AGENDA

1. Preference Learning Tasks
2. **Performance Assessment and Loss Functions**
   a. **Evaluation of Rankings**
   b. Weighted Measures
   c. Evaluation of Bipartite Rankings
   d. Evaluation of Partial Rankings
3. Preference Learning Techniques
4. Complexity of Preference Learning
5. Conclusions

Rank Evaluation Measures

- In the following, we do not discriminate between different ranking scenarios
    - we use the term items for both, objects and labels

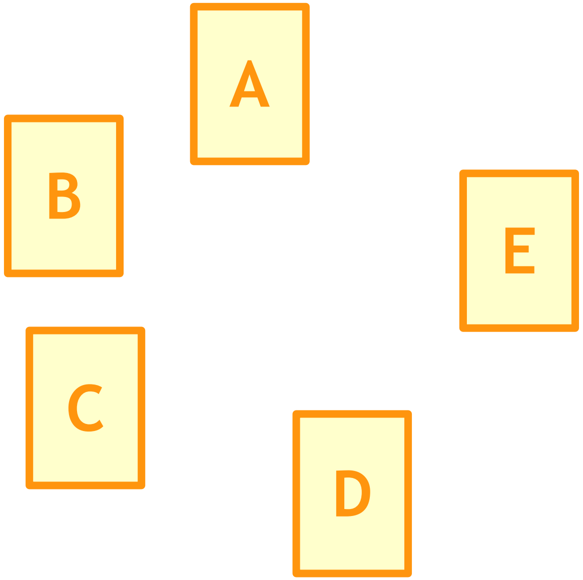
- All measures are applicable to both scenarii
    - sometimes have different names according to context

- Label Ranking
    - measure is applied to the ranking of the labels of each examples
    - averaged over all examples
- Object Ranking
    - measure is applied to the ranking of a set of objects
    - we may need to average over different sets of objects which have disjoint preference graphs
        - e.g. different sets of query / answer set pairs in information retrieval

Ranking Errors

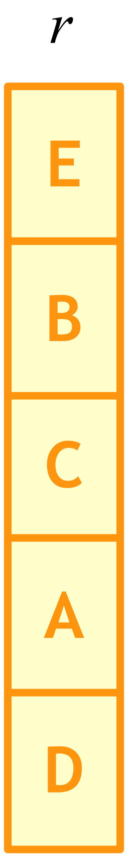
- Given:
  - a set of items $X = \{x_1, \ldots, x_c\}$ to rank
    - Example:
      $X = \{$A, B, C, D, E$\}$

items can be objects or labels



Ranking Errors

- Given:
  - a set of items $X = \{x_1, \ldots, x_c\}$ to rank
    - *Example:*
      X = {A, B, C, D, E}
  - a target ranking r
    - *Example*:
      E $\succ$ B $\succ$ C $\succ$ A $\succ$ D



r

E

B

C

A

D

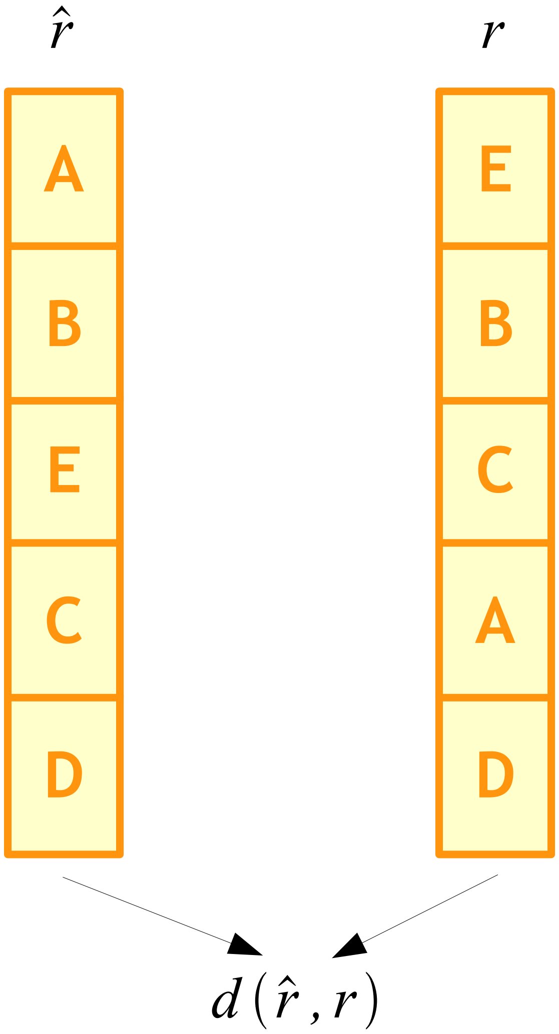
Ranking Errors

- Given:
  - a set of items $X = \{x_1, \ldots, x_c\}$ to rank
    - *Example:*

      X = {A, B, C, D, E}
  - a target ranking r
    - *Example*:

      E $\succ$ B $\succ$ C $\succ$ A $\succ$ D
  - a predicted ranking $\hat{r}$
    - *Example*:

      A $\succ$ B $\succ$ E $\succ$ C $\succ$ D
- Compute:
  - a value $d(r, \hat{r})$ that measures the *distance* between the two rankings



Notation

- r and $\hat{r}$ are functions from $X \rightarrow \mathbb{N}$
  - returning the rank of an item x

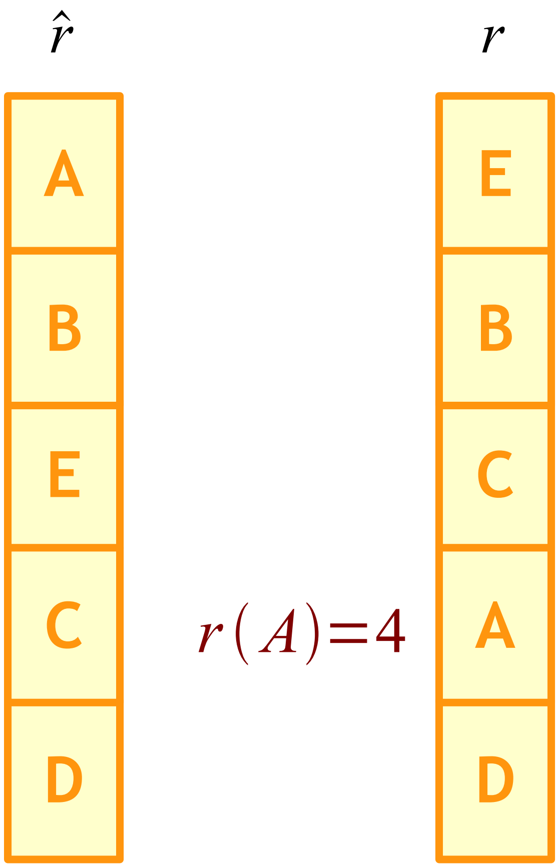
$$\hat{r}(A)=1$$

- the inverse functions r^{-1}: $\mathbb{N} \rightarrow X$
  - return the item at a certain position

$$\hat{r}^{-1}(1)=A \qquad r^{-1}(4)=A$$

- as a short-hand for $r \circ \hat{r}^{-1}$, we also define function R: $\mathbb{N} \rightarrow \mathbb{N}$
  - $R(i)$ returns the true rank of the i-th item in the predicted ranking

$$R(1)=r(\hat{r}^{-1}(1))=4$$



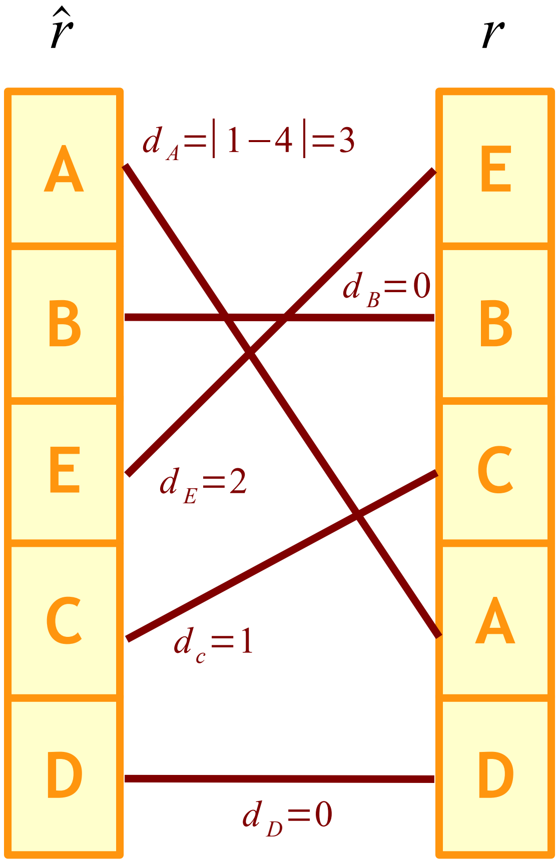
$$r(A)=4$$

Spearman's Footrule

- Key idea:
  - Measure the sum of absolute differences between ranks

$$D_{SF}(r,\hat{r}) = \sum_{i=1}^{c} |r(x_i) - \hat{r}(x_i)| = \sum_{i=1}^{c} |i - R(i)|$$

$$= \sum_{i=1}^{c} d_{x_i}(r,\hat{r})$$



$$\sum_{x_i} d_{x_i} = 3+0+1+0+2 = 6$$

Spearman Distance

- Key idea:
  - Measure the sum of ~~absolute~~ squared differences between ranks

$$D_S(r,\hat{r}) = \sum_{i=1}^{c} (r(x_i) - \hat{r}(x_i))^2 = \sum_{i=1}^{c} (i - R(i))^2$$

$$= \sum_{i=1}^{c} d_{x_i}(r,\hat{r})^2$$

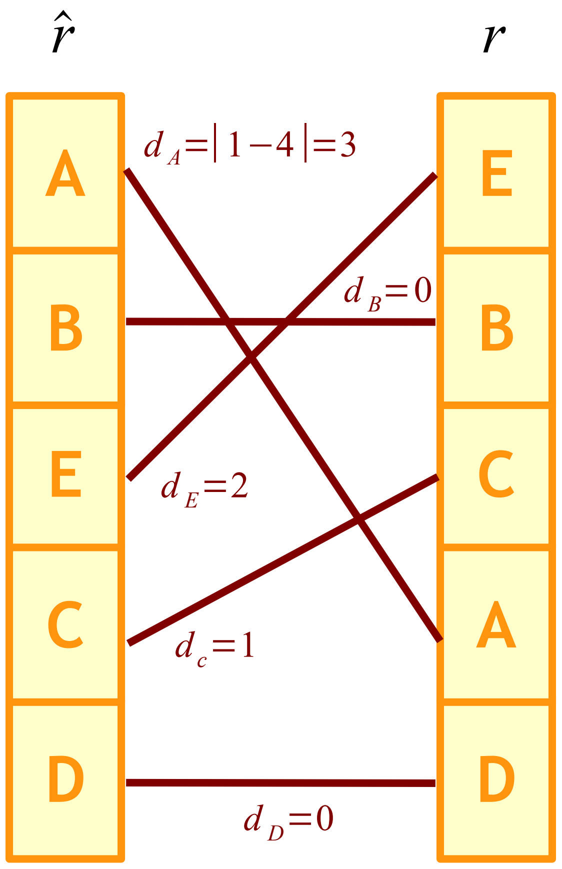
- Value range:

$$\min D_S(r,\hat{r}) = 0$$

$$\max D_S(r,\hat{r}) = \sum_{i=1}^{c} ((c-i)-i)^2 = \frac{c\cdot(c^2-1)}{3}$$

→ Spearman Rank Correlation Coefficient

$$1 - \frac{6\cdot D_S(r,\hat{r})}{c\cdot(c^2-1)} \in [-1,+1]$$

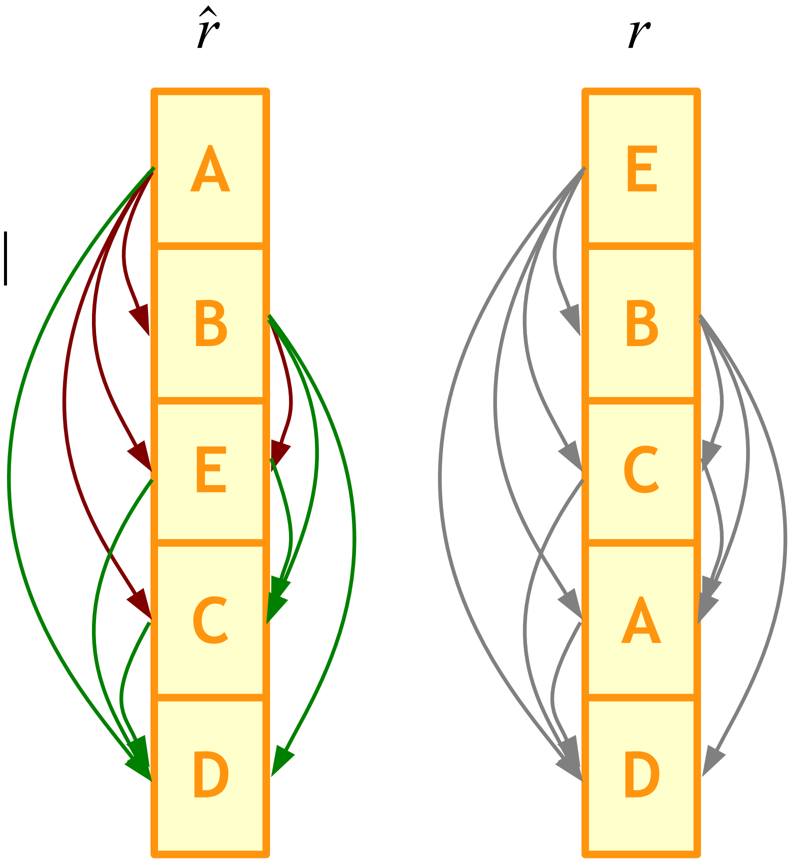


$$\sum_{x_i} d_{x_i}^2 = 3^2 + 0 + 1^2 + 0 + 2^2 = 14$$

Kendall's Distance

- Key idea:
  - number of item pairs that are inverted in the predicted ranking

$$D_\tau(r,\hat{r}) = \left|\left\{(i,j) \mid r(x_i) < r(x_j) \wedge \hat{r}(x_i) > \hat{r}(x_j)\right\}\right|$$



$$D_\tau(r,\hat{r}) = 4$$

- Value range:

$$\min D_\tau(r,\hat{r}) = 0$$

$$\max D_\tau(r,\hat{r}) = \frac{c\cdot(c-1)}{2}$$

→ Kendall's tau

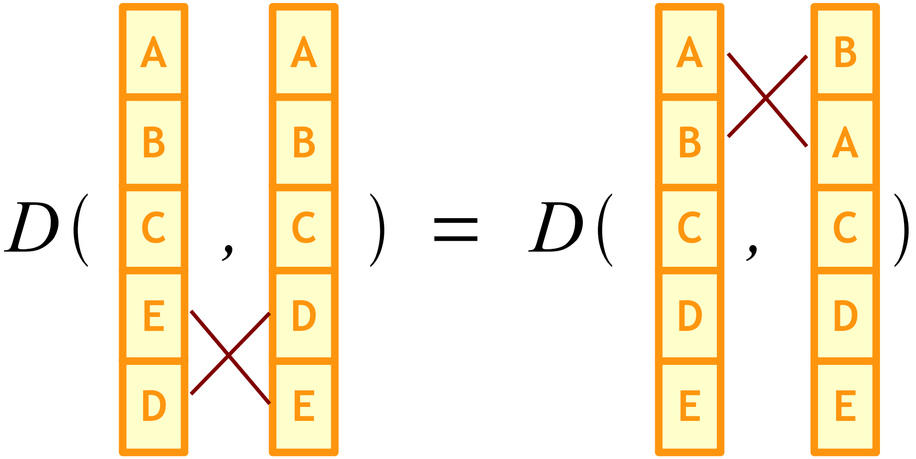
$$1 - \frac{4\cdot D_\tau(r,\hat{r})}{c\cdot(c-1)} \in [-1,+1]$$

AGENDA

1. Preference Learning Tasks
2. **Performance Assessment and Loss Functions**
   a. Evaluation of Rankings
   b. Weighted Measures
   c. Evaluation of Bipartite Rankings
   d. Evaluation of Partial Rankings
3. Preference Learning Techniques
4. Complexity of Preference Learning
5. Conclusions

Weighted Ranking Errors

- The previous ranking functions give equal weight to all ranking positions
  - i.e., differences in the first ranking positions have the same effect as differences in the last ranking positions



- In many applications this is not desirable
  - ranking of search results
  - ranking of product recommendations
  - ranking of labels for classification
  - ...

⇒ Higher ranking positions should be given more weight

Position Error

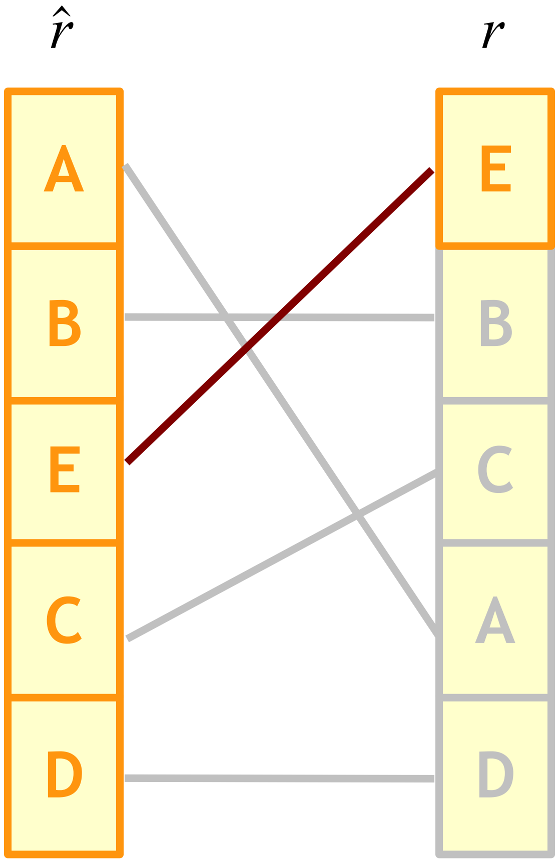
- Key idea:
  - in many applications we are interested in providing a ranking where the target item appears a high as possible in the predicted ranking
    - e.g. ranking a set of actions for the next step in a plan
  - Error is the number of wrong items that are predicted before the target item

$$D_{PE}(r,\hat{r})=\hat{r}(\arg\min_{x\in X} r(x))-1$$

- Note:
  - equivalent to Spearman's footrule with all non-target weights set to 0

$$D_{PE}(r,\hat{r})=\sum_{i=1}^{c} w_i \cdot d_{x_i}(r,\hat{r})$$

with $w_i=[\![x_i=\arg\min_{x\in X} r(x)]\!]$



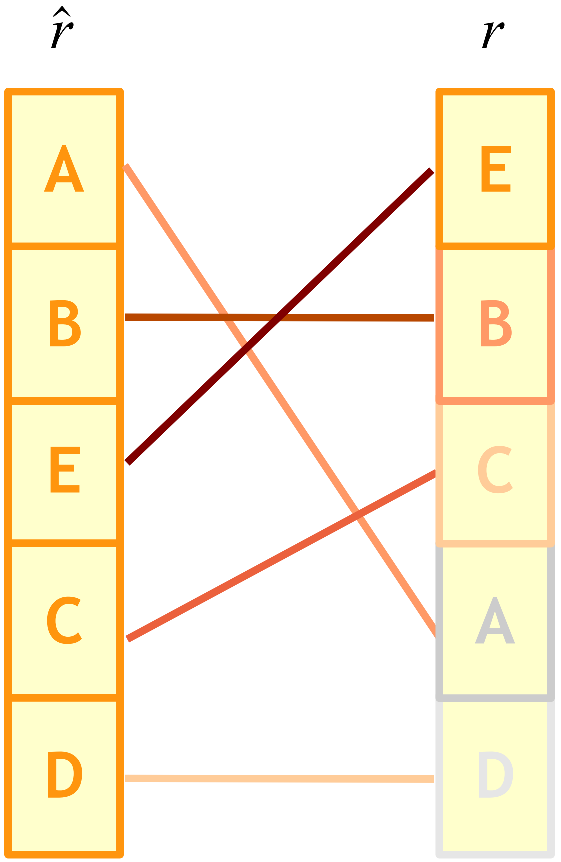
$$D_{PE}(r,\hat{r})=2$$

Discounted Error

- Higher ranks in the target position get a higher weight than lower ranks

$$D_{DR}(r,\hat{r})=\sum_{i=1}^{c} w_i \cdot d_{x_i}(r,\hat{r})$$

with $w_i = \dfrac{1}{\log(r(x_i)+1)}$



$$D_{DR}(r,\hat{r})=\frac{3}{\log 2}+0+\frac{1}{\log 4}+0+\frac{2}{\log 6}$$

(Normalized) Discounted Cumulative Gain

- a “positive” version of discounted error: Discounted Cumulative Gain (DCG)

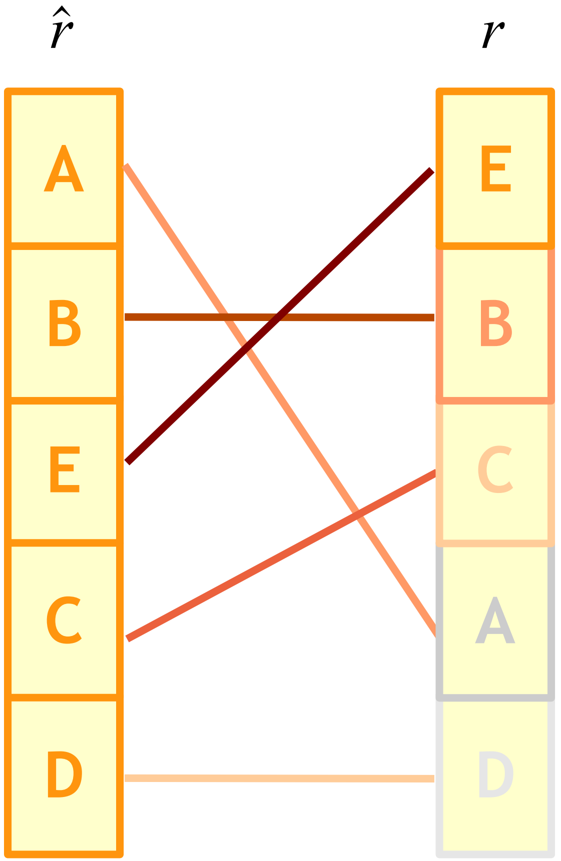
$$DCG(r,\hat{r}) = \sum_{i=1}^{c} \frac{c - R(i)}{\log(i+1)}$$

- Maximum possible value:
  - the predicted ranking is correct, i.e. $\forall i : i = R(i)$
  - Ideal Discounted Cumulative Gain (IDCG)

$$IDCG = \sum_{i=1}^{c} \frac{c-i}{\log(i+1)}$$

- Normalized DCG (NDCG)

$$NDCG(r,\hat{r}) = \frac{DCG(r,\hat{r})}{IDCG}$$



$$NDCG(r,\hat{r}) = \frac{\frac{1}{\log 2} + \frac{3}{\log 3} + \frac{4}{\log 4} + \frac{2}{\log 5} + \frac{0}{\log 6}}{\frac{4}{\log 2} + \frac{3}{\log 3} + \frac{2}{\log 4} + \frac{1}{\log 5} + \frac{0}{\log 6}}$$

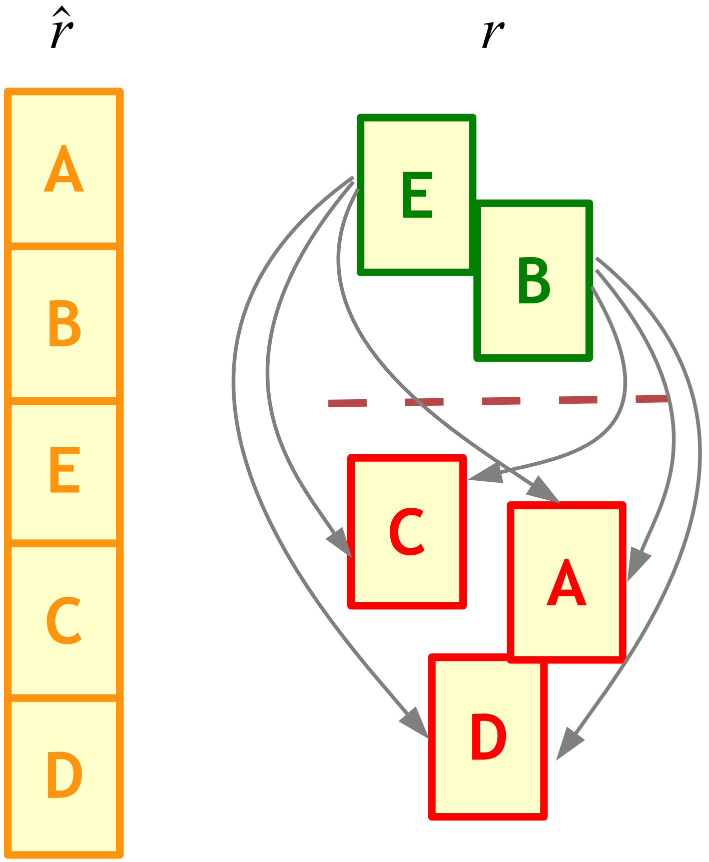
AGENDA

1. Preference Learning Tasks
2. **Performance Assessment and Loss Functions**
   a. Evaluation of Rankings
   b. Weighted Measures
   c. Evaluation of Bipartite Rankings
   d. Evaluation of Partial Rankings
3. Preference Learning Techniques
4. Complexity of Preference Learning
5. Conclusions

Bipartite Rankings

Bipartite Rankings

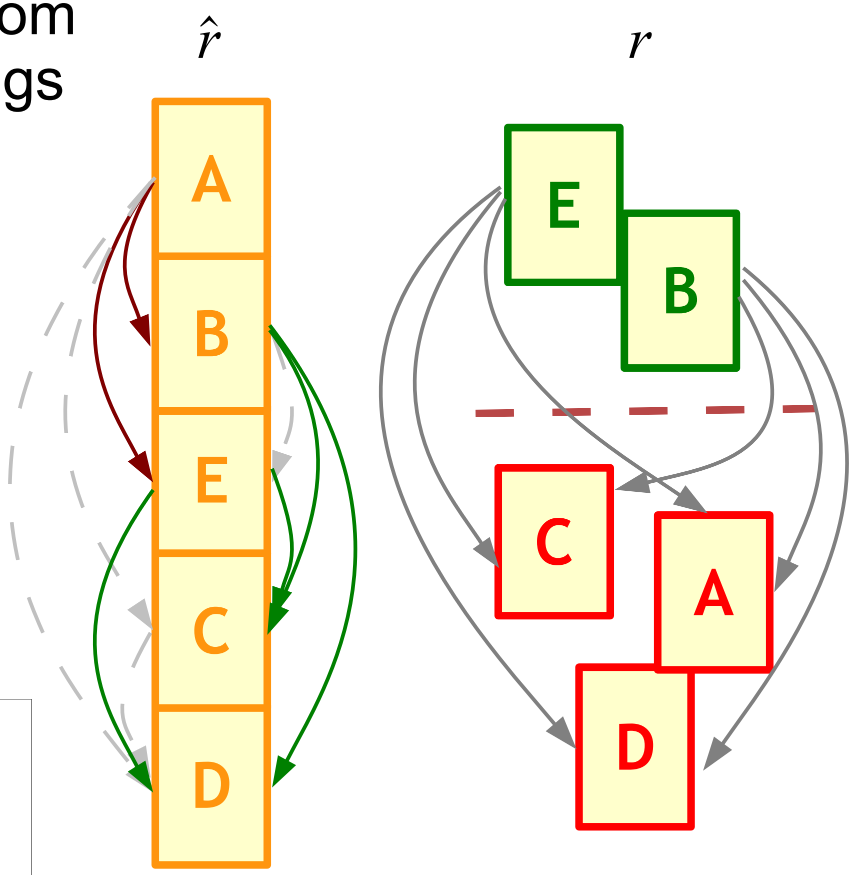
- The target ranking is not totally ordered but a *bipartite graph*
- The two partitions may be viewed as preference levels $L = \{0, 1\}$
  - all c_1 items of level 1 are preferred over all c_0 items of level 0
- We now have fewer preferences
  - for a total order: $\frac{c}{2} \cdot (c-1)$
  - for a bipartite graph: $c_1 \cdot (c - c_1)$



Evaluating Partial Target Rankings

- Many Measures can be directly adapted from total target rankings to partial target rankings
- Recall: Kendall's distance
  - number of item pairs that are inverted in the target ranking

    $$D_\tau(r,\hat{r}) = |\{(i,j) \mid r(x_i) < r(x_j) \wedge \hat{r}(x_i) > \hat{r}(x_j)\}|$$
  - can be directly used
  - in case of normalization, we have to consider that fewer items satisfy $r(x_i) < r(x_j)$
- Area under the ROC curve (AUC)
  - the AUC is the fraction of pairs of (p,n) for which the predicted score $s(p) > s(n)$
    - Mann Whitney statistic is the absolute number
  - This is 1 - normalized Kendall's distance for a bipartite preference graph with $L = \{p,n\}$



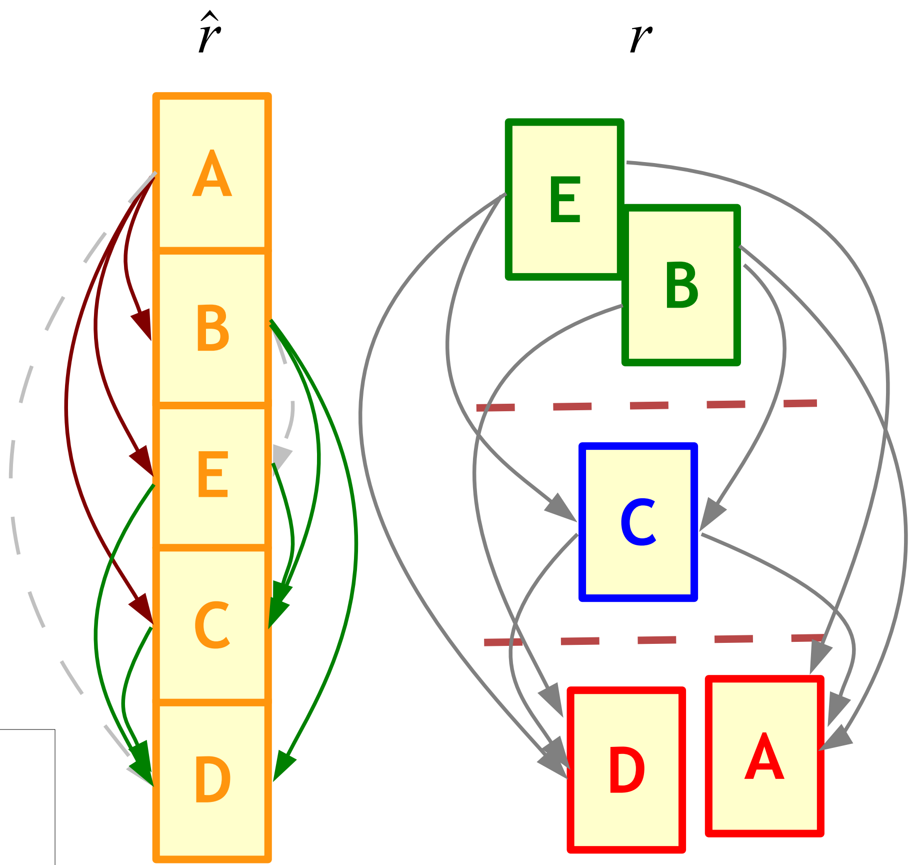
$$D_\tau(r,\hat{r}) = 2$$

$$AUC(r,\hat{r}) = \frac{4}{6}$$

Evaluating Multipartite Rankings

- Multipartite rankings:
  - like Bipartite rankings
  - but the target ranking r consists of *multiple* relevance levels $L = \{1 \ldots l\}$, where $l < c$
  - total ranking is a special case where each level has exactly one item
- # of preferences $= \sum_{(i,j)} c_i \cdot c_j \leq \frac{c^2}{2} \cdot (1 - \frac{1}{l})$
  - c_i is the number of items in level I
- C-Index [Gnen & Heller, 2005]
  - straight-forward generalization of AUC
  - fraction of pairs (x_i, x_j) for which

$$l(i) > l(j) \wedge \hat{r}(x_i) < \hat{r}(x_j)$$



$$D_\tau(r, \hat{r}) = 3$$

$$\text{C-Index}(r, \hat{r}) = \frac{5}{8}$$

Evaluating Multipartite Rankings

C-Index

- the C-index can be rewritten as a weighted sum of pairwise AUCs:

$$\text{C-Index}(r,\hat{r}) = \frac{1}{\sum_{i,j>i} c_i \cdot c_j} \sum_{i,j<i} c_i \cdot c_j \cdot \text{AUC}(r_{i,j}, \hat{r}_{i,j})$$

where $r_{i,j}$ and $\hat{r}_{i,j}$ are the rankings r and $\hat{r}$ restricted to levels i and j.

Jonckheere-Terpstra statistic

- is an *unweighted* sum of pairwise AUCs:

$$\text{m-AUC} = \frac{2}{l \cdot (l-1)} \sum_{i,j>i} \text{AUC}(r_{i,j}, \hat{r}_{i,j})$$

Note:
C-Index and m-AUC can be optimized by optimization of pairwise AUCs

- equivalent to well-known multi-class extension of AUC
[Hand & Till, MLJ 2001]

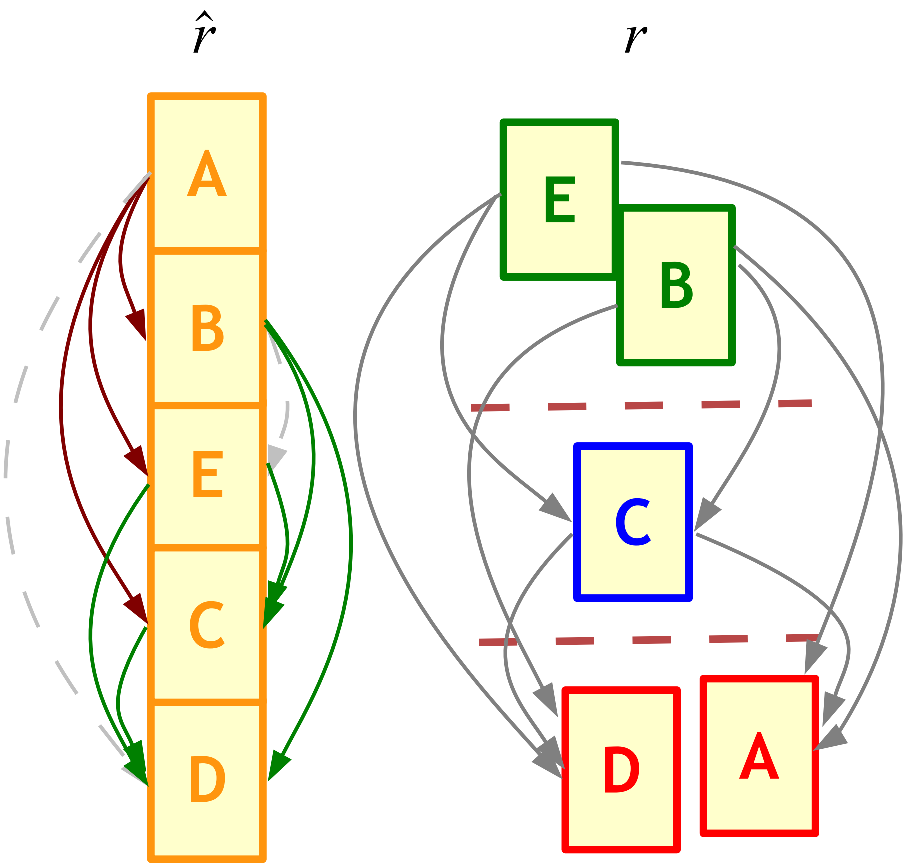
Normalized Discounted Cumulative Gain

[Jarvelin & Kekalainen, 2002]

- The original formulation of (normalized) discounted cumulative gain refers to this setting

$$DCG(r,\hat{r}) = \sum_{i=1}^{c} \frac{l(i)}{\log(i+1)}$$

  - the sum of the true (relevance) levels of the items
  - each item weighted by its rank in the predicted ranking
- Examples:
  - retrieval of relevant or irrelevant pages
    - 2 relevance levels
  - movie recommendation
    - 5 relevance levels



AGENDA

1. Preference Learning Tasks
2. **Performance Assessment and Loss Functions**
   a. Evaluation of Rankings
   b. Weighted Measures
   c. Evaluation of Bipartite Rankings
   d. Evaluation of Partial Rankings
3. Preference Learning Techniques
4. Complexity of Preference Learning
5. Conclusions

Evaluating Partial Structures in the Predicted Ranking

- For fixed types of partial structures, we have conventional measures
  - bipartite graphs → binary classification
    - accuracy, recall, precision, F1, etc.
    - can also be used when the items are labels!
      - e.g., accuracy on the set of labels for multilabel classification
  - multipartite graphs → ordinal classification
    - multiclass classification measures (accuracy, error, etc.)
    - regression measures (sum of squared errors, etc.)

- For general partial structures
  - some measures can be directly used on the reduced set of target preferences
    - Kendall's distance, Gamma coefficient
  - we can also use set measures on the set of binary preferences
    - both, the source and the target ranking consist of a set of binary preferences
    - e.g. Jaccard Coefficient
      - size of interesection over size of union of the binary preferences in both sets

Gamma Coefficient

- Key idea: normalized difference between
  - number of correctly ranked pairs (Kendall's distance)

    $$d = D_\tau(r, \hat{r})$$

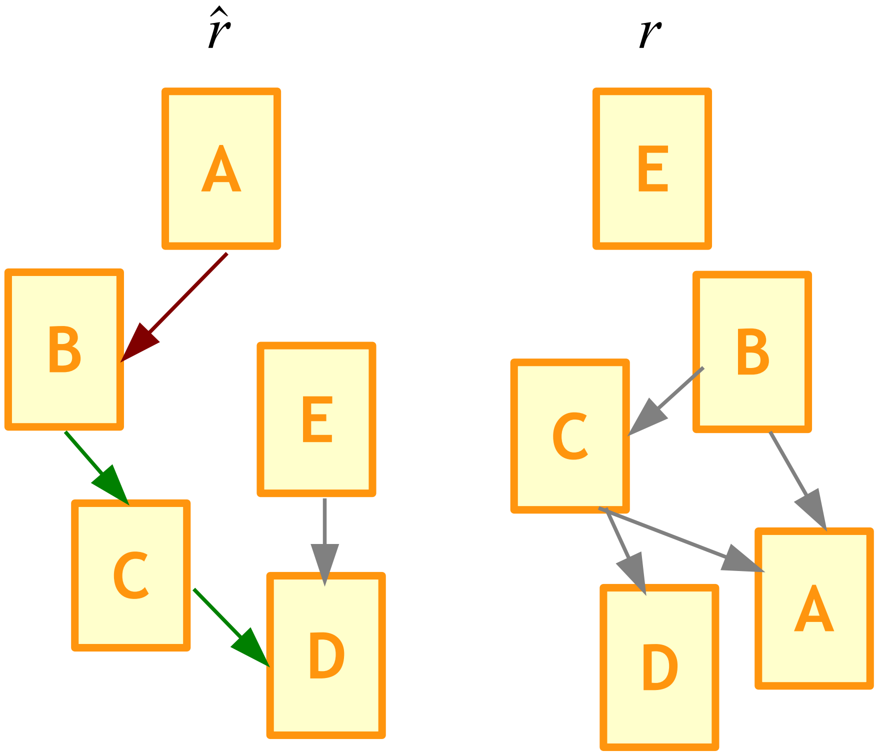
  - number of incorrectly ranked pairs

    $$\bar{d} = \left| \left\{ (i, j) \mid r(x_i) < r(x_j) \wedge \hat{r}(x_i) < \hat{r}(x_j) \right\} \right|$$

- Gamma Coefficient
  [Goodman & Kruskal, 1979]

  $$\gamma(r, \hat{r}) = \frac{d - \bar{d}}{d + \bar{d}} \in [-1, +1]$$

  - Identical to Kendall's tau if both rankings are total
    - i.e., if $d + \bar{d} = \frac{c \cdot (c-1)}{2}$



$$\gamma(r, \hat{r}) = \frac{2-1}{2+1} = \frac{1}{3}$$

References

- Cheng W., Rademaker M., De Baets B., Hüllermeier E.: *Predicting Partial Orders: Ranking with Abstention*. Proceedings ECML/PKDD-10(1): 215-230 (2010)
- Fürnkranz J., Hüllermeier E., Vanderlooy S.: *Binary Decomposition Methods for Multipartite Ranking***.** Proceedings ECML/PKDD-09(1): 359-374 (2009)
- Gnen M., Heller G.: *Concordance probability and discriminatory power in proportional hazards regression*. Biometrika **92**(4):965–970 (2005)
- Goodman, L., Kruskal, W.: *Measures of Association for Cross Classifications*. Springer-Verlag, New York (1979)
- Hand D.J., Till R.J.: *A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems*. Machine Learning **45**(2):171-186 (2001)
- Jarvelin K., Kekalainen J.: *Cumulated gain-based evaluation of IR techniques***.** ACM Transactions on Information Systems **20**(4): 422–446 (2002)
- Jonckheere, A. R.: *A distribution-free k-sample test against ordered alternatives*. Biometrika: 133–145 (1954)
- Kendall, M. *A New Measure of Rank Correlation*. Biometrika 30 (1-2): 81–89 (1938)
- Mann H. B.,Whitney D. R. *On a test of whether one of two random variables is stochastically larger than the other.* Annals of Mathematical Statistics, **18**:50–60 (1947)
- Spearman C. *The proof and measurement of association between two things.* American Journal of Psychology, **15**:72–101 (1904)