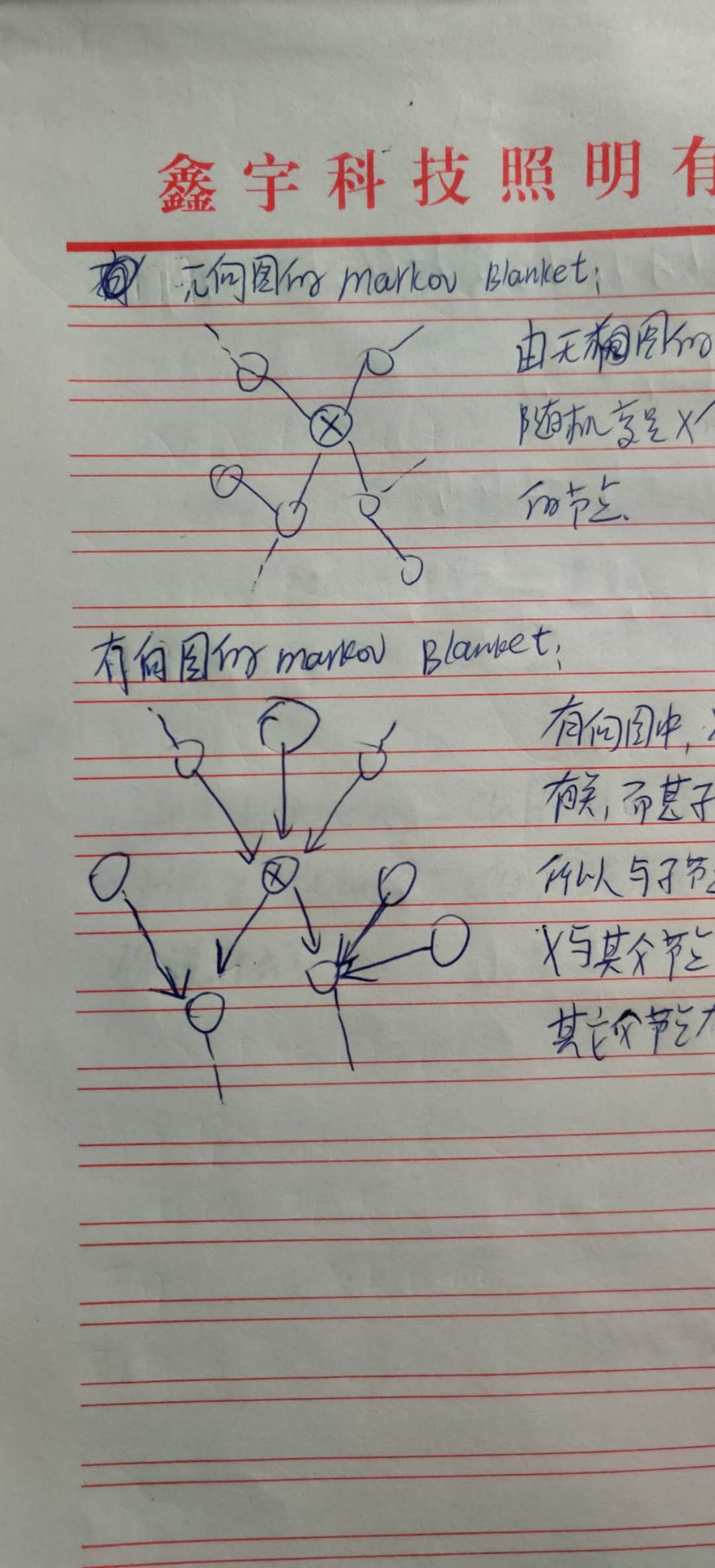
马尔科夫毯：在随机变量的全集U中，对于给定的随机变量X∈U和变量集MB∈U（），若有：



则称能满足上述条件的最小变量集MB为X的Markov Blanket。

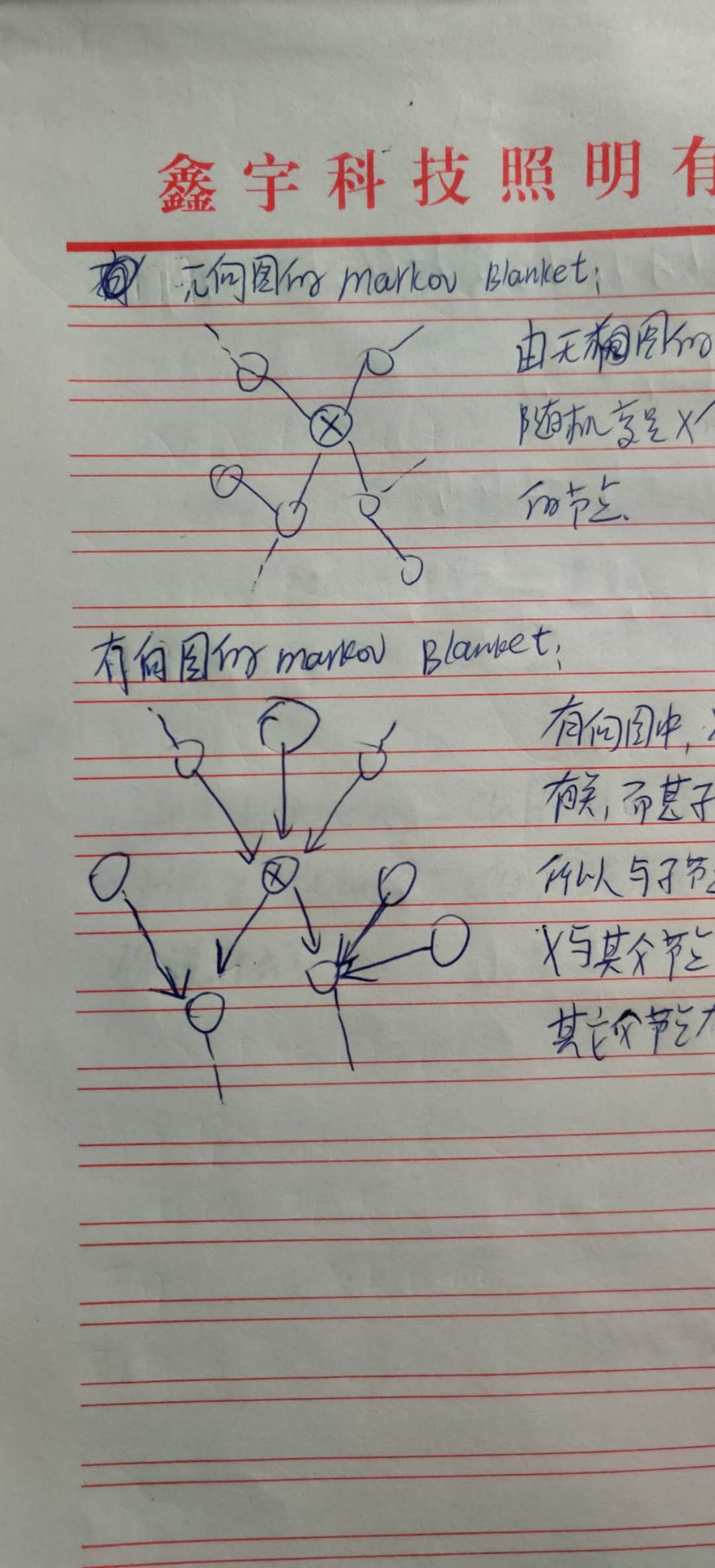
通俗来理解，给定随机变量X的邻居MB，则X与除MB以外的所有变量都独立，则成最小的MB为X的Markov Blanket。

（1）无向图的Markov Blanket



无向图中随机变量X的Markov Blanket是指直接与X相连的随机变量。

（2）有向图的Markov Blanket



有向图的Markov Blanket是指X的父节点、X的子节点以及子节点的父节点。

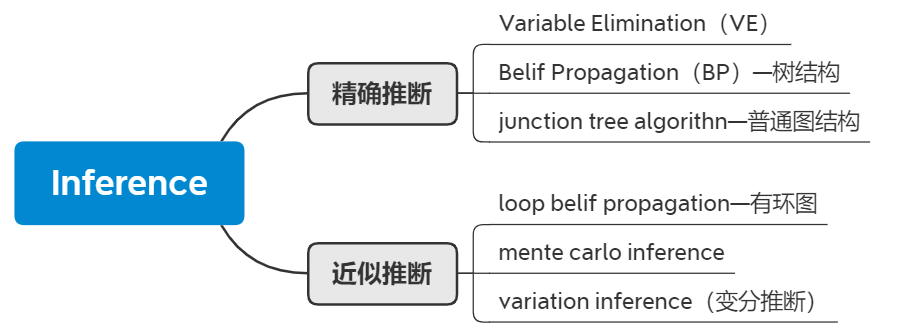
## 9.2 Inference

Inference解决的概率的求解问题。

具体来讲就是通过已知的联合概率密度求解：

边缘概率，条件概率，

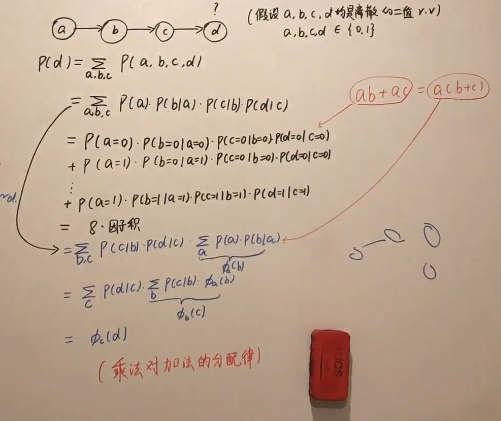
最大后验概率。



### 9.2.1 variation elimination

变量消除的实质其实就是乘法对加法的分配律：ac+bc=c(a+b)

就是先做加法在做乘法。



VE的问题在于：

（1）重复计算（计算过后，如果要计算，只能从头计算过）

（2）计算复杂度与计算顺序有很大关系。

### 9.2.2 belif propagation

Pass

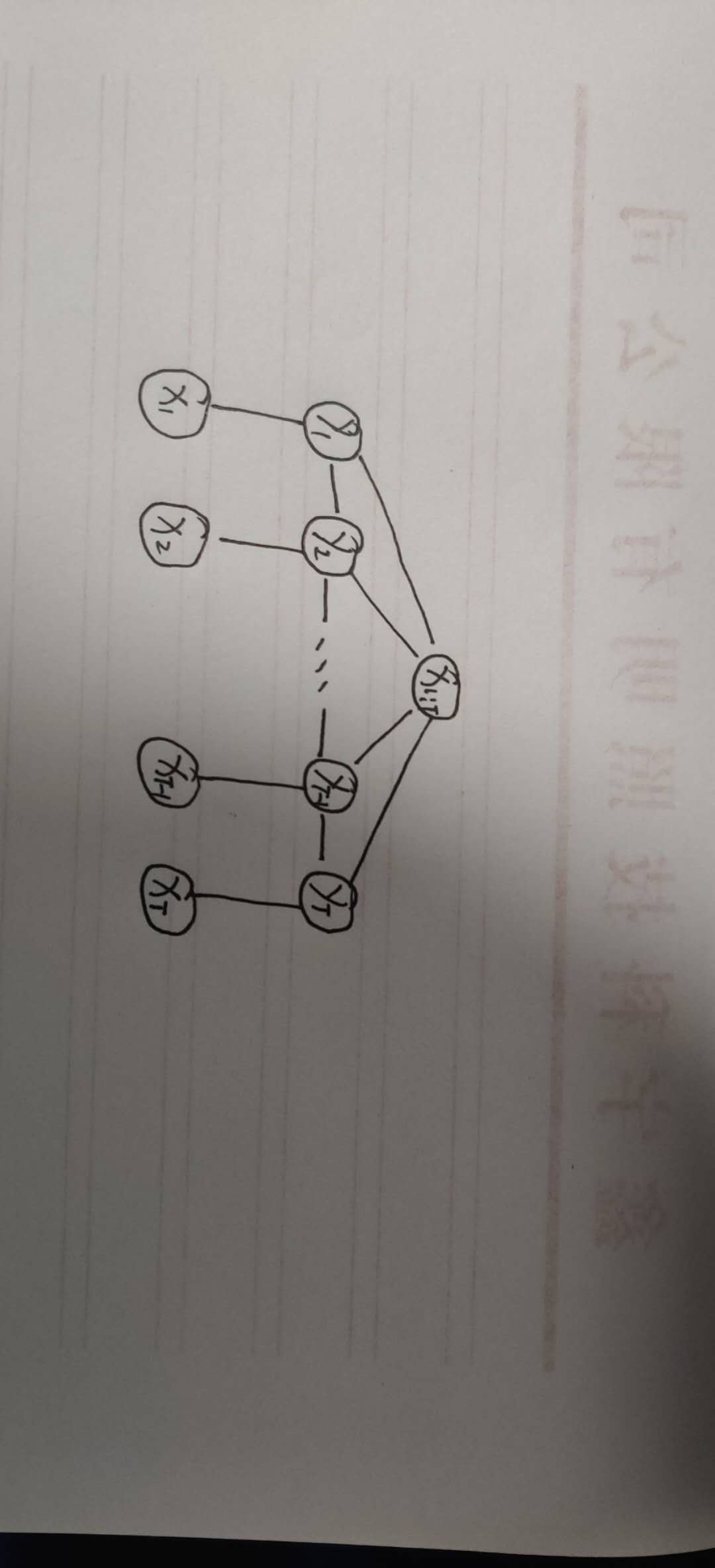
# 10 条件随机场（CRF）

在HMM中有两个假设：观测独立假设和齐次Markov假设。MEMM（maximum entropy Markov model）结合了HMM和最大熵模型（maximum entropy model）的思想，打破了HMM的齐次对立假设（在MEMM中，xi不仅影响yi，而且对其他的应状态都可能有影响），其概率图如下：



但是MEMM也存在问题，那便是**由局部归一化引起的label bias problem**。

所以有提出了以无向图模型为基础的CRF，CRF是天生全局归一化的。由有向图变为无向图，也打破了HMM的齐次Markov假设，其概率图如下：



随机场：随机场是由若干个位置组成的整体，当给每一个位置中按照某种分布随机赋予一个值之后，其全体就叫做随机场。还是举词性标注的例子：假如我们有一个十个词形成的句子需要做词性标注。这十个词每个词的词性可以在我们已知的词性集合（名词，动词...)中去选择。当我们为每个词选择完词性后，这就形成了一个随机场。

Markov随机场：马尔科夫随机场是随机场的特例，它假设随机场中某一个位置的赋值仅仅与和它相邻的位置的赋值有关，和与其不相邻的位置的赋值无关。

条件随机场：设X与Y是随机变量，P(Y|X)是给定X时Y的条件概率分布，若随机变量Y构成的是一个马尔科夫随机场，则称条件概率分布P(Y|X)是条件随机场。

线性链条件随机场：其实就是观察状态序列与隐状态序列长度相等的CRF。

linear-CRF中，特征函数分为两类。第一类是定义在YY节点上的节点特征函数，这类特征函数只和当前节点有关，记为：

sl(yi,x,i),l=1,2,...L

其中LL是定义在该节点的节点特征函数的总个数，ii是当前节点在序列的位置。

第二类是定义在YY上下文的局部特征函数，这类特征函数只和当前节点和上一个节点有关，记为：

tk(yi−1,yi,x,i),k=1,2,...K

　其中KK是定义在该节点的局部特征函数的总个数，ii是当前节点在序列的位置。

之所以只有上下文相关的局部特征函数，没有不相邻节点之间的特征函数，是因为我们的linear-CRF满足马尔科夫性。

无论是节点特征函数还是局部特征函数，它们的取值只能是0或者1。即满足特征条件或者不满足特征条件。同时，我们可以为每个特征函数赋予一个权值，用以表达我们对这个特征函数的信任度。假设tktk的权重系数是λkλk,slsl的权重系数是μlμl,则linear-CRF由我们所有的tk,λk,sl,μltk,λk,sl,μl共同决定。

此时我们得到了linear-CRF的参数化形式如下：

　其中，Z(x)Z(x)为规范化因子。  
回到特征函数本身，每个特征函数定义了一个linear-CRF的规则，则其系数定义了这个规则的可信度。所有的规则和其可信度一起构成了我们的linear-CRF的最终的条件概率分布。