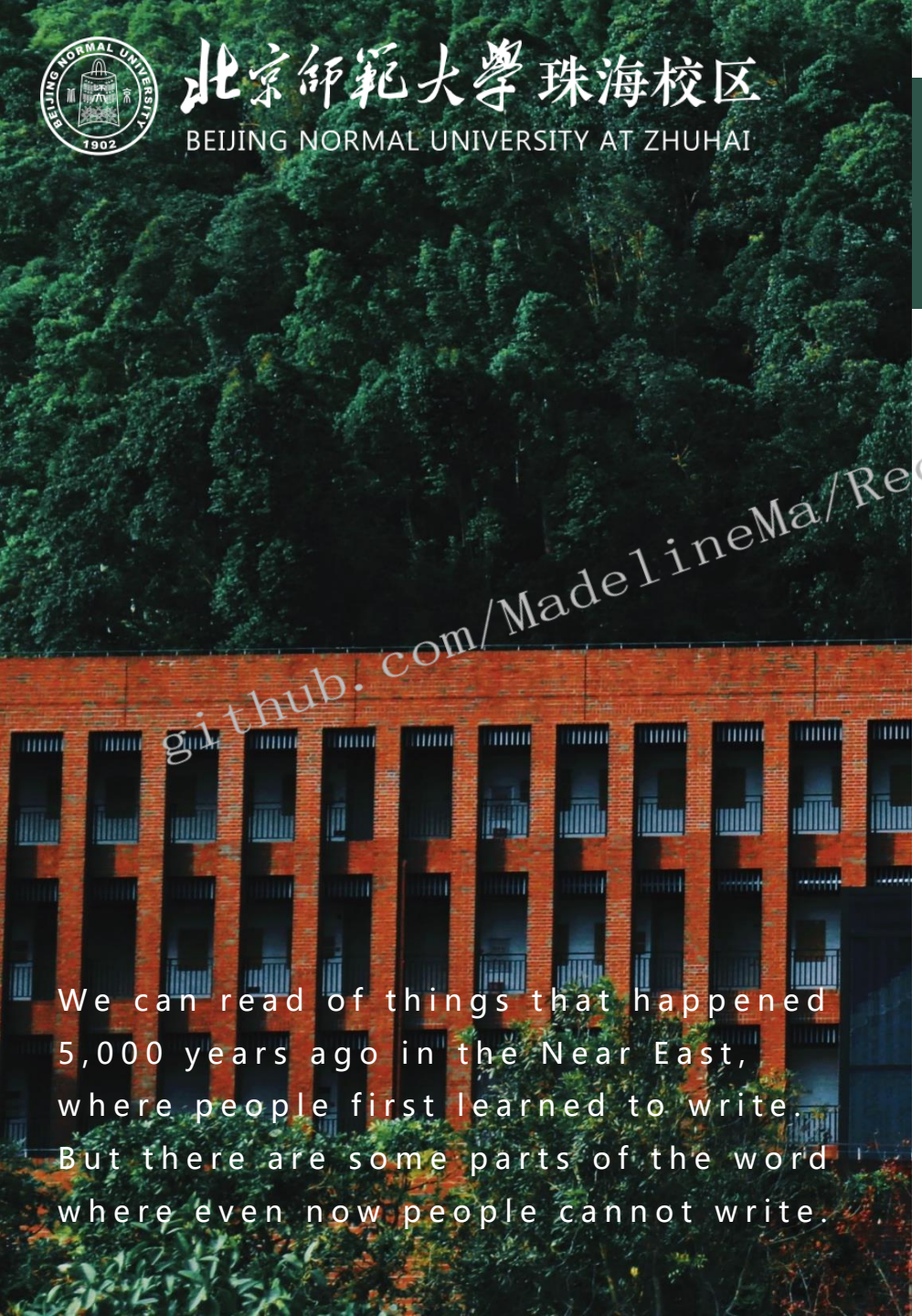

北京師範大學珠海校区
BEIJING NORMAL UNIVERSITY AT ZHUHAI
前深度学习时代
github.com/MadelineMa/Recommender-System
github.com/MadelineMa/Recommender-System
马静



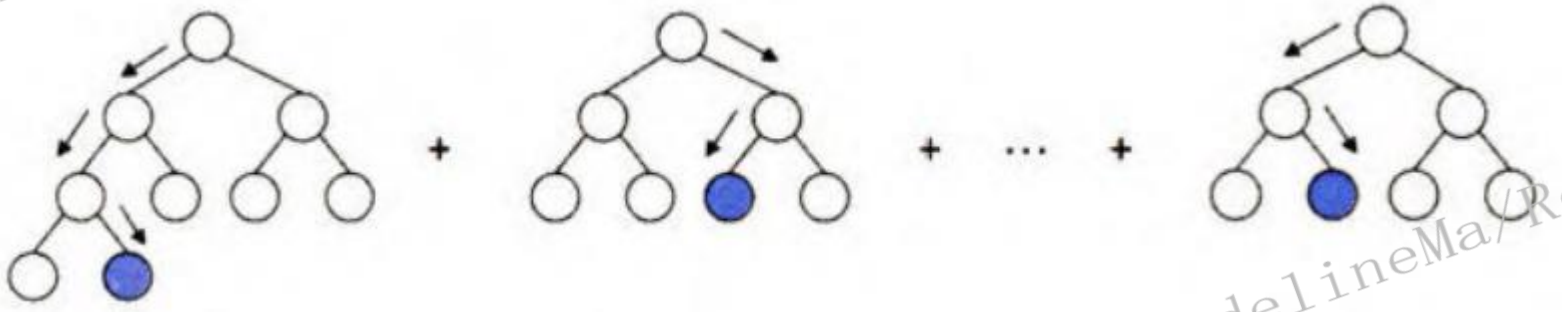
CONTENT

北京師範大學珠海校区
BEIJING NORMAL UNIVERSITY AT ZHUHAI

01GBDT

GBDT(Gradient Boosting Decision Tree)的基本结构是决策树组成的树林

- 求解方法
- 一组弱学习器的集成学习
- if-then 规则的集合



$$D(x) = d_{\text{tree }1}(x) + d_{\text{tree }2}(x) + \cdots$$

GBDT(Gradient Boosting Decision Tree)的基本结构是决策树组成的树林

求解方法

一组弱学习器的集成学习

if-then 规则的集合

面试重点

1. GBDT通过逐一生成决策子树的方式生成整个树林，生成新子树的过程是利用样本标签值与当前树林预测值之间的残差，构建新的子树.

GBDT期望的是构建第i棵子树，使当前树林的预测结果D(x)与第i棵子树的预测结果之和，能进一步逼近理论上的拟合函数，即

$$D(x) + d_{\text{tree}\,4}(x) = f(x)$$

$$R(x) = f(x) - D(x)$$

梯度上升拟合残差

GBDT(Gradient Boosting Decision Tree)的基本结构是决策树组成的树林

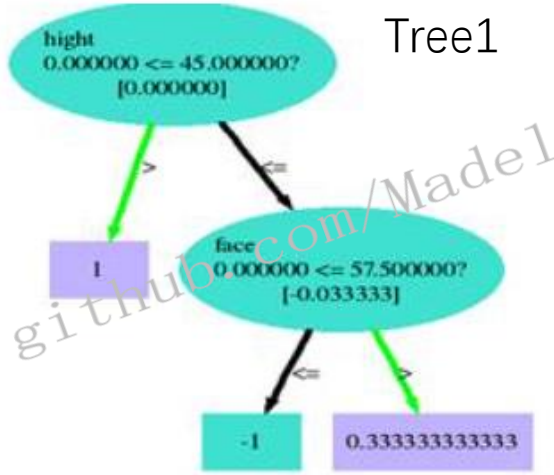
- 求解方法
- 一组弱学习器的集成学习
- if-then 规则的集合

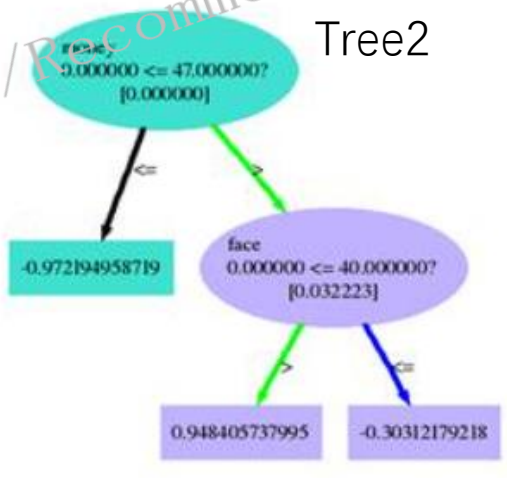
面试重点

2. 每棵树生成的过程是一棵标准的回归树生成过程.

虽然名字叫决策树，但是子树是Cart回归树，因为需要向残差值逼近. 回归树中每个节点的分裂是一个自然的特征选择的过程.

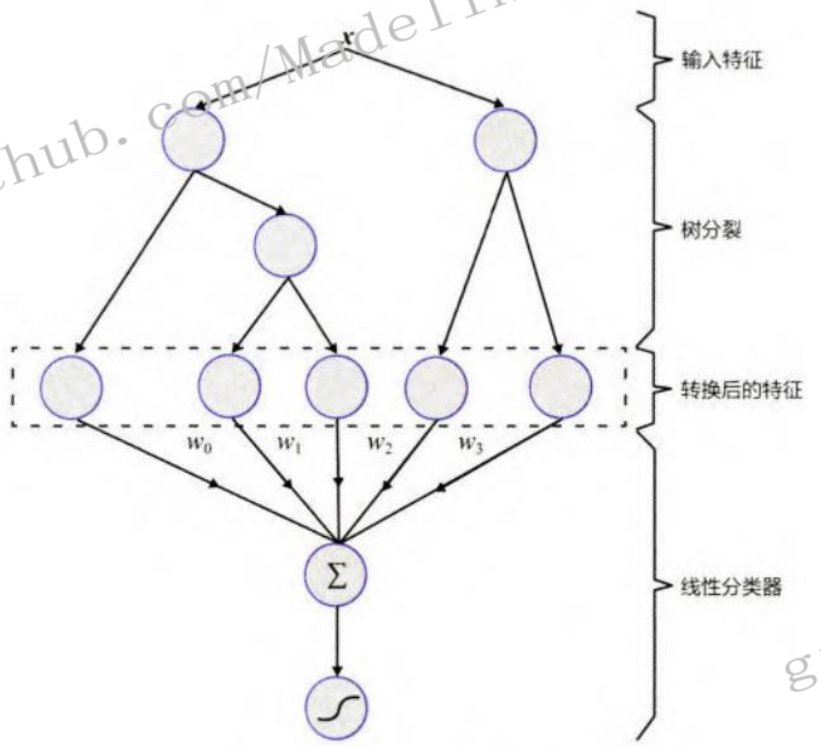
#id	label	hight	money	face
_0	1	20	80	100
_1	1	60	90	25
_2	1	3	95	95
_3	1	66	95	60
_4	0	30	95	25
_5	0	20	12	55
_6	0	15	14	99
_7	0	10	99	2





02
GBDT+LR

Motivation：FM只能做二阶交叉，2014年Facebook提出该方法解决组合爆炸和计算复杂度过高的问题.



GBDT做特征工程

↕ 独立，无梯度回传需求

LR做CTR预估

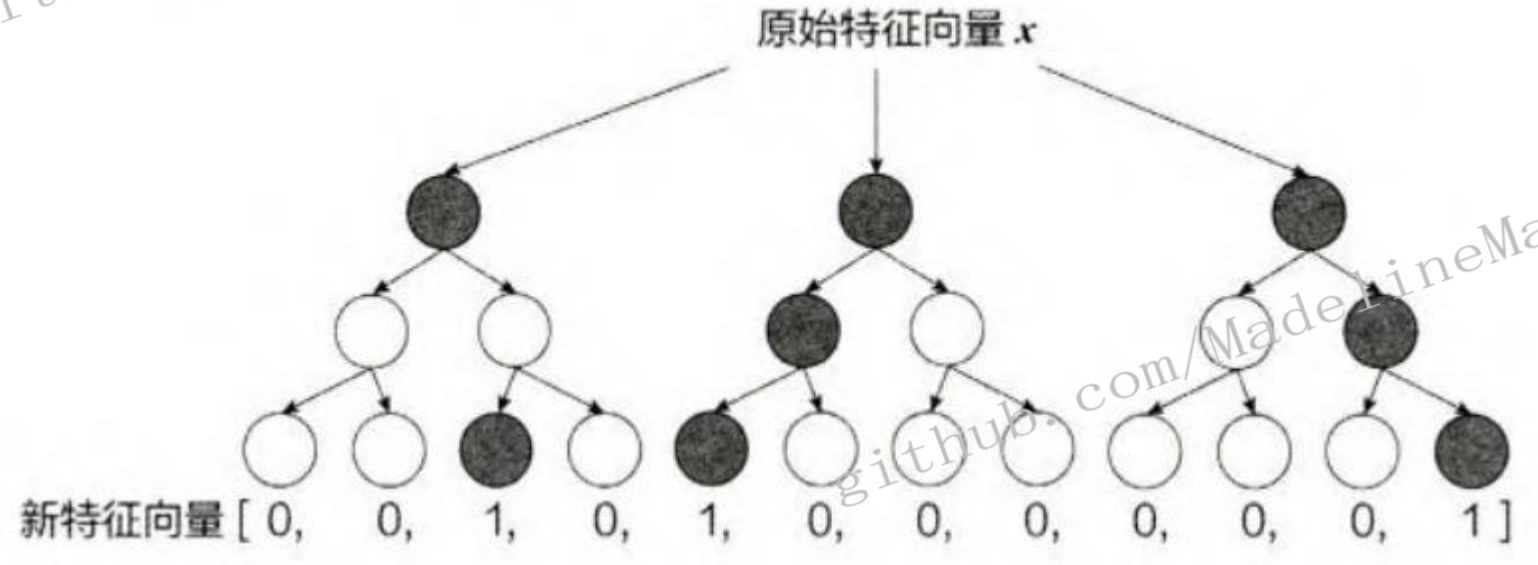
实战体感：

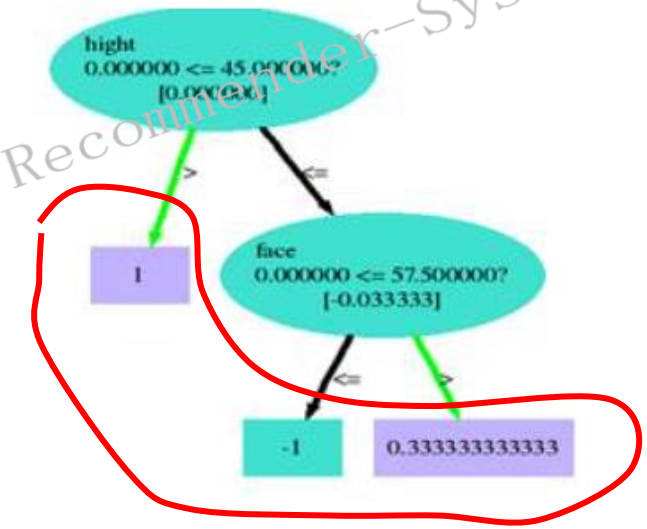
- 虽然该方法可以使效果提升，但是特征工程需要人工定期更新，否则效果会有折扣.
- 增加线上部署复杂度.

GBDT进行特征转换工程

- 利用训练集训练好GBDT模型，完成从原始特征向量到新的离散型特征向量的转化.

一个训练样本在输人GBDT的某一子树后，会根据每个节点的规则最终落人某一叶子节点，把该叶子节点置为1，其他叶子节点置为0,所有叶子节点组成的向量即形成了该棵树的特征向量，把GBDT所有子树的特征向量连接起来，即形成了后续LR模型输人的离散型特征向量





优势:

- 决策树的深度决定了特征交叉的阶数。如果决策树的深度为4,则通过3次节点分裂，最终的叶节点实际上是进行三阶特征组合后的结果，如此强的特征组合能力显然是FM系的模型不具备的.
- 推进了特征工程模型化.
- 深度学习模型通过各类网络结构、Embedding层等方法完成特征工程的自动化，都是GBDT+LR开启的特征工程模型化这一趋势的延续.

劣势:

- GBDT容易产生过拟合
- GBDT的特征转换方式实际上丢失了大量特征的数值信息
- 增加工程量和部署工作

北京師範大學珠海校区
BEIJING NORMAL UNIVERSITY AT ZHUHAI

03 LS-PLM

We can read of things that happened 5,000 years ago in the Near East, where people first learned to write.

LS-PLM

- 阿里巴巴曾经的主流推荐模型“大规模分段线性型(Large Scale Piece-wise Linear Model)
- 在2017年才被阿里巴巴公之于众，但其实早在2012年，它就是阿里巴巴主流的推荐模型，并在深度学习模型提出之前长时间应用于阿里巴巴的各类广告场景.
- LS-PLM的结构与三层神经网络极其相似，在深度学习来临的前夜，可以将它看作推荐系统领域连接两个时代的节点.
- 又被称为MLR(Mixed Logistic Regression,混合逻辑回归）模型，现在被称作XFTRL.

- 分而治之的思路，先对样本进行分片，再在样本分片中应用逻辑回归进行CTR预估.

如果CTR模型要预估的是女性受众点击女装广告的CTR那么显然，我们不希望把男性用户点击数码类产品的样本数据也考虑进来，因为这样的样本不仅与女性购买女装的广告场景毫无相关性，甚至会在模型训练过程中扰乱相关特征的权重.

- 先对全量样本进行聚类，再对每个分类施以逻辑回归模型进行CTR预估.

超参数“分片数”m可以较好地平衡模型的拟合与推广能力

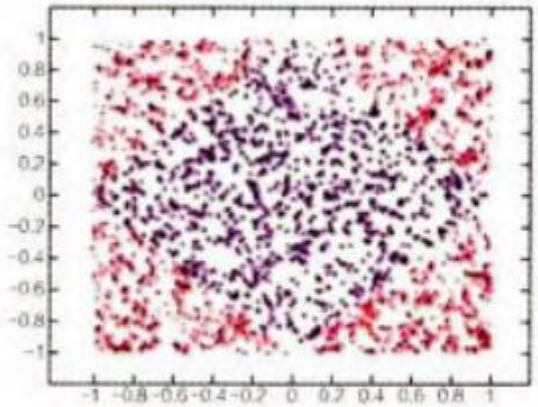
$$f(x)=\sum_{i=1}^{m}\pi_i(x)\cdot\eta_i(x)=\sum_{i=1}^{m}\frac{e^{\mu_i\cdot x}}{\sum_{j=1}^{m}e^{\mu_j\cdot x}}\cdot\frac{1}{1+e^{-w_i\cdot x}}$$

π 采用了 softmax 函数对样本进行多分类

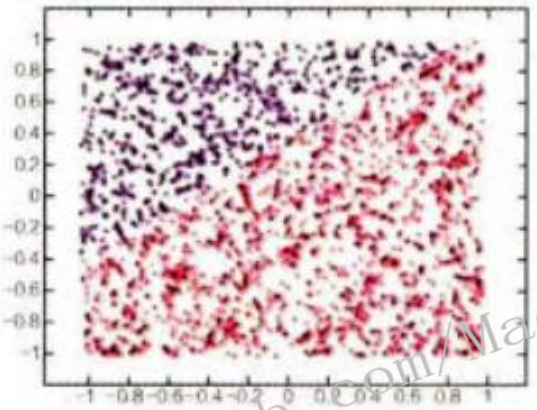
这部分样本的话语权

LS-PLM

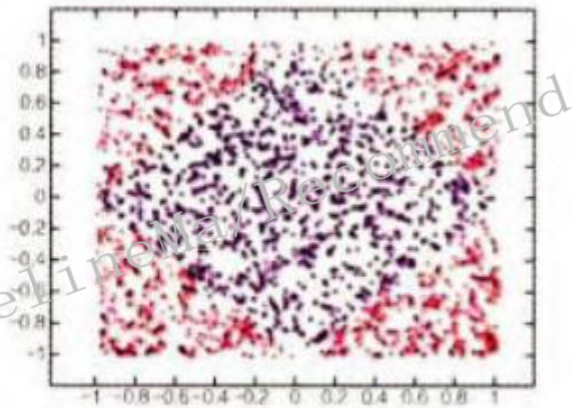
- LS-PLM具有样本分片的能力，因此能够挖掘出数据中蕴藏的非线性模式，省去了大量的人工样本处理和特征工程的过程
- 模型的稀疏性强：LS-PLM在建模时引人了L1和L2l范数，可以使最终训练出来的模型具有较高的稀疏度，使模型的部署更加轻量级



训练数据



LR模型



MLR模型

这里用一个二维的例子来解释为什么 L1 范数更容易产生稀疏性。L2 范数 $|w_1|^2+|w_2|^2$ 的曲线如图 2-19(a)的红色圆形，L1 范数 $|w_1|+|w_2|$ 的曲线如图 2-19(b)红色菱形。用蓝色曲线表示不加正则化项的模型损失函数曲线。

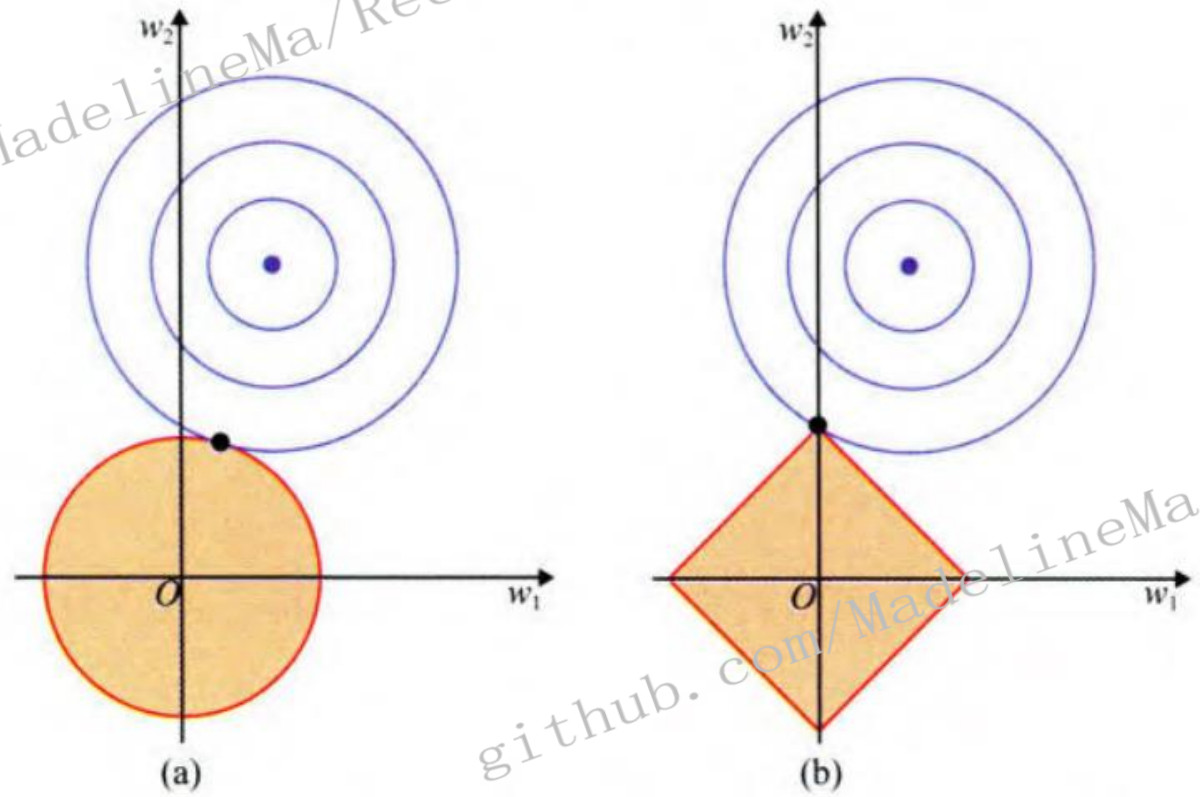


图 2-19　L1 范数和 L2 范数与损失函数“损失等高线”示意图

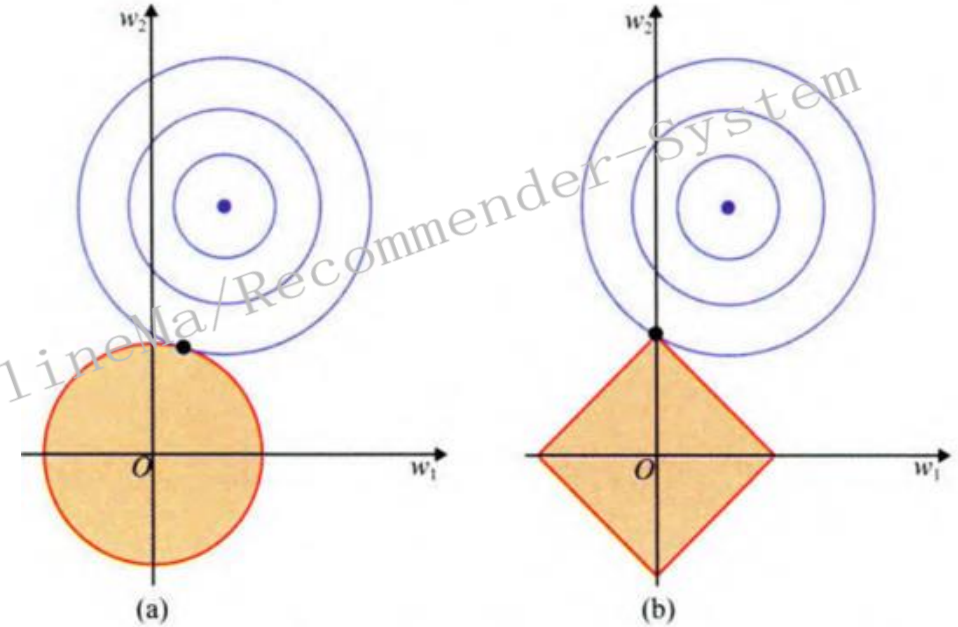


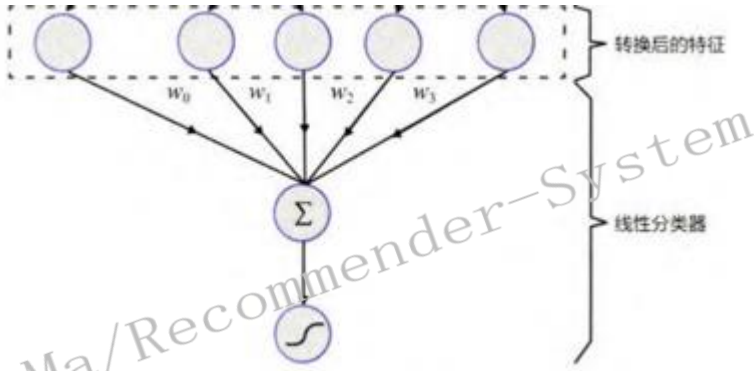
图 2-19　L1 范数和 L2 范数与损失函数“损失等高线”示意图

求解加入正则化项的损失函数最小值，就是求解红圈上某一点和蓝圈上某一点之和的最小值。这个值通常在红色曲线和蓝色曲线的相切处（如果不在相切处，那么至少有两点值相同，与极值的定义矛盾），而 L1 范数曲线更容易与蓝色曲线在顶点处相交，这就导致除了相切处的维度不为零，其他维度的权重均为 0，从而容易产生模型的稀疏解。

LS-PLM

- LS-PLM可以看作一个加入了Attention机制的三层神经网络模型，其中输入层是样本的特征向量，中间层是由m个神经元组成的隐层，其中是分片的个数，对于一个CTR预估问题，LS-PLM的最后一层自然是由单一神经元组成的输出层。

$$f(x) = \sum_{i=1}^{m} \pi_i(x) \cdot \eta_i(x) = \sum_{i=1}^{m} \frac{e^{\mu_i \cdot x}}{\sum_{j=1}^{m} e^{\mu_j \cdot x}} \cdot \frac{1}{1 + e^{-w_i \cdot x}}$$



$$\sigma_2(\sigma_1(f(x)))$$

- 在隐层和输出层之间，神经元之间的权重是由分片函数得出的注意力得分来确定的。也就是说，样本属于哪个分片的概率就是其注意力得分.

北京師範大學珠海校区
BEIJING NORMAL UNIVERSITY AT ZHUHAI

04
总结

We can read of things that happened 5,000 years ago in the Near East, where people first learned to write.

总结

模型名称	基本原理	特　点	局限性
协同过滤	根据用户的行为历史生成用户-物品共现矩阵，利用用户相似性和物品相似性进行推荐	原理简单、直接，应用广泛	泛化能力差，处理稀疏矩阵的能力差，推荐结果的头部效应较明显
矩阵分解	将协同过滤算法中的共现矩阵分解为用户矩阵和物品矩阵，利用用户隐向量和物品隐向量的内积进行排序并推荐	相较协同过滤，泛化能力有所加强，对稀疏矩阵的处理能力有所加强	除了用户历史行为数据，难以利用其他用户、物品特征及上下文特征
逻辑回归	将推荐问题转换成类似CTR预估的二分类问题，将用户、物品、上下文等不同特征转换成特征向量，输入逻辑回归模型得到CTR，再按照预估CTR进行排序并推荐	能够融合多种类型的不同特征	模型不具备特征组合的能力，表达能力较差
FM	在逻辑回归的基础上，在模型中加入二阶特征交叉部分，为每一维特征训练得到相应特征隐向量，通过隐向量间的内积运算得到交叉特征权重	相比逻辑回归，具备了二阶特征交叉能力，模型的表达能力增强	由于组合爆炸问题的限制，模型不易扩展到三阶特征交叉阶段
FFM	在FM模型的基础上，加入“特征域”的概念，使每个特征在与不同域的特征交叉时采用不同的隐向量	相比FM，进一步加强了特征交叉的能力	模型的训练开销达到了$O(n^2)$的量级，训练开销较大

总结

模型名称	基本原理	特　点	局限性
GBDT+LR	利用 GBDT 进行“自动化”的特征组合，将原始特征向量转换成离散型特征向量，并输入逻辑回归模型，进行最终的 CTR 预估	特征工程模型化，使模型具备了更高阶特征组合的能力	GBDT 无法进行完全并行的训练，更新所需的训练时长较长
LS-PLM	首先对样本进行“分片”，在每个“分片”内部构建逻辑回归模型，将每个样本的各“分片”概率与逻辑回归的得分进行加权平均，得到最终的预估值	模型结构类似三层神经网络，具备了较强的表达能力	模型结构相比深度学习模型仍比较简单，有进一步提高的空间

本周作业

1. GBDT防止过拟合的方式（面经）.
2. NN的前向推到和反向递推过程.


北京師範大學珠海校区
BEIJING NORMAL UNIVERSITY AT ZHUHAI
github.com/MadelineMa/Recommender-System
THANKS
github.com/MadelineMa/Recommender-System
DESIGNED BY 2xh