# Contrastive Learning

## Data Preparation Methods

### Pointwise Match

- Data Point: Each sample consists of a user-item pair and its label (positive/negative, 1/0).
- Task: This is a binary classification task, goal is to predict a similarity which is close to 1 for positive pair and close to 0 for negative pair.
- Sampling Strategy: Maintain a positive-to-negative sample ratio, typically 1:2 or 1:3.

### Pairwise Match

- Data Point: Each sample consists of a user, a positive item, and a negative item.
- Task: The goal is to make the similarity for the positive pair is larger than that of the negative pair.
- Loss: hinge loss $\max(0, \text{margin} + s_{\text{neg}} - s_{\text{pos}})$

### Listwise Match

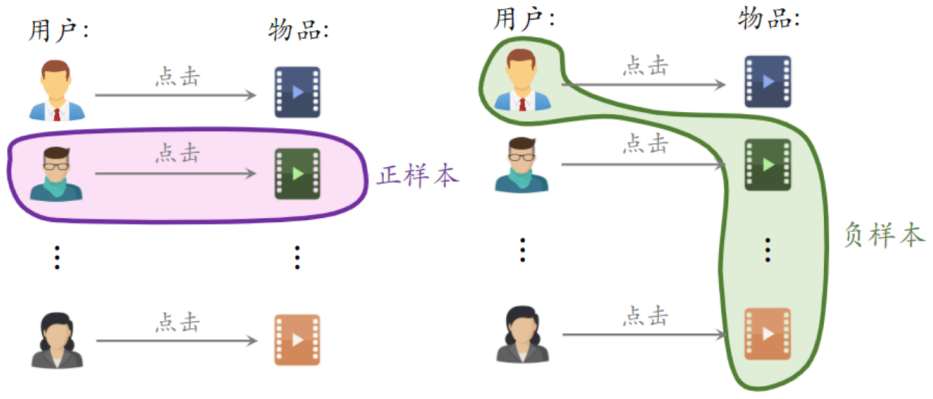
- Data Point: Each sample consists of a user, a positive item, and a list of negative items.
- Task: Each user-item pair is processed independently by the model, but a softmax is applied over all similarity scores to normalize them. The goal is to maximize the score for the positive pair while minimizing the scores for the negatives.
- Loss: Cross-entropy loss, as the positive pair is labeled as 1 and the negatives as 0. This essentially corresponds to −log(softmaxed positive-pair score)$= -\log\left(\frac{\exp(s_{\text{pos}})}{\exp(s_{\text{pos}})+\sum_{i=1}^{n}\exp(s_{\text{neg},i})}\right)$

## Lack Positive Samples

- Under Sampling: removing negative samples.  Over Sampling: duplicating positive samples.
- Weighted Loss: Assigning higher weights to loss of positive samples.
- Use list-match approach: a single positive sample can be paired with multiple negatives, maximizing its use.
- Use hard negative samples as negative: By selecting negative samples whose features are close to those of positive samples, it enhances the model's ability to distinguish between positive and negative samples.

## Sample Negative Samples

- Simple Negative Samples
  - Uniform Sampling: The issue is that most items are long-tail (unpopular). If sampled uniformly, the negative samples will mostly be unpopular items, which is unfair to unpopular items.
  - Non-Uniform Sampling: the probability of being negative items is proportional to their popularity (e.g., number of clicks). 多选 popular items
  - Batch-Based Sampling: Suppose there are $n$ positive samples in a batch, each user is paired with $n-1$ items to form negative samples. This results in a total of $n(n-1)$ negative samples in the batch, all simple negative samples.
    - Problem: The probability of an item appearing in the batch is proportional to its popularity, so popular items are more likely to become negative samples.  对热门物品打压太狠了，容易造成偏差，




- Hard Negative Samples
  - Definition：Items eliminated during the scoring stage (粗排，relatively hard negatives). Items with low scores in the scoring stage (精排，extremely hard negatives).

# Evaluation

## Online Evaluation Strategy

除了A/B testing，还有shallow deployment，interleaving，canary release，multi-armed bandit

- A/B Testing
  - A/B testing involves splitting users into two groups: one group sees the results generated by the old model (control group), and the other sees results from the new model (treatment group). User interactions from both groups are compared to determine which model performs better.
  - [ 缺 ] requires a larger user base
- Shadow Deployment
  - runs a new model in the background, alongside the production model, but does not expose its results to users. The new model's predictions are logged for analysis and comparison.
  - effective for testing system integration and scalability.
  - [ 优 ] safe: no direct impact on users
- Canary Release
  - gradually rolls out the new model to a small subset of users, starting with a small percentage, and slowly increasing the exposure as the model proves reliable.
  - [ 优 ]low-risk: if issues arise, the new model can be rolled back quickly.
- Interleaving Experiments
  - both the old and new models make predictions for the same user, and the results are mixed within the same recommendation list. User interactions help determine which model performs better.
  - [ 优 ] direct comparison: Both models are tested under identical conditions.
  - [ 缺 ] limited to scenarios where interleaving predictions make sense (e.g., search results, recommendations).
- Multi-Armed Bandits
  - dynamically allocate traffic between multiple models based on their real-time performance, favoring the model that performs best while still exploring others.
  - [ 优 ] balances exploration (testing new models) and exploitation (using the best-performing model).
  - [ 优 ] quickly adapts to changing user preferences or environments.
  - [ 缺 ] May miss potential improvements (有些模型的好表现有延时，不能及时反应).

## Offline Evaluation Strategy

- **K-Fold Cross Validation**
  - splits the data into K equal folds, where each fold serves as a validation set once while the remaining folds are used for training. This process is repeated K times, and the overall performance is averaged across all folds.
  - [ 优 ] a robust validation of performance
  - [ 优 ] particularly good for moderate-sized datasets
  - [ 优 ] ensures that every data contributes to both training and validation, minimizing information leakage and reducing the impact of data splits.
- Hold-out Validation
  - It divides the data into training and test sets (e.g., 80/20 or 70/30)
  - [适用] large datasets

    - [ 优 ] simple and efficient
    - [ 缺 ] be cautious of the potential variance introduced by a single random split.
- **Bootstrap**
    - Bootstrap involves sampling the training data with replacement to generate multiple bootstrap samples, each of the same size as the original dataset. The model is then trained on each bootstrap sample and evaluated on the remaining unsampled data, known as the out-of-bag (OOB) data. By repeating this process many times, we can compute the average performance and obtain confidence intervals for the model's metrics, which is particularly useful when assessing the model’s uncertainty.
    - [ 优 ] works for small datasets: allows fully utilize the available data without needing to reserve a large portion for testing.
    - [ 优 ] By providing confidence intervals, it also gives deeper insights into the range of potential performance outcomes,
    - [ 优 ] can exam model's sensitivity to variations in the training data.

# Offline Evaluation Metrics

## Metrics Selection

- Recommendation: (1) Precision@K+Recall@K+mAP，(2) NDCG@K，(3) Intra-List Diversity+Gini Index
- Search: Mean Reciprocal Rank + NDCG@K

  Relevance metric not good for ranking: 只是排序，不一定有 relevant item，也可能 relevant item 太多
    - precision@k and mAP: when the number of relevant item for query is small (e.g., 1), this metric is always low.
    - recall@k：the total number of the relevant item can be very high, leading to low recall.
- Anomaly detection：Precision, Recall, F1 score, PR-AUC (ROC-AUC)

## Metrics Introduction

- **Precision@K** [ Relevance ]
    - The proportion of relevant items in the top K recommendations.
    - 关注推荐系统的准确性，确保推荐的都是用户感兴趣的
- **Recall@K** [ Relevance ]
    - The proportion of the user’s relevant items that appear in the top K recommendations.
    - What it evaluates: relevance 强调覆盖性，保证用户的兴趣都被考虑了
- **mAP (Mean Average Precision)** [ Relevance + Ranking Quality ]
    - mAP is the mean of the AP (Average Precision) scores across all users or queries.

      AP is  the average of Precision@k at all positions where relevant items appear in the recommendation.
    - 缺点：mAP is for binary relevance, NDCG is for continuous relevance. (binary：one item is considered either relevant or not; continuous: relevant degree can be 0, 1, 2 .. 5)
- **NDCG@K (Normalized Discounted Cumulative Gain)** [ Ranking Quality ]
    - evaluates the quality of the ranking by assigning higher weights to the relevance scores of items that appear earlier in the recommendation list. IDCG@K is the ideal number.
    - 关注相关项目的位置，对高度相关的项目进行排名更高的推荐奖励。$\frac{1}{\log_2(i+1)}$表示这个位置每个球值几分，$\mathrm{rel}_i$表示拿了几个球。
    - Example: true relevance score of the ranked items [0, 5, 1, 4, 2]

      $$\mathrm{DCG}_k = \sum_{i=1}^{k} \frac{\mathrm{rel}_i}{\log_2(i+1)} = \frac{0}{\log_2(2)} + \frac{5}{\log_2(3)} + \frac{1}{\log_2(4)} + \frac{4}{\log_2(5)} + \frac{2}{\log_2(6)} = 6.151$$

$$\text{IDCG}_k = \sum_{i=1}^{k} \frac{\text{rel}_i}{\log_2(i+1)} = \frac{5}{\log_2(2)} + \frac{4}{\log_2(3)} + \frac{2}{\log_2(4)} + \frac{1}{\log_2(5)} + \frac{0}{\log_2(6)} = 8.9543$$

$$\text{nDCG}_k = \frac{\text{DCG}_k}{\text{IDCG}_k} = \frac{6.151}{8.9543} = 0.6869$$

    - 缺点：ground truth relevance scores is not always available.
- **Mean Reciprocal Rank (MRR)** [ Ranking Quality ]
    - averages the rank of the first relevant item across all searches.
    - $\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\text{rank}_i}$ ， $\text{rank}_i$ is the rank of the first relevant item in the $i^{\text{th}}$ output list.
- **Intra-List Diversity (ILD)** [ Diversity ]
    - measures how different the items in the recommendation list are from each other. It evaluates the recommendation variety，helping to prevent repetitive or overly similar content.
    - 推荐多样性，推荐的东西之间区别有多大
- **Gini Index** [ Diversity ]
    - measures the balance in the distribution of recommendations across all items in the system. It fairness and diversity of item exposure, helping to prevent over-recommendation of popular content.
    - 项目曝光度，对 item pool的使用有多均匀
- **Coverage** [ Diversity ]
    - User Coverage measure the proportion of users receiving relevant recommendations. (Item Coverage measures the proportion of items recommended.)
    - 另一种曝光度，user coverage看给多少user推荐过。

# Online Evaluation Metrics

- Click-through rate: The ratio between the total number of click and the total number of impressions (recommended videos). but it cannot capture or measure clickbait.
- [Time] Total (watch) time. spend on the timeline during a fixed period, such as 1 week.
- [Item] Complete of Item: item completion rate
- [Feedback] Explicit user feedback rate.
- [Money] Conversion rate:  Number of conversions /Number of impressions, Revenue lift. This measures the percentage of revenue increase over time.