FFM原理及公式推导

上一篇讲了FM（Factorization Machines），今天说一说FFM（Field-aware Factorization Machines ）。

回顾一下FM：

$$\hat{y} = w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} v_i \cdot v_j x_i x_j \tag{1}$$

$\cdot$表示向量的内积。样本x是n维向量，x_i是第i个维度上的值。v_i是x_i对应的长度为K的隐向量，V是模型参数，所以所有样本都使用同一个V，即$x_{1,1}$与$x_{2,1}$都使用v_1。

在FFM（Field-aware Factorization Machines ）中每一维特征（feature）都归属于一个特定的field，field和feature是一对多的关系。比如

field	field1年龄	field2城市			field3性别	
feature	x1年龄	x2北京	x3上海	x4深圳	x5男	x6女
用户1	23	1	0	0	1	0
用户2	31	0	0	1	0	1

1. 对于连续特征，一个特征就对应一个Field。或者对连续特征离散化，一个分箱成为一个特征。比如

field	field1年龄			
feature	小于20	20-30	30-40	大于40
用户1	0	23	0	0
用户2	0	0	31	0

2. 对于离散特征，采用one-hot编码，同一种属性的归到一个Field

不论是连续特征还是离散特征，它们都有一个共同点：同一个field下只有一个feature的值不是0，其他feature的值都是0。

FFM模型认为v_i不仅跟x_i有关系，还跟与x_i相乘的x_j所属的Field有关系，即v_i成了一个二维向量$v_{F \times K}$，F是Field的总个数。FFM只保留了(1)中的二次项.

$$\hat{y} = \sum_{i=1}^{n} \sum_{j=i+1}^{n} v_{i,fj} \cdot v_{j,fi} x_i x_j \tag{2}$$

以上文的表格数据为例，计算用户1的$\hat{y}$

$$\hat{y} = v_{1,f2} \cdot v_{2,f1} x_1 x_2 + v_{1,f3} \cdot v_{3,f1} x_1 x_3 + v_{1,f4} \cdot v_{4,f1} x_1 x_4 + \cdots$$

由于x_2, x_3, x_4属于同一个Field，所以$f2, f3, f4$可以用同一个变量来代替，比如就用$f2$。

$$\hat{y} = v_{1,f2} \cdot v_{2,f1} x_1 x_2 + v_{1,f2} \cdot v_{3,f1} x_1 x_3 + v_{1,f2} \cdot v_{4,f1} x_1 x_4 + \cdots$$

我们来算一下$\hat{y}$对$v_{1,f2}$的偏导。

$$\frac{\partial \hat{y}}{\partial v_{1,f2}} = v_{2,f1} x_1 x_2 + v_{3,f1} x_1 x_3 + v_{4,f1} x_1 x_4$$

等式两边都是长度为K的向量。

注意x_2, x_3, x_4是同一个属性的one-hot表示，即x_2, x_3, x_4中只有一个为1，其他都为0。在本例中$x_3 = x_4 = 0, x_2 = 1$，所以

$$\frac{\partial \hat{y}}{\partial v_{1,f2}} = v_{2,f1} x_1 x_2$$

推广到一般情况：

$$\frac{\partial \hat{y}}{\partial v_{i,fj}} = v_{j,fi} x_i x_j \tag{3}$$

x_j属于Field fj，且同一个Field里面的其他x_m都等于0。实际项目中x是非常高维的稀疏向量，求导时只关注那些非0项即可。

你一定有个疑问：v是模型参数，为了求v我们□采用梯度下降法时需要计算损失函数对v的导数，为什么这里要计算$\hat{y}$对v的导数？看看分割线下方的内容你就明白了。

在实际预测点击率的项目中我们是不会直接使用公式(2)的，通常会再套一层sigmoid函数。公式(2)中的$\hat{y}$我们用z来取代。

$$z=\phi(v,x)=\sum_{i=1}^{n}\sum_{j=i+1}^{n}v_{i,fj}\cdot v_{j,fi}x_ix_j$$

由公式(3)得

$$\frac{\partial z}{\partial v_{i,fj}}=v_{j,fi}x_ix_j$$

用a表示对点击率的预测值

$$a=\sigma(z)=\frac{1}{1+e^{-z}}=\frac{1}{1+e^{-\phi(v,x)}}$$

令$y=0$表示负样本，$y=1$表示正样本，C表示交叉熵损失函数。根据《神经网络调优》中的公式(1)(2)可得

$$\frac{\partial C}{\partial z}=a-y=\begin{cases}-\frac{1}{1+e^{z}} & if\ y\text{是正样本}\\ \frac{1}{1+e^{-z}} & if\ y\text{是负样本}\end{cases}$$

$$\frac{\partial C}{\partial v_{i,fj}}=\frac{\partial C}{\partial z}\frac{\partial z}{\partial v_{i,fj}}$$

看完了本博客再去看论文《Field-aware Factorization Machines for CTR Prediction》中的公式推导应该就比较容易了吧，在该论文中他是以$y=1$代表正样本，$y=-1$代表负样本，所以才有了3.1节中的

$$\kappa=\frac{\partial C}{\partial z}=\frac{-y}{1+e^{yz}}$$

加速计算

关注(2)式，当x都是one-hot时可以写成

$$\sum_{i=1}^{n}\sum_{j=i+1}^{n}V_{i,fj}V_{j,fi}$$

公式通过变形可以减少计算量，这里要分两种情况:i和j是否属于同一个Field。

i和j不属于同一个Field

$$\sum_{i\in Filed1}\sum_{j\in Filed2}V_{i,f2}V_{j,f1}=\sum_{i\in Filed1}V_{i,f2}\sum_{j\in Filed2}V_{j,f1}$$

举个例子，比如a、b、c属于Field1，d、e属于Field2，则$ad+ae+bd+de+cd+ce=(a+b+c)(d+e)$。只需要一次乘法。

i和j属于同一个Field

$$\begin{aligned}\sum_{i=1}\sum_{j=i+1}V_{i,f}V_{j,f}&=\frac{1}{2}\left[\sum_{i=1}\sum_{j=1}V_{i,f}V_{j,f}-\sum_{i=1}V_{i,f}^2\right]\\&=\frac{1}{2}\left[\sum_{i=1}V_{i,f}\sum_{j=1}V_{j,f}-\sum_{i=1}V_{i,f}^2\right]\\&=\frac{1}{2}\left[\left(\sum_{i=1}V_{i,f}\right)^2-\sum_{i=1}V_{i,f}^2\right]\end{aligned}$$

举个例子，比如a、b、c属于同一个Fied，则$ab+ac+bc=\frac{1}{2}[(a+b+c)^2-(a^2+b^2+c^2)]$。乘法计算量由$O(n^2)$ 降为$O(n)$，n表示该Field内有几个特征。

原文来自:博客园（华夏35度）http://www.cnblogs.com/zhangchaoyang

作者:张朝阳

我的新书：工业机器学习算法详解与实战

分类: DataMining