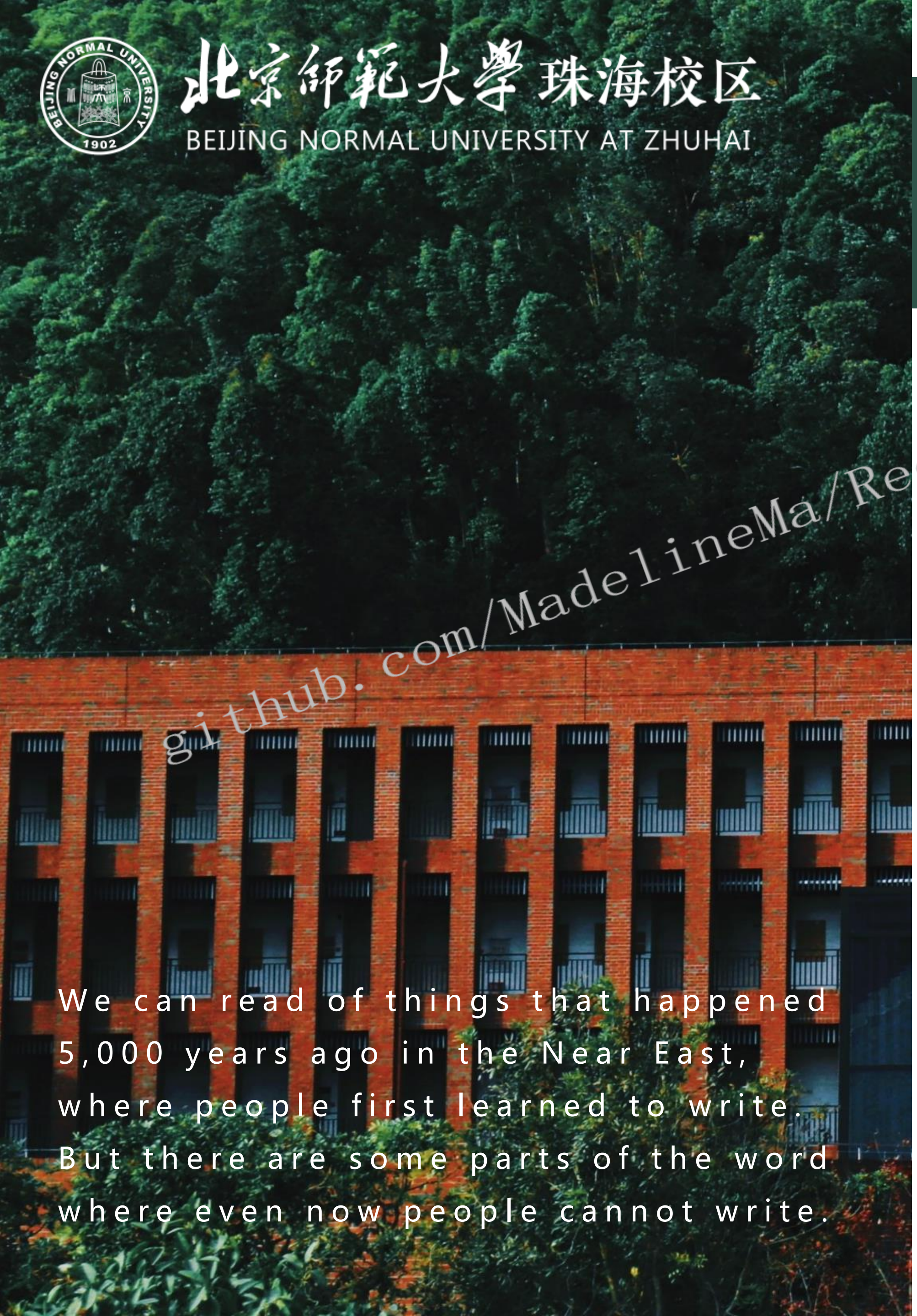
北京师範大學 珠海校区
BEIJING NORMAL UNIVERSITY AT ZHUHAI

前深度学习时代

马静


北京師範大學 珠海校区
BEIJING NORMAL UNIVERSITY AT ZHUHAI
We can read of things that happened
5,000 years ago in the Near East,
where people first learned to write.
But there are some parts of the word
where even now people cannot write.

CONTENT

北京师范大学 珠海校区
BEIJING NORMAL UNIVERSITY AT ZHUHAI

01逻辑回归
Logistic Regression

Logistic regression

- 首次综合利用用户、物品、上下文等多种不同的特征，生成较为“全面”的推荐结果．
- 举例用户物品、上下文特征？
- 算法简单，但是感知机中最简单的一种，深度学习的基础性结构．
- 虽然叫回归，但是利用“分类”思想，对匹配度进行打分．
- 1：点击，观看，购买等隐式正反馈，0：其他．
- LR将将推荐问题转换成了一个点击率（Click Through Rate, CTR)预估问题．

流程

- Data Preprocessing: 将用户年龄、性别、物品属性、物品描述、当前时间、当前地点等特征转换成数值型特征向量。
- Train: 确定逻辑回归模型的优化目标（以优化“点击率”为例），利用已有样本数据对逻辑回归模型进行训练，确定逻辑回归模型的内部参数。
- Predict: 在模型服务阶段，将特征向量输人逻辑回归模型，经过逻辑回归模型的推断，得到用户“点击”（这里用点击作为推荐系统正反馈行为的例子）物品的概率。
- 利用“点击”概率对所有候选物品进行排序，得到推荐列表。

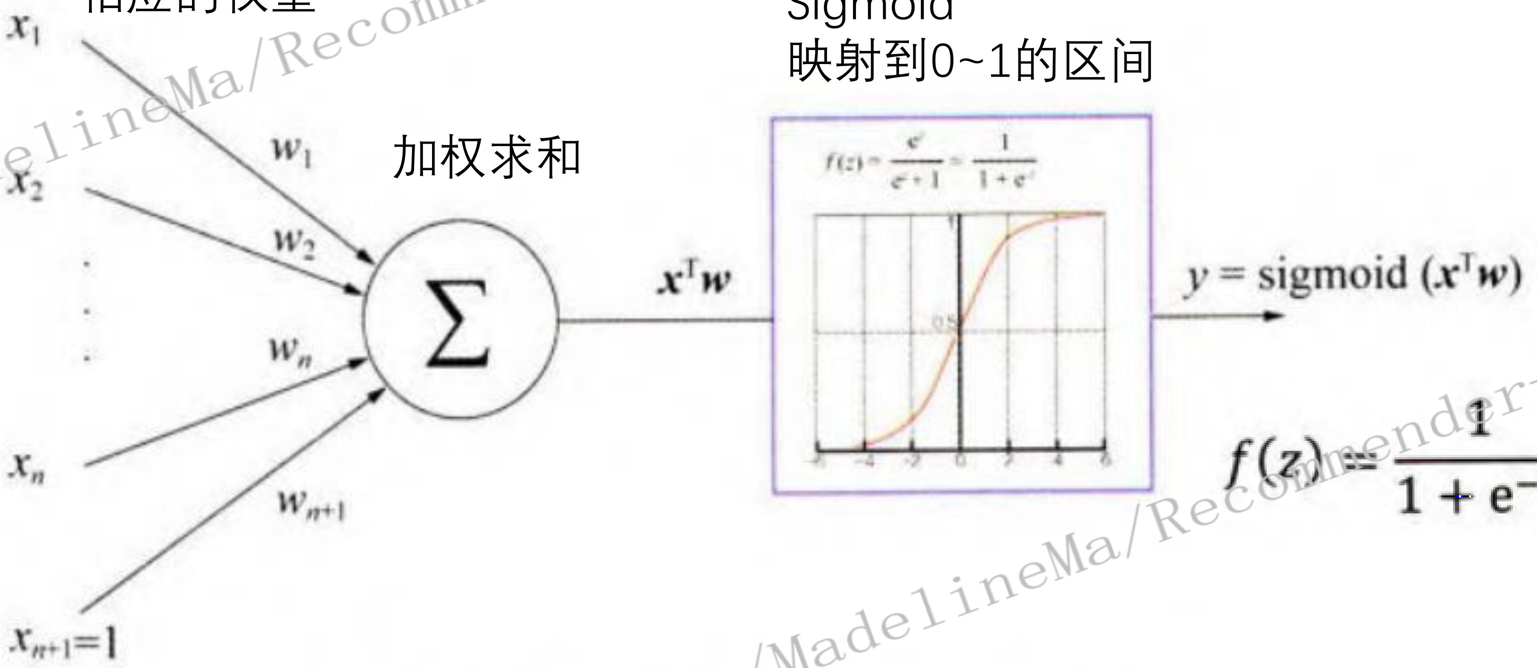
数学形式

为各特征赋予相应的权重

输入特征向量

加权求和

Sigmoid
映射到0~1的区间



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

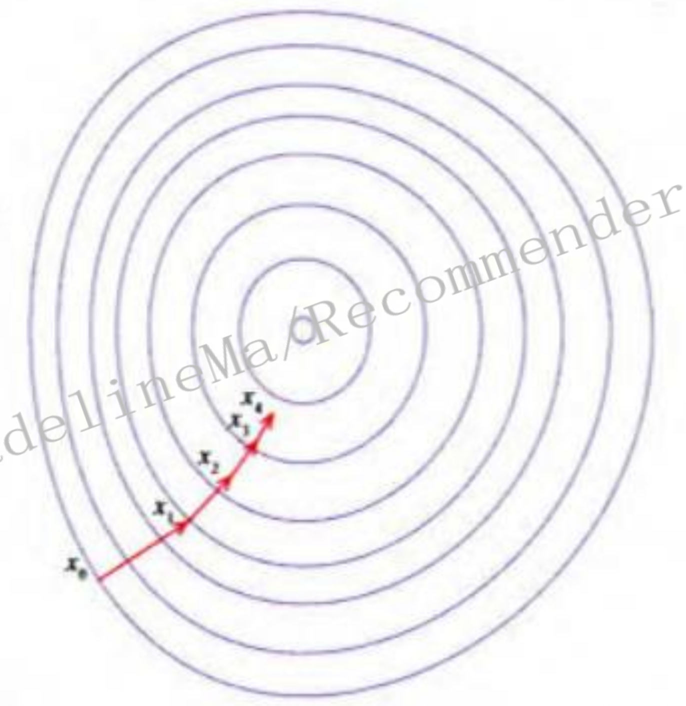
$$f(x) = \frac{1}{1+\mathrm{e}^{-(w\cdot x+b)}}$$

权重向量W的训练方法

- 回顾：梯度下降法
- 一阶最优化算法，也称为最速下降法(梯度下降最大的方向).
- 目标：局部极小值. 必须沿函数上当前点对应梯度（或者是近似梯度）的反方向进行规定步长距离的迭代搜索.

在寻找最低点的过程中，沿哪个方向才是下降最快的方向呢?

"梯度"的性质：如果实值函数$F(x)$在点x_0处可微且有定义，
那么函数F(x)在点处沿着梯度相反的方向$-\nabla F(x_0)$下降最快.

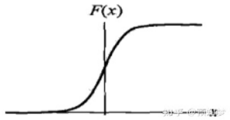


权重向量W的训练方法

- ✓ 确定目标函数
- ✓ 对目标函数进行求导，得到梯度的方向
- ✓ 沿梯度的反方向下降，并迭代此过程直至寻找到局部最小点

基于梯度的ML通用法则！

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



正样本的概率： $P(y=1|x)=\frac{e^{\omega^T x+b}}{1+e^{\omega^T x+b}}$ $P(y=1|x)=\pi(x),$

负样本的概率： $P(y=0|x)=\frac{1}{1+e^{\omega^T x+b}}$ $P(y=0|x)=1-\pi(x).$

综合起来： $P(y|x,\omega)=\pi(x).^{y}(1-\pi(x))^{1-y}$

利用似然函数求极大值，梯度下降法求解

$$l=\prod_{i=1}^{m}[\pi(x^{(i)})]^{y^{(i)}}[1-\pi(x^{(i)})]^{1-y^{(i)}},$$

连乘不利于求导，改为对数形式，求平均：

$$L(\omega)=\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}log\pi\left(x^{(i)}\right)+\left(1-y^{(i)}\right)\log\left(1-\pi\left(x^{(i)}\right)\right)\right],$$

对目标函数求导：

$$\frac{\partial}{\partial\omega_j}L(\omega)=\sum_{i=1}^{m}\left(\pi\left(x^{(i)}\right)-y^{(i)}\right)x_j^{(i)}$$

github.com/MadeLineMa/Recommender-System

权重向量W的训练方法

- ✓ 确定目标函数
- ✓ 对目标函数进行求导，得到梯度的方向
- ✓ 沿梯度的反方向下降，并迭代此过程直至寻找到局部最小点

基于梯度的ML通用法则!

对目标函数求导:

$$\frac{\partial}{\partial \omega_j} L(\omega) = \sum_{i=1}^{m} \left(\pi(x^{(i)}) - y^{(i)}\right) x_j^{(i)}$$

模型更新公式:

$$\omega_j < -\omega_j - \gamma \sum_{i=1}^{m} \left(\pi(x^{(i)}) - y^{(i)}\right) x_j^{(i)}$$

扩展：交叉熵损失函数

交叉熵：信息论中的一个重要概念，主要用于度量两个概率分布间的差异性.

信息奠基人香农（Shannon）认为：“信息是用来消除随机不确定性的东西.

信息量的大小与信息发生的概率成**反比**。概率越大，信息量越小。概率越小，信息量越大.

$$I(x) = -\log(P(x))$$

example1：“太阳从东边升起”

example1： 2022年中国队成功进入世界杯

信息熵也被称为熵，用来表示所有信息量的期望。 $H(\mathbf{X}) = -\sum_{i=1}^{n} P(x_i)\log(P(x_i)) \quad (\mathbf{X} = x_1, x_2, x_3 \ldots, x_n)$

相对熵（KL散度）:如果对于同一个随机变量有两个单独的概率分布，则我们可以使用KL散度来衡量这两个概率分布之间的差异。

$$D_{KL}(p\|q) = \sum_{i=1}^{n} p(x_i)\log\left(\frac{p(x_i)}{q(x_i)}\right)$$

扩展：交叉熵损失函数

交叉熵：信息论中的一个重要概念，主要用于度量两个概率分布间的差异性.

$$D_{KL}(p||q) = \sum_{i=1}^{n} p(x_i)\log\left(\frac{p(x_i)}{q(x_i)}\right)$$

$$= \underbrace{\boxed{\sum_{i=1}^{n} p(x_i)\,log(p(x_i))}}_{\text{信息熵}} - \underbrace{\boxed{\sum_{i=1}^{n} p(x_i)\,log(q(x_i))}}_{\text{交叉熵}}$$

- 在机器学习训练网络时，输入数据与标签常常已经确定，那么真实概率分布P(x)也就确定下来了，所以信息熵在这里就是一个常量.
- KL散度的值表示真实概率分布P(x)与预测概率分布Q(x)之间的差异，值越小表示预测的结果越好，所以需要最小化KL散度，或最小化交叉熵.

权重向量W的训练方法

- ✓ 确定目标函数
- ✓ 对目标函数进行求导，得到梯度的方向
- ✓ 沿梯度的反方向下降，并迭代此过程直至寻找到局部最小点

利用似然函数求极大值，梯度下降法求解

$$l=\prod_{i=1}^{m}[\pi(x^{(i)})]^{y^{(i)}}[1-\pi(x^{(i)})]^{1-y^{(i)}},$$

连乘不利于求导，改为对数形式：

$$L(\omega)=\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}log\pi\left(x^{(i)}\right)+\left(1-y^{(i)}\right)\log\left(1-\pi\left(x^{(i)}\right)\right)\right],$$

对目标函数求导：

$$\frac{\partial}{\partial\omega_j}L(\omega)=\sum_{i=1}^{m}\left(\pi\left(x^{(i)}\right)-y^{(i)}\right)x_j^{(i)}$$

$$H(p,q)=-\sum_{i=1}^{n}p(x_i)\,log\,(q(x_i))$$

以交叉熵为损失函数：

$$L(\omega)=-\sum_{i=1}^{m}\left[y^{(i)}\log(\pi\left(x^{(i)}\right))+\left(1-y^{(i)}\right)\log\left(1-\pi\left(x^{(i)}\right)\right)\right],$$

似然函数极大化 ⇔ 交叉熵极小化！

权重向量W的训练方法

- ✓ 确定目标函数
- ✓ 对目标函数进行求导，得到梯度的方向
- ✓ 沿梯度的反方向下降，并迭代此过程直至寻找到局部最小点

为什么不选用平方误差?

$$L = \frac{(y-\hat{y})^2}{2}$$

$$\frac{\partial L}{\partial w} = (\hat{y}-y)\sigma'(w\cdot x)x$$

$$\sigma'(w\cdot x) = w\cdot x(1-w\cdot x)$$

参数初始化后，导数值可能很小，导致收敛变慢，训练过程也可能会发生梯度消失

交叉熵的梯度:

$$g' = \sum_{i=1}^{N} x_i(y_i - p(x_i))$$

当模型输出概率偏离真实概率时，梯度大，加快收敛
反之，训练速度变缓慢，防止震荡。

Logistic regression

算法优势

- LR的假设服从伯努利分布与CTR契合
- Sigmoid映射到0~1符合CTR物理意义
- 不仅预测出类别，还可得到近似概率预测；
- 因为结果是概率，可用作排序.

工程优势

- 对率函数是任意阶可导凸函数，有很好得数学性质，很多数值优化算法可直接用于求取最优解；
- 容易使用和解释，计算代价低；
- LR对时间和内存需求上相当高效；

业界倾向

- 可应用于分布式数据，并且还有在线算法实现，用较小资源处理较大数据；
- 灵活进行特征选取和交叉，方便特征工程的迭代.

千亿特征流式学习在大规模推荐排序场景的应用 - 知乎 (zhihu.com)
CTR预测算法之FTRL-Proximal - 知乎 (zhihu.com)

github.com/MadelineMa/Recommender-System

北京师范大学 珠海校区
BEIJING NORMAL UNIVERSITY AT ZHUHAI

02

FM、FFM

github.com/MadelineMa/Recommender-System

扩展

基础知识——什么是辛普森悖论

在对样本集合进行分组研究时，在分组比较中都占优势的一方，在总评中有时反而是失势的一方，这种有悖常理的现象，被称为“辛普森悖论”。

表 2-1　男性用户

视　频	点击（次）	曝光（次）	点击率
视频 A	8	530	1.51%
视频 B	51	1520	3.36%

表 2-2　女性用户

视　频	点击（次）	曝光（次）	点击率
视频 A	201	2510	8.01%
视频 B	92	1010	9.11%

从以上数据中可以看出，无论男性用户还是女性用户，对视频 B 的点击率都高于视频 A，显然推荐系统应该优先考虑向用户推荐视频 B。

扩展

那么，如果忽略性别这个维度，将数据汇总（如表 2-3 所示）会得出什么结论呢？

表 2-3　数据汇总

视　　频	点击（次）	总曝光（次）	点击率
视频 A	209	3040	6.88%
视频 B	143	2530	5.65%

在“辛普森悖论”的例子中，分组实验相当于使用“性别”+“视频 id”的组合特征计算点击率，而汇总实验则使用“视频 id”这一单一特征计算点击率。汇总实验对高维特征进行了合并，损失了大量的有效信息，因此无法正确刻画数据模式。

FM

- 针对LR无法做特征交叉和筛选的局限性和辛普森悖论可能性.
- 免去工程师手动组合的经验局限
- POLY2的暴力组合的稀疏加剧和复杂度上升n->n^2

$$\varnothing \mathrm{POLY2}(\boldsymbol{w},\boldsymbol{x}) = \sum_{j_1=1}^{n} \sum_{j_2=j_1+1}^{n} w_{h(j_1,j_2)} x_{j_1} x_{j_2}$$

Category数据one-hot编码,
相乘后更加稀疏

$$\underbrace{[0, 1, 0, 0, 0, 0, 0]}_{\text{Weekday = Tuesday}} \quad \underbrace{[0, 1]}_{\text{Gender = Male}} \quad \underbrace{[0, 0, 1, 0, \dots, 0, 0]}_{\text{City = London}}$$

FM

- 2010年提出，用两个向量内积取代了权重系数.

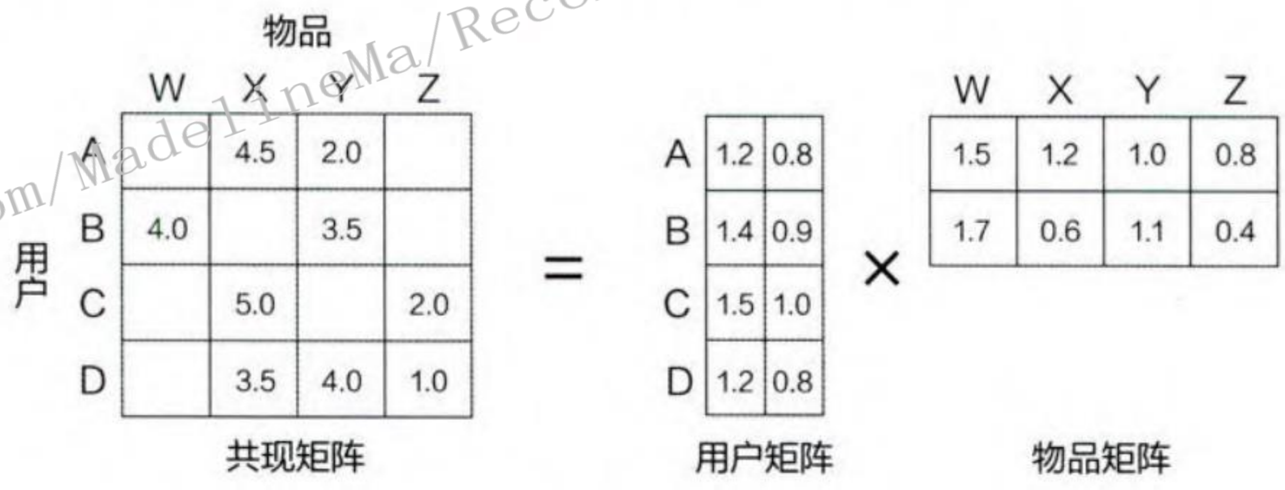
$$\phi POLY2(\boldsymbol{w}, \boldsymbol{x}) = \sum_{j_1=1}^{n} \sum_{j_2=j_1+1}^{n} w_{h(j_1,j_2)} x_{j_1} x_{j_2} \qquad \phi FM(\boldsymbol{w}, \boldsymbol{x}) = \sum_{j_1=1}^{n} \sum_{j_2=j_1+1}^{n} (\boldsymbol{w}_{j_1} \cdot \boldsymbol{w}_{j_2}) x_{j_1} x_{j_2}$$

FM为每个特征学习了一个隐权重向量(latent vector)

- FM是将矩阵分解隐向量的思想进行了进一步扩展，从单纯的用户、物品隐向量扩展到了所有特征上.
- 参数规模由n^2降至nk.
- 解决了稀疏问题，增强了泛化能力.

矩阵分解概览

- 目标：用隐向量表达用户和物品，还要保证相似的用户及用户可能喜欢的物品的距离相近.
- **用户和物品的隐向量是通过分解协同过滤生成的共现矩阵得到的.**



- 将mxn维的共现矩阵及分解为mxk维的用户矩阵U, kxn维的物品矩阵V相乘的形式.
- k的大小决定了隐向量表达能力的强弱和泛化能力强弱?
- 最终打分 $\hat{r}_{ui} = q_i^T p_u$

FM

隐向量的引入使 FM 能更好地解决数据稀疏性的问题。举例来说，在某商品推荐的场景下，样本有两个特征，分别是频道（channel）和品牌（brand），某训练样本的特征组合是(ESPN, Adidas)。在 POLY2 中，只有当 ESPN 和 Adidas 同时出现在一个训练样本中时，模型才能学到这个组合特征对应的权重；而在 FM 中，ESPN 的隐向量也可以通过(ESPN, Gucci)样本进行更新，Adidas 的隐向量也可以通过(NBC, Adidas)样本进行更新，这大幅降低了模型对数据稀疏性的要求。甚至对于一个从未出现过的特征组合(NBC，Gucci)，由于模型之前已经分别学习过 NBC 和 Gucci 的隐向量，具备了计算该特征组合权重的能力，这是 POLY2 无法实现的。相比 POLY2，FM 虽然丢失了某些具体特征组合的精确记忆能力，但是泛化能力大大提高。

北京师范大学珠海校区
BEIJING NORMAL UNIVERSITY AT ZHUHAI

03 FFM

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

FFM

- 2015年提出，引入特征域的概念，在多项CTR预估大赛中夺魁，并被Criteo、美团等公司深度应用在推荐系统、CTR预估等领域.

$$\text{ØFM}(\boldsymbol{w},\boldsymbol{x}) = \sum_{j_1=1}^{n}\sum_{j_2=j_1+1}^{n}(\boldsymbol{w}_{j_1}\cdot\boldsymbol{w}_{j_2})x_{j_1}x_{j_2} \qquad \text{ØFFM}(\boldsymbol{w},\boldsymbol{x}) = \sum_{j_1=1}^{n}\sum_{j_2=j_1+1}^{n}(\boldsymbol{w}_{j_1,f_2}\cdot\boldsymbol{w}_{j_2,f_1})x_{j_1}x_{j_2}$$

- 每个特征对应的不是唯一一个隐向量，而是一组隐向量；
- 特征1与特征2进行交叉时，特征1会挑出与特征2的域对应的隐向量进行交叉；
- 比如淘宝有很多场景，特征1：用户喜好类目，特征2：用户喜好场景，隐向量1：在场景域下的喜好类目，隐向量2：在类目域下的喜好场景.

FFM

域	Publisher(P)	Advertiser(A)	Gender(G)
特征值	ESPN	NIKE	Male

"域"代表特征域；
域内的特征一般是采用one-hot编码形成的一段one-hot特征向量；

如果按照 FM 的原理，特征 ESPN、NIKE 和 Male 都有对应的隐向量 $\boldsymbol{w}_{\mathrm{ESPN}}, \boldsymbol{w}_{\mathrm{NIKE}}, \boldsymbol{w}_{\mathrm{Male}}$，那么 ESPN 特征与 NIKE 特征、ESPN 特征与 Male 特征做交叉的权重应该是 $\boldsymbol{w}_{\mathrm{ESPN}} \cdot \boldsymbol{w}_{\mathrm{NIKE}}$ 和 $\boldsymbol{w}_{\mathrm{ESPN}} \cdot \boldsymbol{w}_{\mathrm{Male}}$。其中，ESPN 对应的隐向量 $\boldsymbol{w}_{\mathrm{ESPN}}$ 在两次特征交叉过程中是不变的。

而在 FFM 中，ESPN 与 NIKE、ESPN 与 Male 交叉特殊的权重分别是 $\boldsymbol{w}_{\mathrm{ESPN,A}} \cdot \boldsymbol{w}_{\mathrm{NIKE,P}}$ 和 $\boldsymbol{w}_{\mathrm{ESPN,G}} \cdot \boldsymbol{w}_{\mathrm{Male,P}}$。

FFM

域	Publisher(P)	Advertiser(A)	Gender(G)
特征值	ESPN	NIKE	Male

"域"代表特征域;
域内的特征一般是采用one-hot编码形成的一段one-hot特征向量;

- 在FFM模型的训练过程中，需要学习n个特征在f个域上的k维隐向量，参数数量共n*k*f个.
- 复杂度和参数个数上升，需要在效果和工程投入之间权衡.

回顾

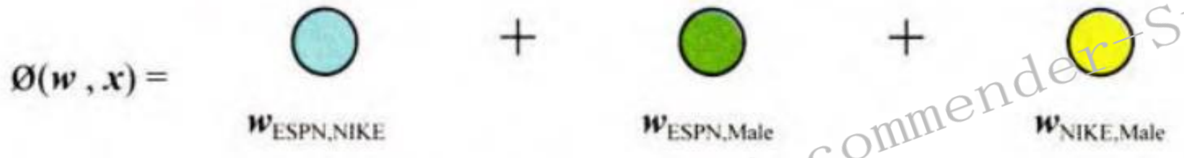


图 2-12 POLY2 模型示意图

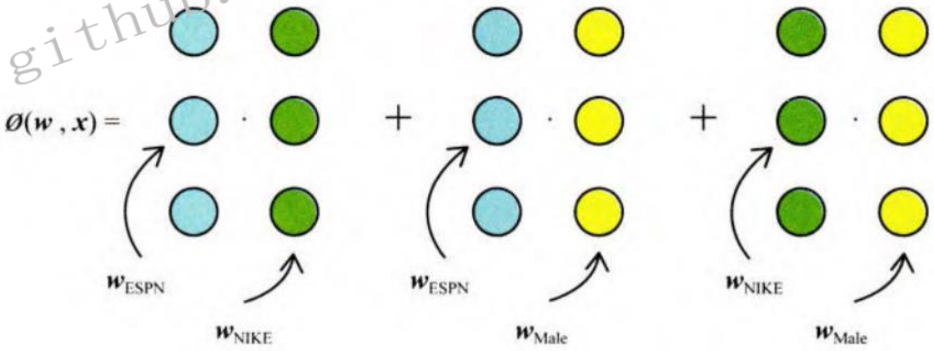


图 2-13 FM 模型示意图

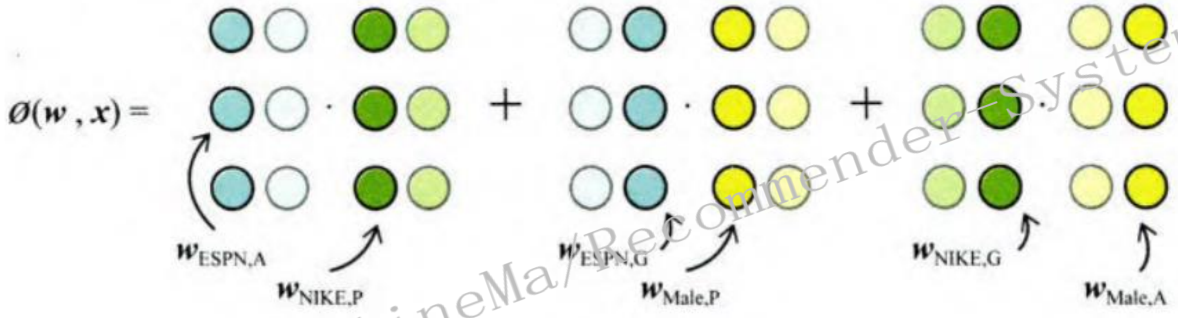


图 2-14 FFM 模型示意图

泛化能力强，但记忆能力有所减弱

权重数量和训练复杂度升高

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

本周作业

1. 自学SVM，Baysian，Tree类的ML方法，对比LR与其他方法区别.
2. AutoFIS课题小组可扩展本节内容进行资料阅读.

Reference：

逻辑回归的原理及Python实现 - 知乎 (zhihu.com)

【机器学习】逻辑回归（非常详细） - 知乎 (zhihu.com)

带你理解朴素贝叶斯分类算法 - 知乎 (zhihu.com)

推荐系统召回四模型之：全能的FM模型 - 知乎 (zhihu.com) -> 可能后半部分读不懂，可课堂讨论.

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

THANKS

github.com/MadelineMa/Recommender-System

DESIGNED BY 2xh