Deep Learning 读书会第三次讨论记录（由@极视角小助手整理）

下面为 2016 年 11 月 28 日 Deep Learnin 读书会第三章 **Probability and Information Theory 概率论与信息论**讨论交流笔记。由极视角小助手整理，**九问**和**万元芳**提出话题组织讨论。如有想加入读书会讨论的，请联系小助手（微信：Extreme-Vision）.

讨论话题

引言

这章主要介绍了概率论与信息论的一些基础知识。概率论主要对学习过程中的不确定性进行一个量化，而信息论主要是告诉如何用熵计算有用的信息，深度学习一个比较常用的损失函数就是交叉熵。有别于传统的统计机器学习，我们不需要了解的太深入。如果你已经学过概率论与数理统计，还是建议大家把这章看一下。最主要这章除了基本概念，还有有一些英语名词，符号，会在后面几章经常出现。当然如果有兴趣，想深入了解一下本章节的知识，推荐另外两份资料

PRML 的第一章和第二章 :如果想了解传统的统计机器学习，仔细看一下。否则大致了解即可，至少知道啥事先验概率啥是共轭先验。当然，深度学习已经不会在涉及到这些知识了。
图解信息论(http://colah.github.io/posts/2015-09-Visual-Information/)。这篇文章很形象的用图给大家介绍了信息论的基础知识。不难，而且很清楚。这个人的博客很有价值，以后想了解啥是 LSTM，或者计算图是如何工作的，也可以

找对应的博客看。
*（*读书会成员 **清** 的表述 *）*

话题一.什么是熵，熵代表是什么意义 （九问）

九问
针对熵这个概念，我理解为一个分布面包含信息的度量，这个分布可以使连续的也可以离散的

$$H(\mathrm{x}) = \mathbb{E}_{\mathrm{x}\sim P}[I(x)] = -\mathbb{E}_{\mathrm{x}\sim P}[\log P(x)].$$

这个公式中,P(x)意思是指这个 distribution

安兴乐
信息的分散程度

九问
实际上是对这个分布概率取 Log 然后求期望，一般认为这个 H(X)值越小（负的越大）代表信息越乱，其中这个书中代表 self-information

$$\mathrm{x}\sim P}[\log P(x)].$$

针对离散的概率分布，也就是 x 的取值是离散的时候，这个叠加计算即可，然而当 x 是连续的时候 ,书中称为 differential entropy ,这个计算就需要用到积分？

平
嗯嗯，这个是信息论的内容，这边只是介绍了一下罢

Pascal
期望的定义

九问

期望的定义是指？

Pascal
针对离散的概率分布，也就是 x 的取值是离散的时候，这个叠加计算即可，然而当 x 是连续的时候 ,书中称为 differential entropy ,这个计算就需要用到积分。这个就是期望定义呀

James Liu
因为本质上就是期望

陆婷婷
这里的自信息应该怎么理解呢？（能否理解为某一个信号取某个具体值时，该值能够提供的信息？）

平
信息量是 - log Pi，信息量的期望就是信息熵

印第安老斑鸠
期望有什么意义呢

九问

$$E(X) = \int_{-\infty}^{+\infty} x f(x) dx$$

一个分布的期望代表着，针对一个 x，最可能的取值？或者平均取值

Clear figure
某种程度上的加权平均吧

平
应该是加权平均

AG-Group 元芳
熵的含义，其实后面有一个图讲的很明白，举了个栗子，用二项分布

平
感觉信息论里面讲得很清楚了，包括互信息量，联合熵这些都有

九问
这个熵比较好理解，应用的地方也比较多，主要是决策树这些

平
InfoGAN 就用了互信息量的概念

徐胜伟
信息的混乱程度

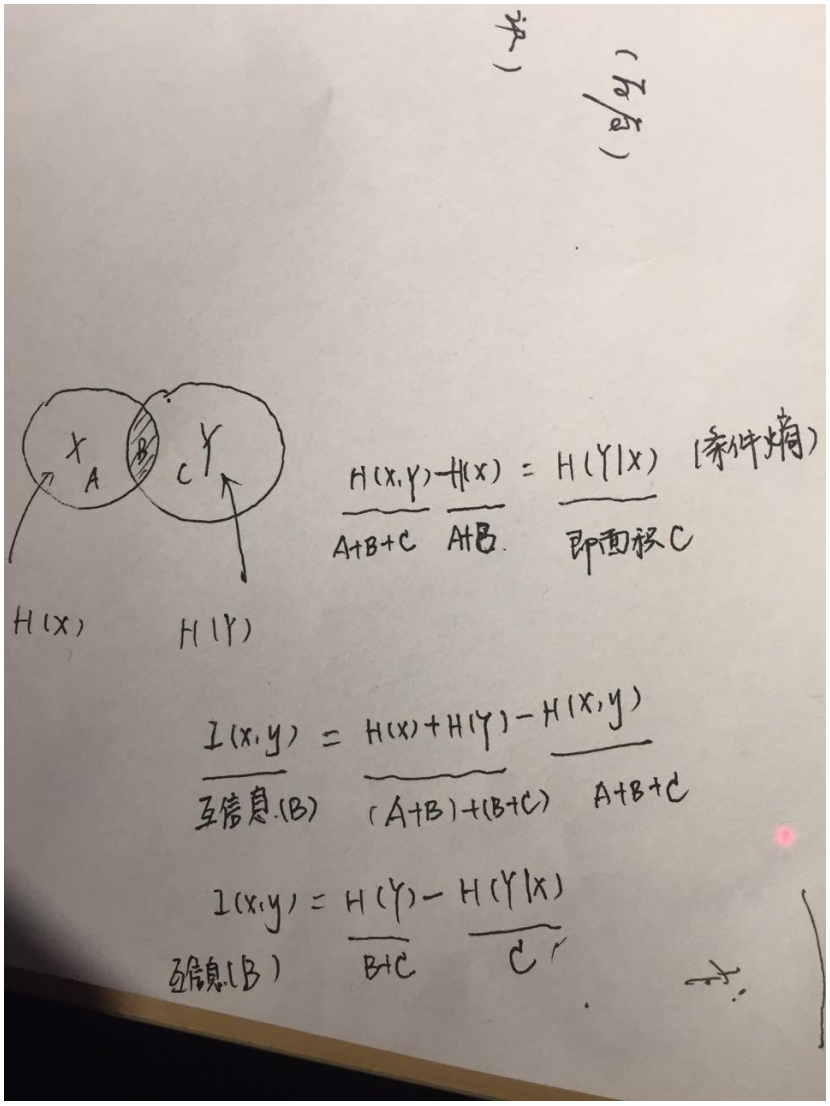
yc
我记得有句话 达到 1/e 确定性的信息量

九问
这个句话什么意思？

果冻儿
用韦恩图很好解释的，信息熵，条件熵，互信息，三者关系一目了然

九问
有例子嘛

果冻儿



话题二.熵与 KL 散度有什么关系（九问）

九问
嗯，我们可以讨论下一个问题，主要是关于 KL 散度的问题

李彬@机器学习
kl 散度就是互信息吧

九问
对，是这么理解。我不是很明白为什么是不对称的。

$$D_{\mathrm{KL}}(P\|Q) = \mathbb{E}_{\mathrm{x}\sim P}\left[\log\frac{P(x)}{Q(x)}\right] = \mathbb{E}_{\mathrm{x}\sim P}\left[\log P(x) - \log Q(x)\right].$$

清
不对称是因为公式

九问
要知道物理意义呀

清
把 pq 换个位置结果就不一样

九问
这是书中公式 ,是否可以这么理解 ,D(P||Q)代表 Q 中包含的 P 中的信息量？

清
应该是距离，两个分布之间的距离

九问
距离一般是对称的

清
其实是有对称的形式

AG-Group 元芳
期望的极值计算是可以互换的啊

清
把公式改一下。算两个 pq 和 qp kl 散度然后加起来，就对称了

Clear figure
因为是相对距离不是绝对距离？

清

距离只是一个形象的说法

话题三.KL 散度代表什么意义，为什么是不对称的（九问）

九问
这个我看到过，对称的公式。KL 散度的应用意义有哪些呢？我比较孤陋寡闻。

bf Q
相对熵（KL 散度）可以用于比较文本的相似度，先统计出词的频率，然后计算 KL 散度就行了。另外，在多指标系统评估中，指标权重分配是一个重点和难点，通过相对熵可以处理

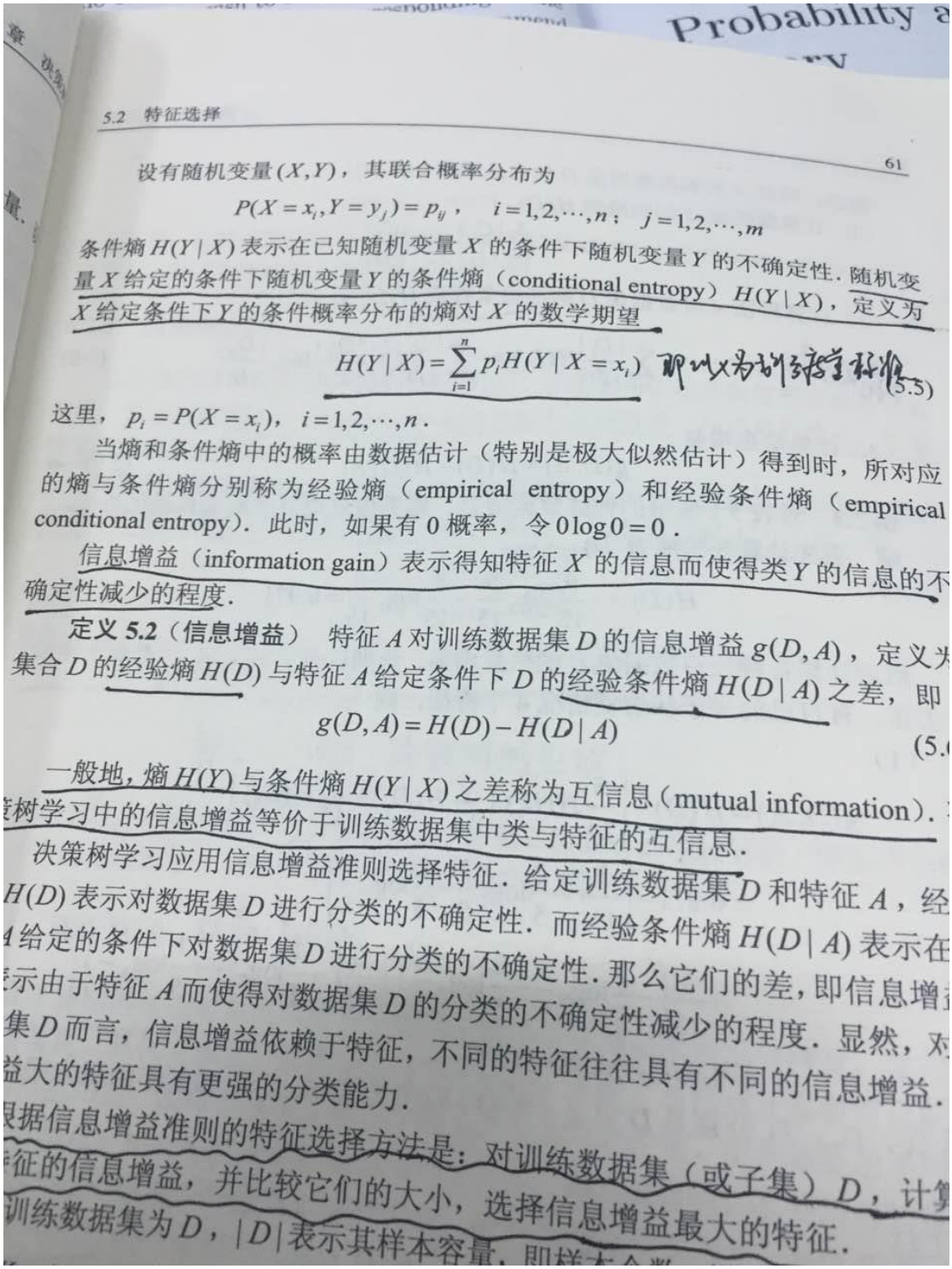
AG-Group 元芳
那我在做的时候是应该 pq 取还是 qp 取呢？

yc
In applications, P typically represents the "true" distribution of data, observations, or a precisely calculated theoretical distribution, while Q typically represents a theory.

Clear figure
em 算法里也用到了

清
对 prml 9.4

李彬@机器学习
我在李航的书里面找到了互信息的解释



5.2 特征选择 61

设有随机变量 (X,Y)，其联合概率分布为

$$P(X=x_i,Y=y_j)=p_{ij},\quad i=1,2,\cdots,n;\quad j=1,2,\cdots,m$$

条件熵 $H(Y|X)$ 表示在已知随机变量 X 的条件下随机变量 Y 的不确定性. 随机变量 X 给定的条件下随机变量 Y 的条件熵（conditional entropy）$H(Y|X)$，定义为 X 给定条件下 Y 的条件概率分布的熵对 X 的数学期望

$$H(Y|X)=\sum_{i=1}^{n}p_iH(Y|X=x_i) \tag{5.5}$$

这里，$p_i=P(X=x_i)$，$i=1,2,\cdots,n$.

当熵和条件熵中的概率由数据估计（特别是极大似然估计）得到时，所对应的熵与条件熵分别称为经验熵（empirical entropy）和经验条件熵（empirical conditional entropy）. 此时，如果有 0 概率，令 $0\log 0=0$.

信息增益（information gain）表示得知特征 X 的信息而使得类 Y 的信息的不确定性减少的程度.

定义 5.2（信息增益） 特征 A 对训练数据集 D 的信息增益 $g(D,A)$，定义为集合 D 的经验熵 $H(D)$ 与特征 A 给定条件下 D 的经验条件熵 $H(D|A)$ 之差，即

$$g(D,A)=H(D)-H(D|A) \tag{5.}$$

一般地，熵 $H(Y)$ 与条件熵 $H(Y|X)$ 之差称为互信息（mutual information）. 决策树学习中的信息增益等价于训练数据集中类与特征的互信息.

决策树学习应用信息增益准则选择特征. 给定训练数据集 D 和特征 A，经验熵 $H(D)$ 表示对数据集 D 进行分类的不确定性. 而经验条件熵 $H(D|A)$ 表示在给定的条件下对数据集 D 进行分类的不确定性. 那么它们的差，即信息增益，表示由于特征 A 而使得对数据集 D 的分类的不确定性减少的程度. 显然，对于数据集 D 而言，信息增益依赖于特征，不同的特征往往具有不同的信息增益. 信息增益大的特征具有更强的分类能力.

根据信息增益准则的特征选择方法是：对训练数据集（或子集）D，计算各特征的信息增益，并比较它们的大小，选择信息增益最大的特征.

设训练数据集为 D，$|D|$ 表示其样本容量，即样本个数.

陆婷婷

看到了一个关于 KL 非对称的解释，但不知道这个解释是否合理

Kullback-Leibler Divergence is asymmetric because it measures the departure of the distribution p from q. In this case p is a candidate model and q is viewed as the "truth." A symmetric measure in this case wouldn't make sense at all.

李彬@机器学习

可不可以理解为给定条件 A，X 的不确定性程度减小的多少？

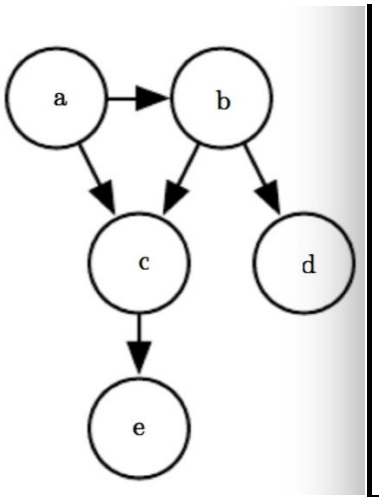
Clear figure

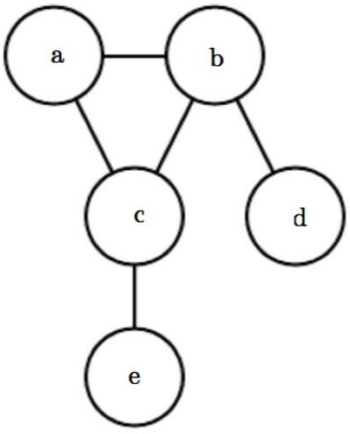
就是这意思

话题四. 有向概率图模型和无向概率图模型的区别（九问）

九问

下面我还有一个问题，这个问题我也不太懂，欢迎大家讨论哈
Structured Probabilistic Models 这一个章节里面主要是大概介绍了概率图模型，其中提到了有向图模型，以及无向图模型，有人理解他们之间真正的区别以及意义么





一个是有向图，一个是无向图

$$p(\mathrm{a},\mathrm{b},\mathrm{c},\mathrm{d},\mathrm{e}) = p(\mathrm{a})p(\mathrm{b} \mid \mathrm{a})p(\mathrm{c} \mid \mathrm{a},\mathrm{b})p(\mathrm{d} \mid \mathrm{b})p(\mathrm{e} \mid \mathrm{c}).$$

$$p(\mathrm{a},\mathrm{b},\mathrm{c},\mathrm{d},\mathrm{e}) = \frac{1}{Z}\phi^{(1)}(\mathrm{a},\mathrm{b},\mathrm{c})\phi^{(2)}(\mathrm{b},\mathrm{d})\phi^{(3)}(\mathrm{c},\mathrm{e}).$$

清

这是概率图里面的内容啊

九问

书中第三节最后就是这个内容

清

第一个公式最主要展示一个依赖关系，第二个公式 z 是用来归一化的。

李彬@机器学习

看起来像决策树

九问

第二个公式我理解是，无向图中所有相连的线代表一个分布

清

嗯 那有个名字叫势函数

九问

back propagation 里面会用到= =

AG-GROUP 元芳

我的理解，q learning 学习完后每一条线上都会有一个 reward 可以类比这里的概率，有向图是我从房间 2 走到 5，最后的总分数可以用链式法则计算（这里的总概率），无向图没有起点和终点，但是我们有每一个 action 的奖励，也可以算各种路径的分数，不知道这样理解对不对

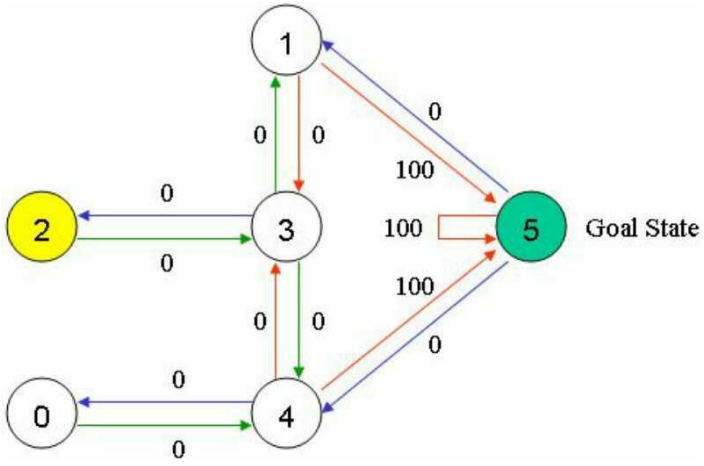
清

是吗？因为概率图课 后向好像是在计算图里面弄的，感觉有点差别吧

yc

感觉很简单啊，就链式法则，然后独立的拿掉

AG-GROUP 元芳



这个是 Q 里的图. 我们确实可以根据这个图用链式法则算出 P（2|5）. 可是无向图又怎么理解呢？

九问

对，是把独立分布的拿掉。无向图就是把独立的分布单独拿出来

yc

phi(x,y,z,...) 表示是两两相关

九问
yep

$$p(\mathrm{a},\mathrm{b},\mathrm{c},\mathrm{d},\mathrm{e}) = p(\mathrm{a})p(\mathrm{b}\mid\mathrm{a})p(\mathrm{c}\mid\mathrm{a},\mathrm{b})p(\mathrm{d}\mid\mathrm{b})p(\mathrm{e}\mid\mathrm{c}).$$

这里有向图，为啥后面的都是 p。按理说后面的分布应该都不一样啊

缨宁
有一种思路是，无向图联合概率分布为什么要是所有最大团的势函数的累乘，要联系其局部独立性看，一个最大团内所有的节点都不可能是条件独立的，必须表示成联合分布

清
大家可以看一下 prml 中文版的第八章，专门讲的概率图，之前那个问题可以看一下第九章前几节

话题五. 概率密度函数的数值意义是什么？（元芳）

AG-GROUP 元芳

我这里的第一个问题是 PDF 的数值意义是什么？尤其是当 PDF > 1 的时候

清
有可能大于 1 吗

yc
有的啊

清
概率密度不是有两个性质，一个是加起来等于 1，一个是都大于 0 吗

yc
那个是面积

AG-GROUP 元芳

书上说因为连续型分布，不能直接求出 PMF，所以引入 PDF 来积分得到 PMF, 可是 PDF 的具体意义代表什么？他只是为了计算 PMF 而存在的么？

清
pdf 似乎是连续的 看一个点没意义.pmf 好像是针对离散. pdf 算概率好像要积分才有意义

AG-GROUP 元芳

单位脉冲函数这样的特例积分也是 1, 可是有很多地方概率 > 1,是 PDF > 1，不是 PMF > 1，怎么解释？

James Liu
PDF 的值不是概率值

AG-GROUP 元芳

说错了上面

清
在信号 里单位脉冲信号不是概率

yc
单位脉冲是 dirac 函数

WWN
对，在某些阶跃处值完全有可能大于 1

James Liu
因为连续分布时没有说 x 取 xx 值的概念,只有 x 在某区间的概率

yc
它在 0 点的值不是 1, 是无穷

清
不过信号课基本还给老师了,只记得那个是用来计算系统冲击响应的

AG-GROUP 元芳

如果我们的 PDF 函数出现 > 1，怎么解释

清
如果一定要个解释,那就是 pdf 看一点没意义,看积分才有意义

James Liu
这个值是密度值,如果一直纠结密度值是否> 1,那可能就需要仔细想想概率的定义

清
因为他是连续的

James Liu
楼上+1 概率是积分. 是密度对区间的积分

AG-GROUP 元芳

这样说哈，我们计算定积分，也就是概率处在 A 和 B 之间的概率，如果 AB 足够接近，是否可以理解为 PDF 上的一个点？

James Liu
取个极限我觉得可以啊

AG-GROUP 元芳

微元的思想，并不是没有意义. 所以我一直不明白他这里的数值大小代表什么

James Liu
这样定义最好了,对着这个定义再看脉冲响应应该就能明白了

AG-GROUP 元芳

哪个定义啊

清

我忽然又不是很懂

James Liu
这样说哈，我们计算定积分，也就是概率处在 A 和 B 之间的概率，如果 AB 足够接近，是否可以理解为 PDF 上的一个点？

果冻儿
密度和概率是不一样的

清
嗯，密度积分才是概率

pascal
对于连续函数求积分 计算某点概率是无意义的

AG-GROUP 元芳

可是微元也是积分啊

清
积分之后是不可能大于 1 的，密度可能大于 1 积分不可能大于 1

James Liu
你看 脉冲响应的 p(0)=\integral _{\delta x}^{\delta x} pdf(x)=1，虽然 pdf(0)=\inf, 但是 p(0)=1

果冻儿
是的

清
这样就清楚了

话题六.We can thus think of the normal distribution as being the one that inserts the least amount of prior knowledge into a model. 原文中这句话怎么理解（元芳）

AG-GROUP 元芳

是因为绝大多数的自然事件都符合正态分布么？所以说他含有最多的先验知识？在神经网络里确实很多时候参数初始化（Weight&Bias）用的是这个，因为含有最多的先验，这里的先验指的是熵么？

清

好像熵最大。高斯分布的熵最大

James Liu

搞贝叶斯推断的时候需要设计先验分布，但不一定都是高斯。。

清

最少的先验吧. 那是为了共轭

James Liu

有不包含先验先验的 prior，好多形式，不一定是高斯

AG-GROUP 元芳

就是说我们能否证明参数初始化时候高斯分布效果是否好于别的分布？读到这里时候感觉完全不懂，这里的插入先验知识怎么理解？

清

大家研究一下 prml 第二章

纸鸢

看到一个解释说，在所有具有相同均值和方差的分布中，高斯分布的熵最大，所以，它具有的不确定就越大，也就是包含的先验最小，不知对不？

AG-GROUP 元芳

有道理。熵最大表示混乱度最大，也就是确定性结论越少，先验知识越少。这样理解对不？

缨宁
@纸鸢 是对的。先验表示的是对已知事件的描述，然而在对某些模型描述的实际情况不清楚，即难以给出有用先验时，可以选用高斯分布，以免引入强指导性但不正确的先验。

yc
能不能通俗的解释一下先验概率和后验概率啊

AG-GROUP 元芳

事情还没有发生,要求这件事情发生的可能性的大小,是先验概率. 事情已经发生,要求这件事情发生的原因是由某个因素引起的可能性的大小,是后验概率.

Jame Liu
Normal distributions are a sensible choice for many applications. In the absence of prior knowledge about what form a distribution over the real numbers should take, the normal distribution is a good default choice for two major reasons.
soga,在说参数初始化。。

清
其实还有一个初始化方式，lecun 提的
Uniform initialization scaled by the square root of the number of inputs

Jame Liu
其实有好多初始化方式，还有 xavier 什么的，有个神奇的根号 6

清
对，这个我用了。效果最好

AG-GROUP 元芳

对啊，但是并不知道怎么选，就是选的依据，所以每次都高斯分布随机。。。

yc
那怎么和书上的概率和条件概率对上呢？ 最大熵的连续分布：已知区间= = > 均

匀分布已知均值==>指数分布已知均值和标准差（方差）==>正态分布

Jame Liu
magic number. 我的理解是，如果你对这个网络的结构没什么理解，那就用 gaussian 吧，反正你也没有 prior。

纸鸢
@yc 这个最大熵的问题哪里有讲呢？怎么理解呢？

yc
http://www.newsmth.net/nForum/#!article/AI/3059
用拉格朗日数乘可以证明

纸鸢
还有个疑问，比如网络初始化时，利用了高斯分布，因为认为此时的高斯分布熵最大，那这个时候难道是默认了均值和方差是已知的吗？

Jame Liu
最原始的是均值 0 方差 0.01. 后续改进的方案中 方差会根据网络的一些性质进行计算。但总体均值是需要为 0 的

纸鸢
@James Liu 长见识了！之前确实不知道网络的权值参数具有这种特征

Jame Liu
方差你可以考虑一下，方差越大这个网络参数可能的取值范围就越大，打到一定程度之后会不会导致梯度爆炸呢？均值如果为+100，会出现什么情况呢~

yc
是不是这就是为什么要 batch norm

纸鸢
那如果网络在学习过程中，出现大均值或者大方差的情况，该怎么处理呢？实际中，会考虑网络的权值的均值和方差吗？

Jame Liu

进一步的,为什么要把初始化参数限制在一个比较小的范围内呢？同理是否需要把输入变量进行归一化呢？
batch norm 其实是另一个问题了。。。他解决的是神经元分布在训练过程中乱变的问题（个人理解 求大佬指正）
大均值/大方差是在初始化的时候你决定的，可以做个实验嘛
参数的数值过大，会不会导致过拟合呢~

纸鸢
这个问题有点复杂呀.....

Jame Liu
其实主要问题是 为什么过大的初始化不好？我倾向于认为过大的初始化会导致过拟合，见 prml 某章。对神经网络来看会引起梯度爆炸，使优化过程不稳定

缨宁
由于网络训练一般有个正则项，控制网络的稀疏性，也就是网络的简洁性。网络的权值基本都会靠近小于 1 来做收敛。以 sigmoid 函数为例，当输入 x 越来越大的时候，函数越来越平滑，梯度也越来越小，训练会越来越慢。从这两个角度看，初始化都不宜过大

纸鸢
假如，初始化的时候，网络权值取值都很小；可能会出现学习过程中，这些本来初始化很小的参数又变大（均值和方差都变大）的可能吗？实际的训练过程中，需要对均值和方差进行实时计算分析吗？

Jame Liu
你们实际训练的时候用过正则项嘛

清
Dropout。没有对参数进行过正则

Jame Liu
对 dropout 是正则项
l2norm 在神经网络训练里不太常见

缨宁

是的，dropout 起了正则的作用
@纸鸢 激活函数一般都是 s 型，你输入再大也会被激活函数控制在一定范围内。比如输入 4 已经可以让激活函数产生正向输出了，没必要输入 1000 了，到了很大的值，神经元反而过饱和了，不好调整了。还有你问可不可以用均匀分布初始化，答案是不可以，失去随机性，网络一开始就注定不会收敛。

盛夏的午后~
激活函数用双曲正切函数和 s 型有什么差距吗

清
均匀分布似乎可以收敛 只不过效果不是很好 如果是都初始化为一个数 是不能收敛的@缨宁

缨宁
@清 ，你说的对，我脑补成一个数了，囧。有正有负的小数字就行

badrobot
relu 激活函数有什么特殊的地方啊

唧唧歪歪
relu 函数收敛快,不像 sigmoid，x 越大梯度变化越小

缨宁
1.计算代价小（两段线性）2.relu 导数>0 时为 1，sigmoid 导数最值为 1/4,减轻梯度消失 3.稀疏的激活性

徐金龙
有比 relu 更好的激活函数吗

daiwk
近年来应该都是 relu 吧, tanh 之后

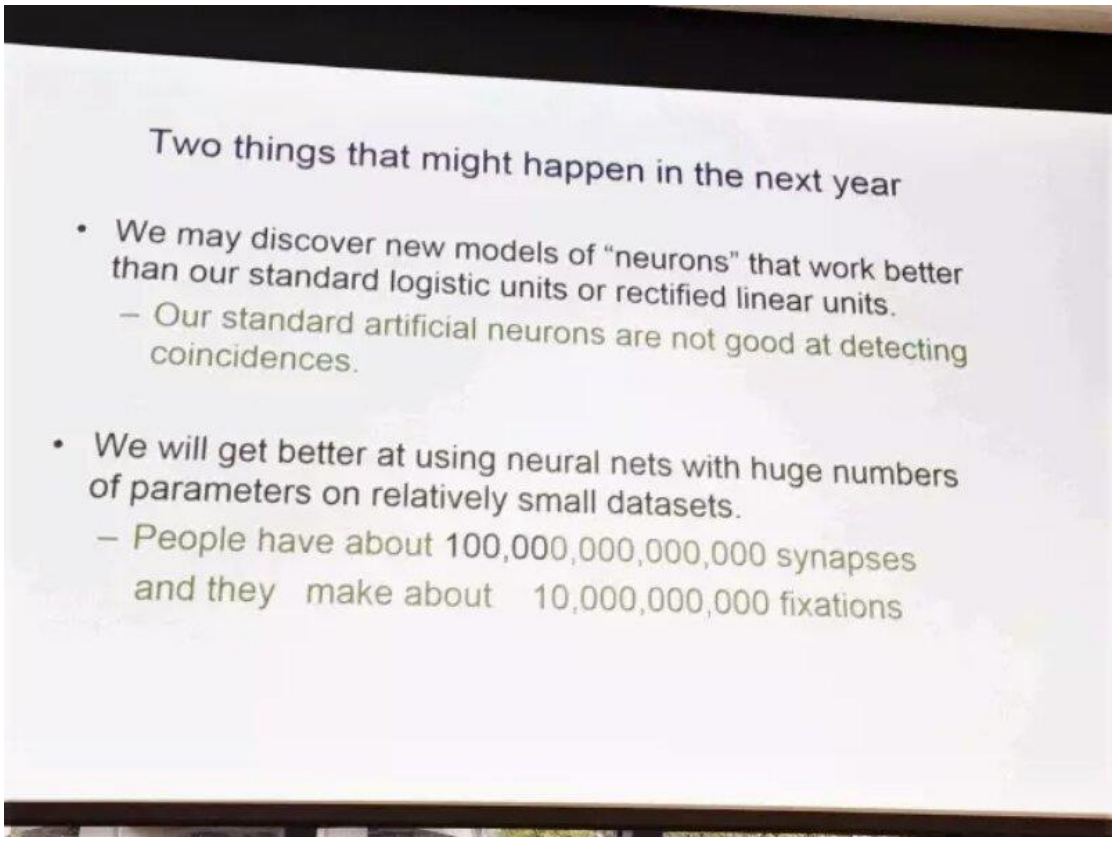
AG-GROUP 元芳

有一个函数叫 softplus，和 reLu 类似，也可以用。

Jame Liu

今年 ICML 有一篇 noisy activation functions。目前还没多少人 follow。

缨宁



hinton 在今年十月的 ppt，可见他还没找到更好的

写在最后

非常感谢此次进行讨论交流的朋友们 ,在讨论的过程中不可避免遇到了更多的问题，这就引导我们更深入得去思考，去研究，从而对书籍有一个更本质的理解，也更好得应用于自己的学习和工作.谢谢本次讨论会参与的每一个人。

#广告时间#

免费线上技术分享，视觉前沿资讯关注请关注极市平台公众号。