最大熵与EM算法

Machine Learning Engineer

机器学习工程师

讲师：加号

目录

CONTENTS

01

统计学基础回顾

1.1 先验概率与后验概率

1.2 极大似然估计(MLE)

1.1 先验概率与后验概率

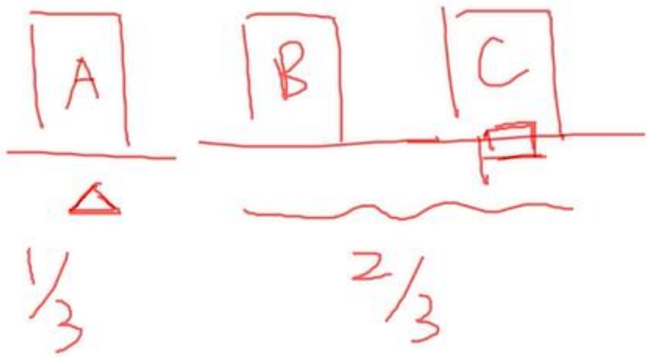
先验概率：根据以往经验和分析得到的概率，如全概率公式，它往往作为”由因求果”问题中的”因”出现。

后验概率：依据得到”结果”信息所计算出的最有可能是那种事件发生，如贝叶斯公式中的，是”执果寻因”问题中的”因”。后验概率可以根据通过贝叶斯公式，用先验概率和似然函数计算出来。

贝叶斯定理：假设B1,B2,...,Bn互斥且构成一个完全事件，已知它们的概率P(Bi),i=1,2,...,n，现观察到某事件A与B1,B2,...,Bn相伴随机出现，且已知条件概率P(A|Bi)，求P(Bi|A)。

$$P(B_i|A)=\frac{P(B_i)P(A|B_i)}{\sum_{j=1}^{n} P(B_j)P(A|B_j)}$$

先验概率与后验概率



$$P(\text{不换}) = \frac{1}{3} \times 100\% + \frac{2}{3} \times 0\% = \frac{1}{3}$$

$$P(\text{换}) = \frac{1}{3} \times 0\% + \frac{2}{3} \times 100\% = \frac{2}{3}$$

先验概率与后验概率

A　B　C

△　田

$\frac{1}{3}$　$\frac{1}{3}$　$\frac{1}{3}$　先验

① A车，100% C开

② B车，K C开，1-K A开

③ C车，0%

$$P(B车|C开) = \frac{P(C开|B车) * P(B车)}{P(C开)}$$

$$= \frac{}{}$$

$$= P(C开|A车) * P(A车) + P(C开|B车) * P(B车) + 0\%$$

$$= \frac{K \times \frac{1}{3}}{1 \times \frac{1}{3} + K \times \frac{1}{3}}$$

$$= \frac{K}{K+1} \ ? \qquad K = \frac{1}{2}$$

1.1 先验概率与后验概率

A B C

$$P(B|C) = \frac{k}{k+1}$$

$(k<1)$

$$P(A车|C开) = \frac{P(C开|A车) * P(A车)}{P(C开)}$$

$$= \frac{}{P(C|A) \times P(A) + P(C|B) \times P(B) + 0\%}$$

$$= \frac{1 \times \frac{1}{3}}{1 \times \frac{1}{3} + k \times \frac{1}{3}} = \frac{1}{k+1} > \frac{k}{k+1}$$

1.2 极大似然估计(MLE)

极大似然估计：已知某个随机样本满足某种概率分布，但是其中具体的参数不清楚，参数估计就是通过若干次试验，观察其结果，利用结果推出参数的大概值。最大似然估计是建立在这样的思想上：已知某个参数能使这个样本出现的概率最大，我们当然不会再去选择其他小概率的样本，所以干脆就把这个参数作为估计的真实值。

定义：设总体分布为f(x,θ)，x1,x2,...,xn为该总体采用得到的样本。因为x1,x2,...,xn独立同分布，于是，它们的联合密度函数为：

$$L(x_1,x_2,\ldots,x_n;\theta_1,\theta_2,\ldots,\theta_k)=\prod_{i=1}^{n} f(x_i;\theta_1,\theta_2,\ldots,\theta_k)$$

1.2 极大似然估计(MLE)

求最大似然函数估计值的一般步骤：

1）写出似然函数；

2）对似然函数取对数，得到对数似然函数；

3）若对数似然函数可导，求导，解方程组logL(θ1,θ2,...,θk)=∑ni=1f(xi;θ1,θ2,...,θk)，得到驻点；

4）分析驻点是极大值点。

举例：抛硬币

要点总结

要点1

贝叶斯定理与应用

要点2

MLE的步骤与使用

02

熵

2.1 信息与熵

2.2 熵之间的关系

2.1 信息与熵

信息：i(x)=−log(p(x)) 如果说概率p是对确定性的度量，那么信息就是对不确定性的度量。

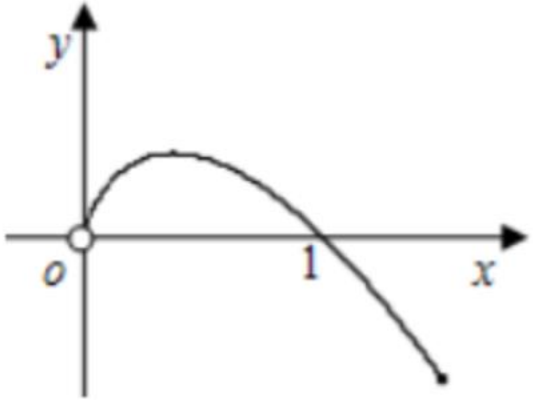
独立事件的信息：如果两个事件X和Y独立，即p(xy)=p(x)p(y) ，假定X和y的信息量分别为i(x)和i(y)，则二者同时发生的信息量应该为i(x^y)=i(x)+i(y)。

熵：是对随机变量平均不确定性的度量。1948年，香农Claude E. Shannon引入信息（熵），将其定义为离散随机事件的出现概率。一个系统越是有序，信息熵就越低；反之，一个系统越是混乱，信息熵就越高。所以说，信息熵可以被认为是系统有序化程度的一个度量。不确定性越大，熵值越大；若随机变量退化成定值，熵为0。熵是自信息的期望。

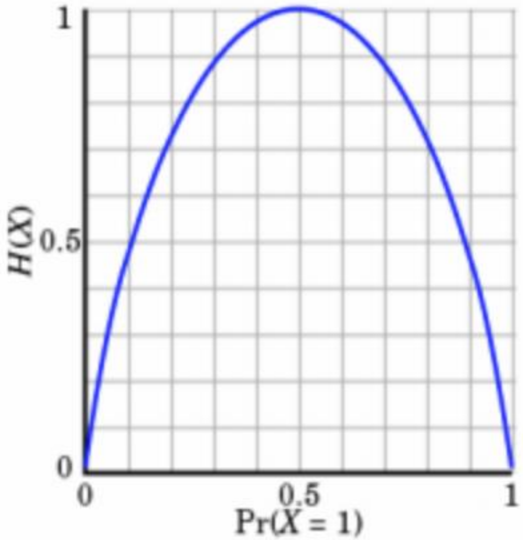
$$H(X) = -\sum_{x \in X} P(x) \log P(x)$$

2.1 信息与熵

熵其实是定义了一个函数（概率分布函数）到一个值（信息熵）的映射。



$-x\log(x)$



$-x\log(x)-(1-x)\log(1-x)$

2.1 信息与熵

1 2

互信息：i(y,x)=i(y)-i(y|x)=log(p(y|x)/p(y))

$$= \log\left(p(yx) / p(y) \times p(x)\right)$$

$$= \log\left(p(x|y) / p(x)\right)$$

$$= i(x) - i(x|y) = i(x,y)$$

平均互信息：决策树中的"信息增益"其实就是平均互信息 I(X,Y)

$$I(X;Y) = \sum_{x\in X, y\in Y} P(x,y)\log\frac{P(x,y)}{P(x)P(y)}$$

2.2 熵之间的关系

联合熵：两个随机变量X，Y的联合分布，可以形成联合熵Joint Entropy，用H(X,Y)表示。不能做误差衡量。

条件熵：在随机变量X发生的前提下，随机变量Y发生所新带来的熵定义为Y的条件熵，用H(Y|X)表示，用来衡量在已知随机变量X的条件下随机变量Y的不确定性。可用来计算交叉熵。H(Y|X)=H(X,Y)-H(X)，表示(X,Y)发生所包含的熵减去X单独发生包含的熵。

平均互信息：I(X;Y) 衡量相似性。

交叉熵：H(T;Y)。衡量两个概率分布的差异性。逻辑回归中的代价函数用到了交叉熵。

相对熵：KL散度，也是衡量两个概率分布的差异性。

Name	Formula	(Dis)similarity	(A)symmetry
Joint Information	$H(T,Y) = -\sum_t \sum_y p(t,y)\log_2 p(t,y)$	Inapplicable	Symmetry
Mutual Information	$I(T,Y) = \sum_t \sum_y p(t,y)\log_2 \frac{p(t,y)}{p(t)p(y)}$	Similarity	Symmetry
Conditional Entropy	$H(Y\|T) = -\sum_t \sum_y p(t,y)\log_2 p(y\|t)$	Dissimilarity	Asymmetry
Cross Entropy	$H(T;Y) = -\sum_z p_t(z)\log_2 p_y(z)$	Dissimilarity	Asymmetry
KL Divergence	$KL(T,Y) = \sum_z p_t(z)\log_2 \frac{p_t(z)}{p_y(z)}$	Dissimilarity	Asymmetry

要点总结

要点1

熵的定义

要点2

各种熵之间的关系

03

最大熵模型

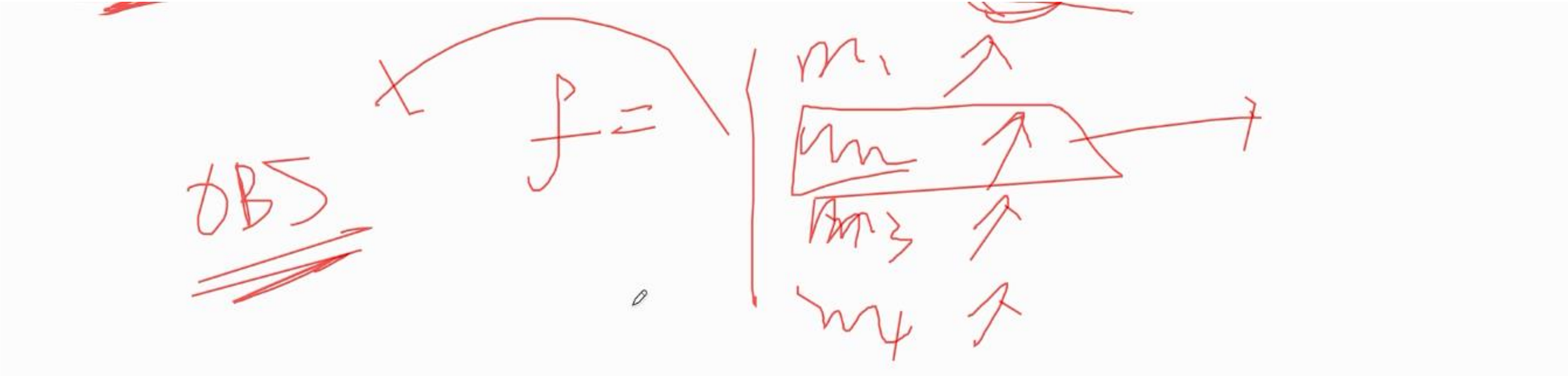
3.1 最大熵原理

3.2 最大熵模型

最大熵原理

最大熵理论：在无外力作用下，事物总是朝着最混乱的方向发展。事物是约束和自由的统一体。事物总是在约束下争取最大的自由权，这其实也是自然界的根本原则。在已知条件下，熵最大的事物，最可能接近它的真实状态。最大熵原理的一个基本假设就是：认为已知的事物是一种约束，未知的条件是均匀分布且没有偏见的。

最大熵原理是统计学的一般原理，也是概率模型学习的一个准则。最大熵原理认为，学习概率模型时，在所有可能的概率模型中，熵最大的模型是最好的模型。



3.1 最大熵原理

最大熵原理进行机器学习：比如用最大条件熵求最大条件概率。

1. 定义条件熵：最大熵的目标是定义一个熵，条件熵实际上就是要找的模型。最大化的条件熵得到的结果是要找到一个条件概率对应的分布，条件概率的分布就是要求的模型，所以，要求的就是条件概率，可以用条件熵定义目标函数。条件熵最大的时候对应的条件概率就是要求的条件概率。公式中x表示特征，y表示标签。

$$H(y|x) = -\sum_{(x,y)\in z} p(y,x)\log p(y|x)$$

2. 模型目的：找到条件熵。公式中P*表示理想概率，公式表示最大化条件熵对应的自变量

$$p^*(y|x) = \arg\max_{p(y|x)\in P} H(y|x)$$

3. 定义特征函数：把其他约束条件写成期望相等的形式。

$$f_i(x,y) \in \{0,1\} \quad i=1,2,...,m$$

4. 约束条件：定义如下

$$\begin{cases} \sum_{y \in Y} p(y|x) = 1 & (1) \\ E(f_i) = \tilde{E}(f_i) \quad i=1,2,...,m & (2) \end{cases}$$

其中，

$$\tilde{E}(f_i) = \sum_{(x,y) \in z} \tilde{p}(x,y) f_i(x,y) = \frac{1}{N} \sum_{(x,y) \in T} f_i(x,y) \quad N = |T|$$

$$E(f_i) = \sum_{(x,y) \in z} p(x,y) f_i(x,y) = \sum_{(x,y) \in z} p(x) p(y|x) f_i(x,y)$$

最大熵模型实例：

(t1) “抓住”：The mother takes her child by the hand.母亲抓住孩子的手。

(t2) “拿走”：Take the book home. 把书拿回家。

(t3) “乘坐”：to take a bus to work. 乘坐公共汽车上班。

(t4) “量”：Take your temperature. 量一量你的体温。

(t5) “装”：The suitcase wouldn’t take another thing. 这个衣箱不能装别的东西了。

(t6) “花费”：It takes a lot of money to buy a house. 买一所房子要花一大笔钱。

(t7) “理解、领会”：How do you take this package? 你怎么理解这段话？

要点总结

要点1

最大熵理论的理解

要点2

最大熵模型的理解

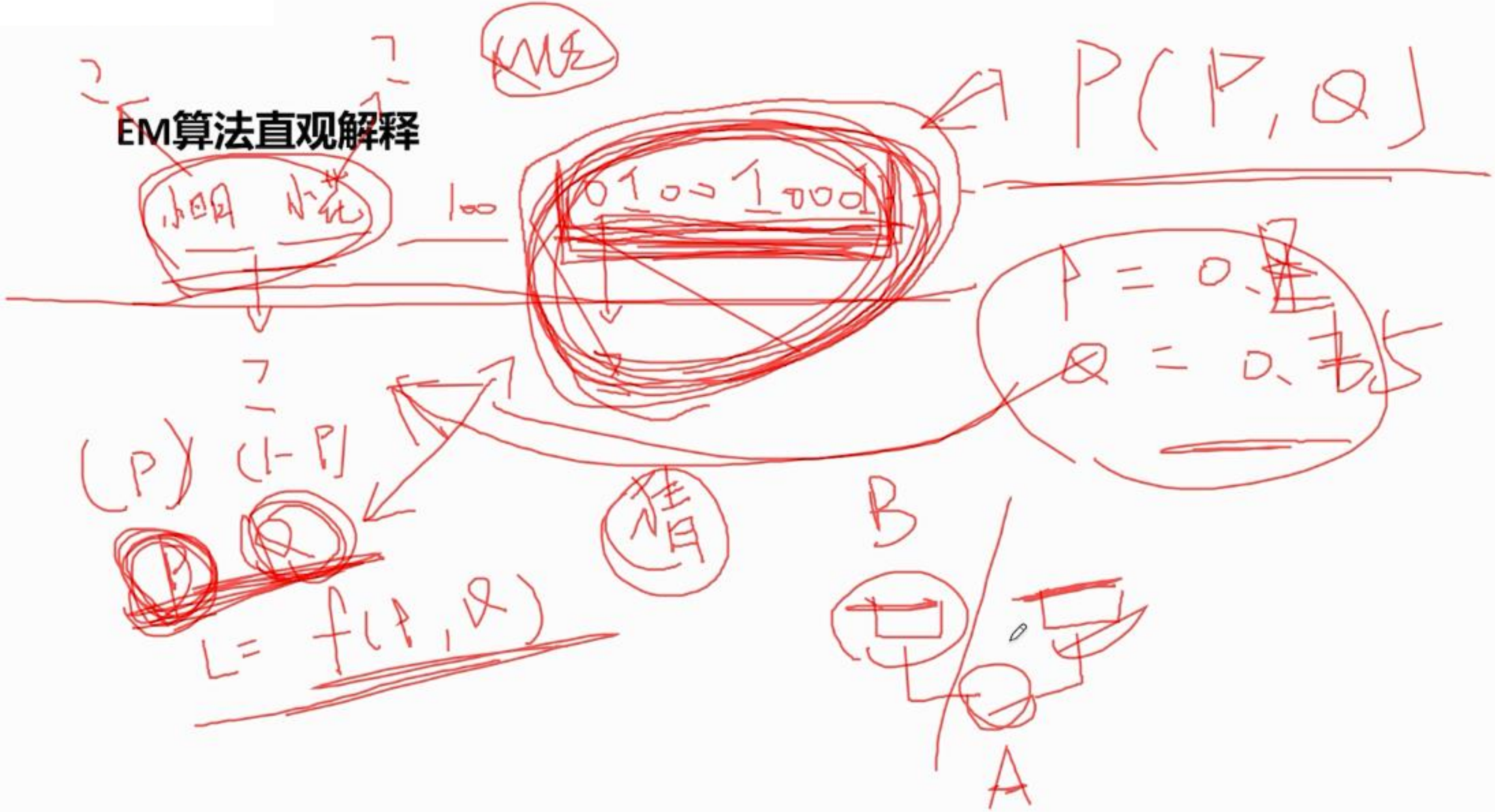
04

EM算法

4.1 EM算法直观解释

4.2 EM算法框架

EM算法直观解释



4.2 EM算法框架

算法整体框架：
（ 推导及细节详见随堂附加课件 ）

Repeat until convergence {

(E-step) For each i, set

$$Q_i(z^{(i)}) := p(z^{(i)}|x^{(i)};\theta).$$

(M-step) Set

$$\theta := \arg\max_{\theta} \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})}$$

}

要点总结

要点1

EM算法理解

要点2

EM算法框架

05

实战案例

5.1 详见随堂附加课件

THANK YOU！

Machine Learning Engineer
机器学习工程师微专业

稀牛学院 網易 NETEASE