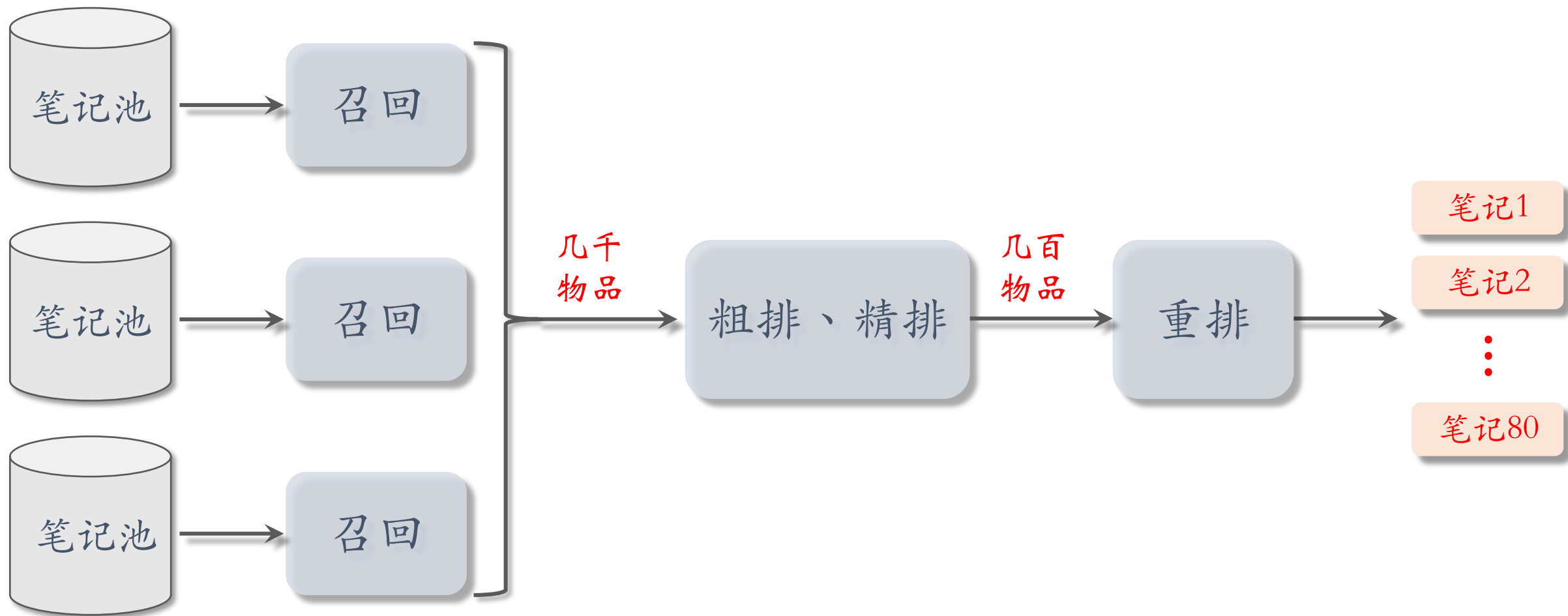


推荐系统的链路



用户—笔记的交互

- 对于每篇笔记，系统记录：
 - 曝光次数 (number of impressions)
 - 点击次数 (number of clicks)
 - 点赞次数 (number of likes)
 - 收藏次数 (number of collects)
 - 转发次数 (number of shares)

用户—笔记的交互

➡ • 点击率 = 点击次数 / 曝光次数

➡ • 点赞率 = 点赞次数 / 点击次数

➡ • 收藏率 = 收藏次数 / 点击次数

➡ • 转发率 = 转发次数 / 点击次数

排序的依据

- 排序模型预估点击率、点赞率、收藏率、转发率等多种分数。
- 融合这些预估分数。（比如加权和。）
- 根据融合的分數做排序、截断。

多目标模型



用户特征



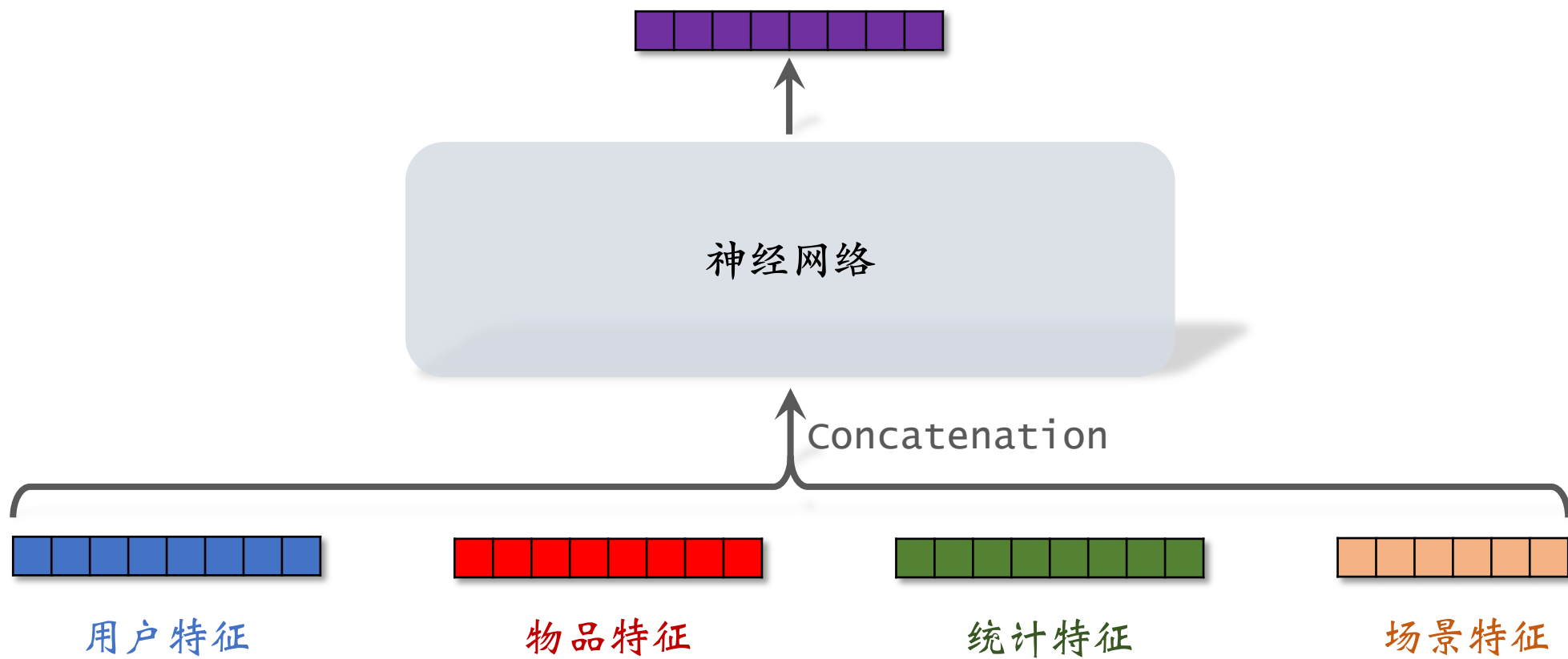
物品特征

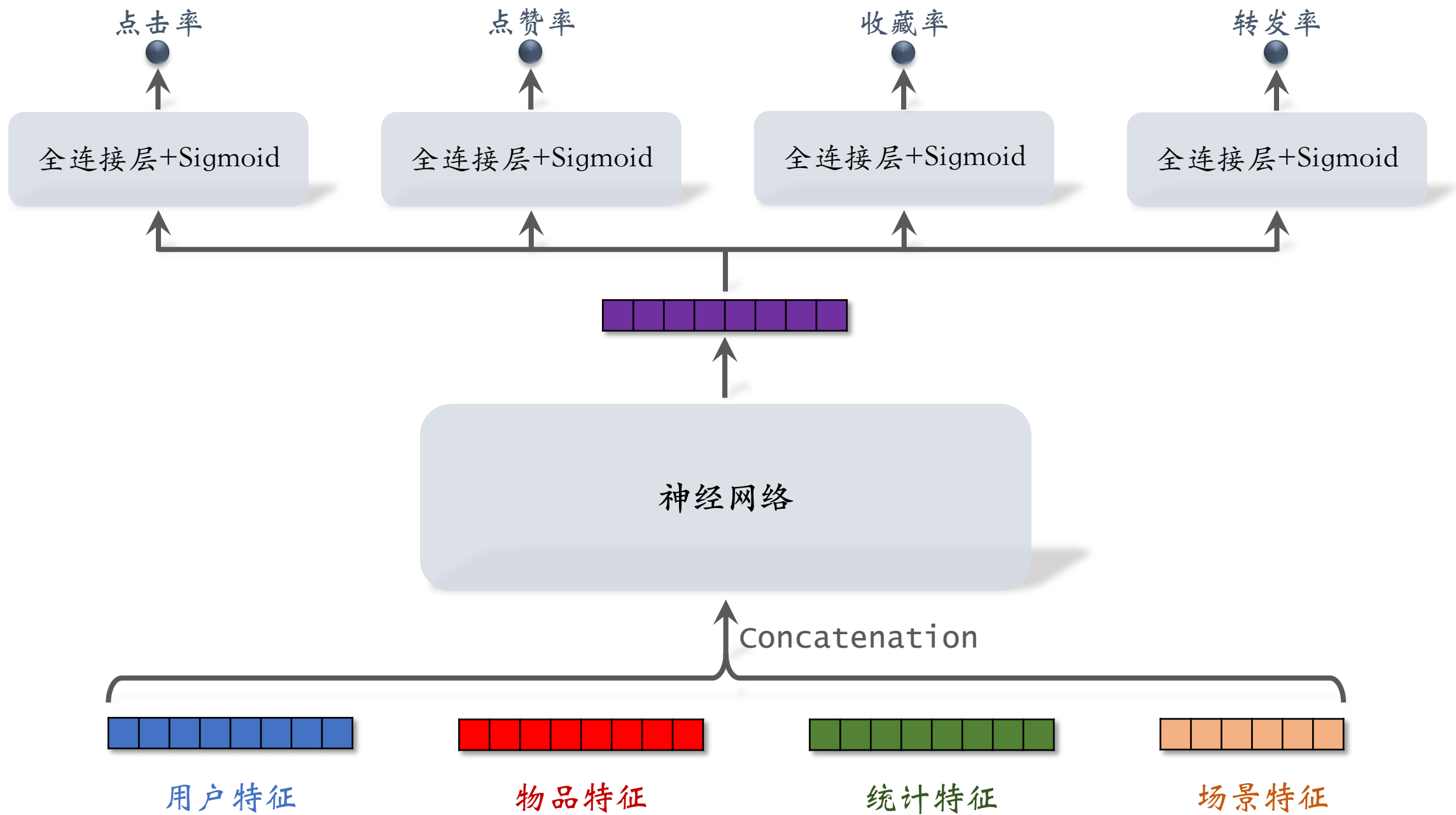


统计特征



场景特征





预估:

点击率



p_1

点赞率



p_2

收藏率



p_3

转发率



p_4

预估:

点击率

 p_1

点赞率

 p_2

收藏率

 p_3

转发率

 p_4

目标:


 y_1


 y_2


 y_3


 y_4

1

有点击

0

无点赞

0

无收藏

1

有转发


预估:

点击率
●
 p_1

点赞率
●
 p_2

收藏率
●
 p_3

转发率
●
 p_4


$$\text{CrossEntropy}(\mathbf{y}_1, \mathbf{p}_1) = -(\mathbf{y}_1 \cdot \ln \mathbf{p}_1 + (1 - \mathbf{y}_1) \cdot \ln(1 - \mathbf{p}_1))$$

目标:

●
 y_1

●
 y_2

●
 y_3

●
 y_4

训练:

- 总的损失函数: $\sum_{i=1}^4 \alpha_i \cdot \text{CrossEntropy}(\mathbf{y}_i, \mathbf{p}_i)$ 。
- 对损失函数求梯度, 做梯度下降更新参数。

训练

- 困难：类别不平衡。
 - 每100次曝光，约有10次点击、90次无点击。
 - 每100次点击，约有10次收藏、90次无收藏。
- 解决方案：负样本降采样 (down-sampling) 。
 - 保留一小部分负样本。
 - 让正负样本数量平衡，节约计算。

注：不是小红书真实数据

预估值校准

预估值校准

- 正样本、负样本数量为 n_+ 和 n_- 。
- 对负样本做降采样，抛弃一部分负样本。
- 使用 $\alpha \cdot n_-$ 个负样本， $\alpha \in (0, 1)$ 是采样率。
- 由于负样本变少，**预估点击率**大于**真实点击率**。

预估值校准

- 真实点击率： $p_{\text{true}} = \frac{n_+}{n_+ + n_-}$ （期望）。
- 预估点击率： $p_{\text{pred}} = \frac{n_+}{n_+ + \alpha \cdot n_-}$ （期望）。

预估值校准

- 真实点击率： $p_{\text{true}} = \frac{n_+}{n_+ + n_-}$ （期望）。
- 预估点击率： $p_{\text{pred}} = \frac{n_+}{n_+ + \alpha \cdot n_-}$ （期望）。
- 由上面两个等式可得校准公式[1]：

$$\underline{p_{\text{true}}} = \frac{\alpha \cdot p_{\text{pred}}}{(1 - p_{\text{pred}}) + \alpha \cdot p_{\text{pred}}}$$

参考文献：

1. Xinran He et al. [Practical lessons from predicting clicks on ads at Facebook](#). In *the 8th International Workshop on Data Mining for Online Advertising*.

Thank You!

<http://wangshusen.github.io/>