逻辑回归: 逻辑回归的步骤:

1. 将特征向量 $x = (x_1, x_2, ..., x_n)^T$ 作为模型引入

2. 通过为各特征向量赋予权重 $(w_1, w_2, ...w_n)$ 来表示各特征重要性的差异, 将各特征进行加权求和, 得到 $x^T w$。

3. 将 $x^T w$ 输入到 sigmoid 函数, 使之映射到 0 1 的区间, 得到最终“点击率”。

$$\hat{y} = \sigma(w^T x + b), \sigma(y) = \frac{1}{1 + e^{-y}}$$

损失函数 (loss function):

$$\begin{cases} p(y = 1|x; w) = f_w(x) \\ p(y = 0|x; w) = 1 - f_w(x) \end{cases}$$

$$P(y|x; w) = (f_w(x))^y (1 - f_w(x))^{1-y}$$

极大似然估计:

$$L(w) = \prod_{i=1}^{m} P(y|x; w)$$

取对数:

$$L(\hat{y}, y) = -(ylog\hat{y} + (1 - y)log(1 - \hat{y}))$$

成本函数 (cost function): 基于参数的总成本

$$J(w, b) = \frac{1}{m} \sum_{i=1}^{m} L(\hat{y}^{(i)}, y^{(i)})$$

梯度下降法: 目的: 找到参数 w,b 使 J(w,b) 最小, 局部极小值 $(-\nabla F(x))$; 以梯度的正方向迭代将获得局部极大值点 $(\nabla F(x))$

$$w := w - \alpha \frac{\partial J(w, b)}{\partial w}$$

优点: 1. 数学含义上的支撑

2. 可解释性强

3. 工程化需要

简单, 易理解, 上线快, 速度快

缺点： LR 模型中没有特征组合，特征间关系没有被发现；因此把任意两个特征组合，把一个特征组合当作一个新特征融合进去可以得到 FM 模型。

逻辑回归的应用： *Predicting Clicks: Estimating the Click-Through Rate for New Ads*
数据集构造：对于一个广告会有多个竞价词 (term), 数据集包括广告主、广告、竞价词。
根据每个广告的累计曝光次数、累计点击次数达到 CTR。
采用训练集所有广告的平均 CTR
测试集上每个广告的 CTR 和成本函数评估模型的准确性
特征工程
Related Term CTR Feature Set: 在预测某个广告的点击率时，相关竞价词的其他类似广告也能提供帮助
Ad Quality Feature Set: appearance, attention capture, reputation, landing page quality, relevance

POLY2 模型：

辛普森悖论： 分组实验中使用“性别”+“视频”的特征组合，而汇总实验使用“视频 ID”这一单一特征来计算，损失了大量有效信息。

表 1: 男性用户

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

表 2: 女性用户

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

表 3: 数据汇总

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

优点： LR 模型只考虑单一特征，POLY2 模型对每个特征进行了两两组合。

缺点： 样本参数过多，达到百万级别时，产生的特征交叉组合会爆炸性增长，导致无法承担的计算量和内存需求。

特征交叉后使得特征向量更加稀疏；使得训练样本不足导致过拟合。特征向量稀疏的原因：在将信息转换为向量时采用 one-hot 编码，导致存在大量的数值 0，在特征交叉组合后特征维度为 0 的数量会更多，造成特征向量更加稀疏。

$$\emptyset POLY2(w,x) = \sum_{j_1=1}^{n-1} \sum_{j_2=j_1+1}^{n} w_h(j_1,j_2) x_{j_1} x_{j_2}$$

FM 模型：

$$\emptyset FM(w,x) = \sum_{j_1=1}^{n-1} \sum_{j_2=j_1+1}^{n} (w_{j_1} \cdot w_{j_2}) x_{j_1} x_{j_2}$$

优点： 与 POLY2 模型相比，将 POLY2 模型 n^2 级别的权重参数减少到 nk(k 为隐向量维度，$n >> k$)。FM 模型将 POLY2 模型中的权重系数 $w_{h(j_1,j_2)}$ 取代为两个向量的内积 $(w_{j_1} \cdot w_{j_2})$，当数据集非常稀疏时效果好于 POLY2. 可以学习到更多特征之间的交叉关系，提高泛化能力。

FFM 模型：

$$\emptyset FFM(w,x) = \sum_{j_1=1}^{n-1} \sum_{j_2=j_1+1}^{n} (w_{j_1,f_2} \cdot w_{j_2,f_1}) x_{j_1} x_{j_2}$$

优点： 相较于 FM 模型，FFM 模型引入了特征域感知 (filed-aware)，即从 FM 模型中每个特征对应一个隐向量变为对应一组隐向量；每一个特征 x_j 都有对应的域 f。

因此在 FFM 模型训练过程中，需要学习 n 个特征在 f 个域上的 k 维隐向量，复杂度维 kn^2。

缺点： 由于 FFM 模型复杂度高于 FM 模型，需要在实际应用中加以权衡。

FFM 优化-Bi-FFM

$$\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(v_i W v_j)x_i x_j$$

v_i, v_j 分别用一个隐向量来表达，把两个特征交互的信息用一个共享参数矩阵 W 表示，能减少参数量。

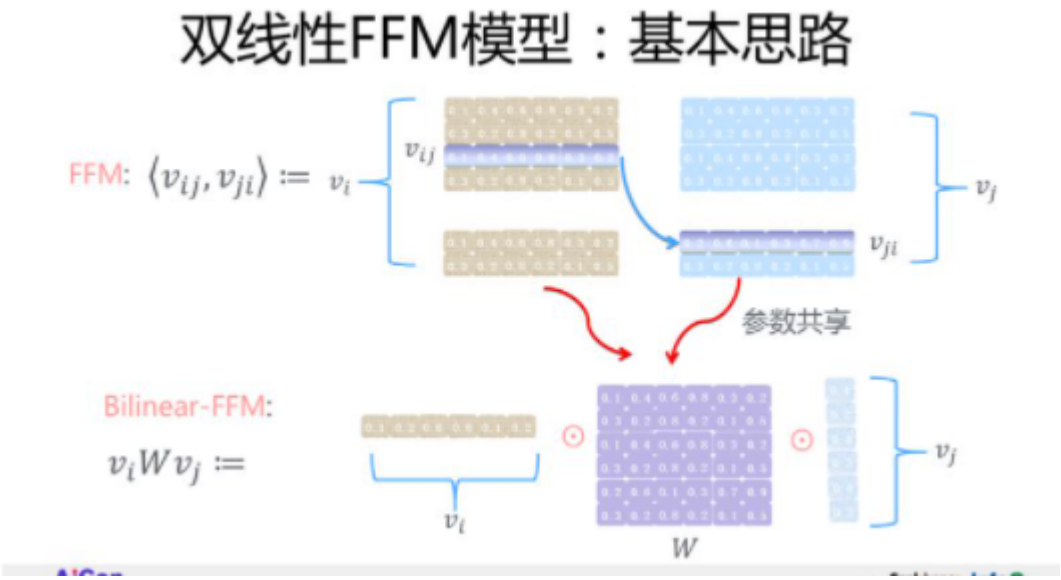

图 1: Bi-FFM