相比于 SVM，FM 模型如何学习交叉特征？其如何优化？

原创 去远方看太阳 随时学丫 3月25日

这是小娌的第 9 期分享

作者 | 小娌

来源 | 小娌（ID: a18581391726）

本文共 2572 字，预计阅读需 7 分钟

在计算广告和推荐系统中，CTR 预估是非常重要的一个环节，判断物品是否进行推荐，需要根据 CTR 预估的点击率排序决定。业界常用的方法有人工特征 + LR，GBDT + LR，FM 和 FFM 等模型。

近几年提出了很多基于 FM 改进的方法，如 DeepFM，FNN，PNN，DCN，xDeepFM 等，今天给大家分享 FM。

Factorization Machine (FM) 是由 Steffen Rendle 在 2010 年提出的，模型主要通过特征组合来解决大规模稀疏数据的分类问题。

Part.1

什么是 Factorization Machine?

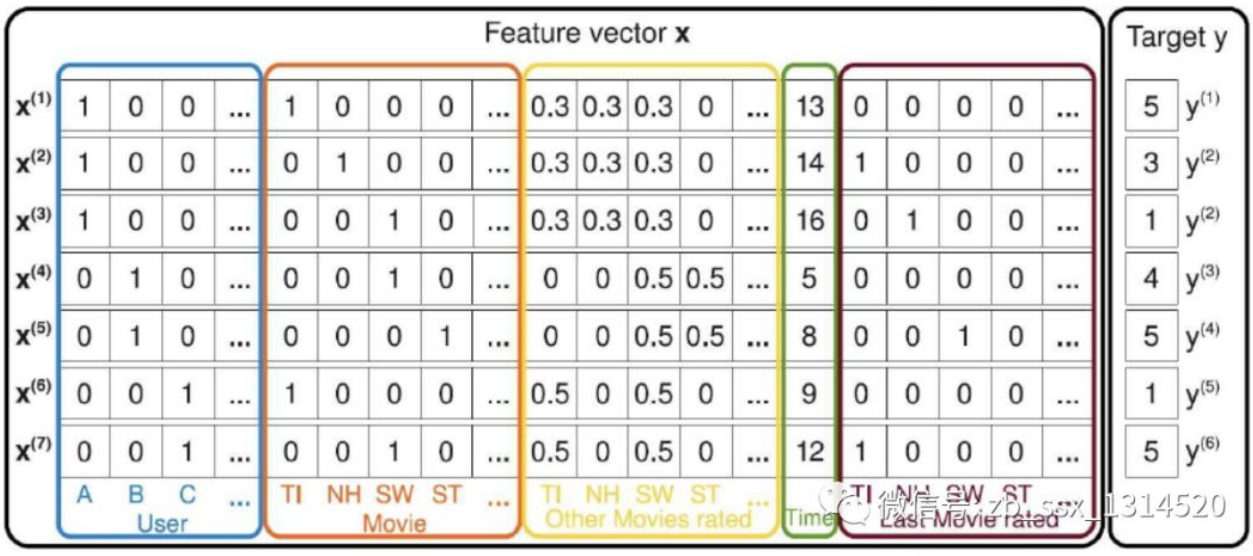
One-Hot 带来的问题？

在面对 CTR 预估的问题的时候，我们常常会转化为下面这种类型的二分类问题。

点击	性别	国别
1	男	中国
0	女	美国
1	女	法国

由于 性别，国别 等特征都是类别特征，所以在使用的时候常常采用 One-Hot Encoding 将其转化为数值类型。

点击	性别 = 男	性别 = 女	国别 = 中国	国别 = 美国	国别 = 法国
1	1	0	1	0	0
0	0	1	0	1	0
1	0	1	0	0	1



上图可以看出，在经过 One-Hot 编码之后，每个样本的特征空间都变大了许多，特征矩阵变得非常稀疏，在现实生活中，我们常常可以看见超过 10^7 维的特征向量。

如果我们采用单一的线性模型来学习用户的点击，打分习惯，我们很容易忽略特征潜在的组合关系，比如：女性喜欢化妆品，男性喜欢打游戏，买奶粉的用户常常买尿不湿等。

二阶多项式核 SVM

SVM 为了学习交叉特征，引入了核函数的概念，最简单直接的做法就是为两两的特征组合分配一个权重参数。这些新的权重参数和原始特征对应的参数一样，交给模型去在训练阶段学习。如此一来就形成了如下的预测函数：

$$\bar{y} = f(x) = w_0 + \sum_{i=1}^{n} + \sum_{0<i<j<=n} w_{i,j} x_i x_j$$

这实际上就是核函数选择为二阶多项式核的 SVM 模型。这样设计的模型看起来能够学到特征的两两交叉带来的信息，但这只是理论上的改进，但是模型在处理大量稀疏数据时，没有很好的泛化能力。

由于 $w_{i,j}$ 的取值完全取决于 x_i 和 x_j 的乘积，在数据稀疏的场景下，可能存在训练集中 $x_i x_j$ 始终为零的情况，这样一来，模型就无法有效的更新权重 $w_{i,j}$ 了，更进一步，在预测阶段，模型遇到 $x_i x_j$ 不为零的情况就很难有效的泛化。

因子分解机模型

既然二阶多项式核 SVM 泛化性能不足的原因是 $w_{i,j}$ 的取值完全取决于 x_i 和 x_j 的乘积，那么最直接的办法就是突破这一限制了。

FM 模型的解决办法是为每个维度的特征 (x_i) 学习一个表征向量 (v_i， 其实可以理解为特征 ID 的 Embedding 向量)。而后将 x_i 和 x_j 的乘积的权重设定为各自表征向量的点积，也就是有如下形式的预测函数：

$$\bar{y} = f(x) = w_0 + \sum_{i=1}^{n} + \sum_{0<i<j<=n} < v_i, v_j > x_i x_j$$

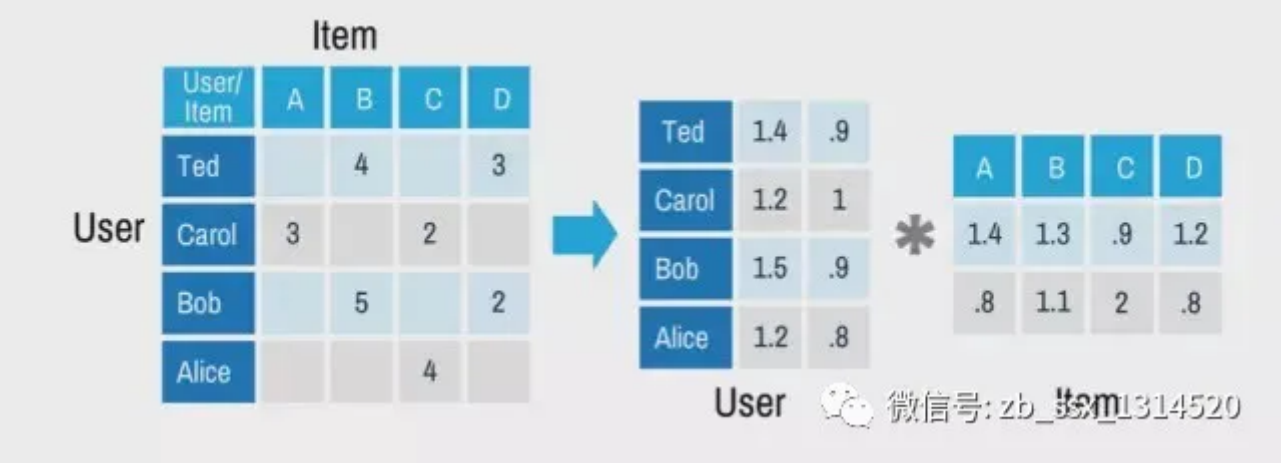
显然，FM 模型也具有二阶多项式核 SVM 的优点：能够学习到特征两两交叉带来的信息。

通过下面这个表达可以看出，FM 很像是计算每个经过 One-Hot 编码变化的特征 Embedding，然后学习不同特征之间 Embedding 相似度对于最后预测结果的影响。由于我们可以将上千万维的稀疏向量压缩为几十，或者几百维的 Embedding，极大地减小了模型的参数数量级，从而增强模型的泛化能力，得到更好的预测结果。

我们回到上一小节举的例子：训练集中 $x_i x_j$ 始终为零。在二阶多项式核 SVM 中，由于参数权重 $w_{i,j}$ 得不到更新，模型无法学到 x_i 和 x_j 交叉带来的信息。但是在 FM 中，x_i 和 x_j 的参数并不完全由 x_i 和 x_j 的乘积决定。具体来说，每一维特征的表征向量由该维特征与其它所有维度特征的交叉共同决定。于是，只要存在某个 k 使得 x_i 和 x_k 的乘积不总是为零，那么第 i 维特征的表征向量 $v_i^{\rightarrow}$ 就能够学到有效的信息——同理对 $v_j^{\rightarrow}$ 也有同样的结论。于是乎，

哪怕在训练集中，$x_i x_j$ 始终为零，其参数 $\langle v_i^{\rightarrow}, v_j^{\rightarrow} \rangle$ 也是经过了学习更新的，因此能够表现出很好的泛化性能。

FM 和矩阵分解



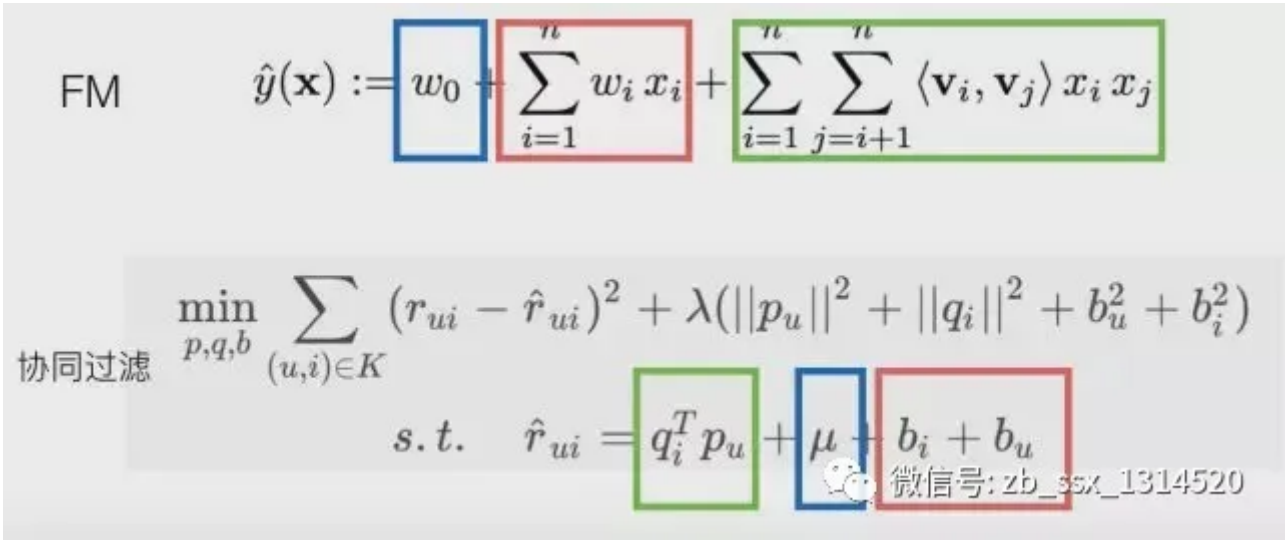
基于矩阵分解的协同过滤是推荐系统中常用的一种推荐方案，从历史数据中收集 user 对 item的 评分，可以是显式的打分，也可以是用户的隐式反馈计算的得分。由于 user 和 item 数量非常多，有过打分的 user 和 item 对通常是十分稀少的，基于矩阵分解的协同过滤是来预测那些没有过行为的 user 对 item 的打分，实际上是一个评分预测问题。

矩阵分解的方法假设 user 对 item 的打分 R 由 User Embedding 和 Item Embedding 相似性以及用户，物品的偏见决定。

这些参数可以通过最小化经验误差得到:

$$min_{p,q,b} \sum_{(u,i) \in K} (r_{ui} - \bar{r}_{ui})^2 + \lambda(||p_u||^2 + ||q_i||^2 + b_u^2 + b_i^2)$$

从上面的叙述来看，FM 的二阶矩阵也用了矩阵分解的技巧，那么基于矩阵分解的协同过滤和 FM 是什么关系呢？以 user 对 item 评分预测问题为例，基于矩阵分解的协同过滤可以看做 FM的一个特殊例子，对于每一个样本，FM 可以看做特征只有 userid 和 itemid 的 onehot 编码后的向量连接而成的向量。另外，FM 可以采用更多的特征，学习更多的组合模式，这是单个矩阵分解的模型所做不到的！因此，FM 比矩阵分解的方法更具普遍性！事实上，现在能用矩阵分解的方法做的方案都直接上 FM 了!



Part.2

FM 如何解决效率问题？

考虑到 FM 模型会对特征进行二阶组合，在有 n 个原始特征时，交叉特征就会有 $(n^2 - n)/2$ 个。因此，如果不做任何优化，FM 模型的复杂度会是 $O(n2)$，具体来说是 $O(kn^2)$（其中 k 是表征向量的长度）。在特征规模非常大的场景中，这是不可接受的。

那么问题来了，是否有办法将复杂度降低到 $O(kn)$ 呢？答案是可以的，我们来看针对特征交叉项的一系列变换。

$$
\begin{aligned}
&\sum_{i=1}^{n}\sum_{j=i+1}^{n} < v_i, v_j > x_i x_j \\
&= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} < v_i, v_j > x_i x_j - \frac{1}{2}\sum_{i=1}^{n} < v_i, v_i > x_i x_i \\
&= \frac{1}{2}(\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{f=1}^{k} v_{i,f} v_{j,f} x_i x_j - \sum_{i=1}^{n}\sum_{f=1}^{k} v_{i,f} v_{i,f} x_i x_i) \qquad (4) \\
&= \frac{1}{2}\sum_{f=1}^{k}((\sum_{i=1}^{n} v_{i,f} x_i)(\sum_{j=1}^{n} v_{j,f} x_j) - \sum_{i=1}^{n} v_{i,f}^2 x_i^2) \\
&= \frac{1}{2}\sum_{f=1}^{k}((\sum_{i=1}^{n} v_{i,f} x_i)^2 - \sum_{i=1}^{n} v_{i,f}^2 x_i^2)
\end{aligned}
$$

可以看到这时的时间复杂度为 $O(kn)$。

Part.3

参数学习

从上面的描述可以知道FM可以在线性的时间内进行预测。因此模型的参数可以通过梯度下降的方法（例如随机梯度下降）来学习，对于各种的损失函数。FM模型的梯度是：

$$
\frac{\partial}{\partial \theta}\hat{y}(x) = \begin{cases} 1, & if\ \theta\ is\ w_0 \\ x_i, & if\ \theta\ is\ w_i \\ x_i \sum_{j=1}^{n} v_{j,f} x_j - v_{i,f} x_i^2, & if\ \theta\ is\ v_{i,f} \end{cases} \qquad (5)
$$

由于 $\sum_{j=1}^{n} v_{j,f} x_j$ 只与 f 有关，与 i 是独立的，可以提前计算出来，并且每次梯度更新可以在常数时间复杂度内完成，因此 FM 参数训练的复杂度也是 $O(kn)$ 。综上可知，FM可以在线性时间训练和预测，是一种非常高效的模型。

Part.4

总结

FM模型有两个优势：

1. 在高度稀疏的情况下特征之间的交叉仍然能够估计，而且可以泛化到未被观察的交叉
2. 参数的学习和模型的预测的时间复杂度是线性的

FM模型的优化点：

1. 特征为全交叉，耗费资源，通常 user 与 user，item 与 item 内部的交叉的作用要小于 user 与 item 的交叉。
2. 使用矩阵计算，而不是 for 循环计算。
3. 高阶交叉特征的构造。

END

点「在看」，就是态度