The Ranking Poset Framework for Classification and Ranking

Guy Lebanon

Carnegie Mellon University

Introduction

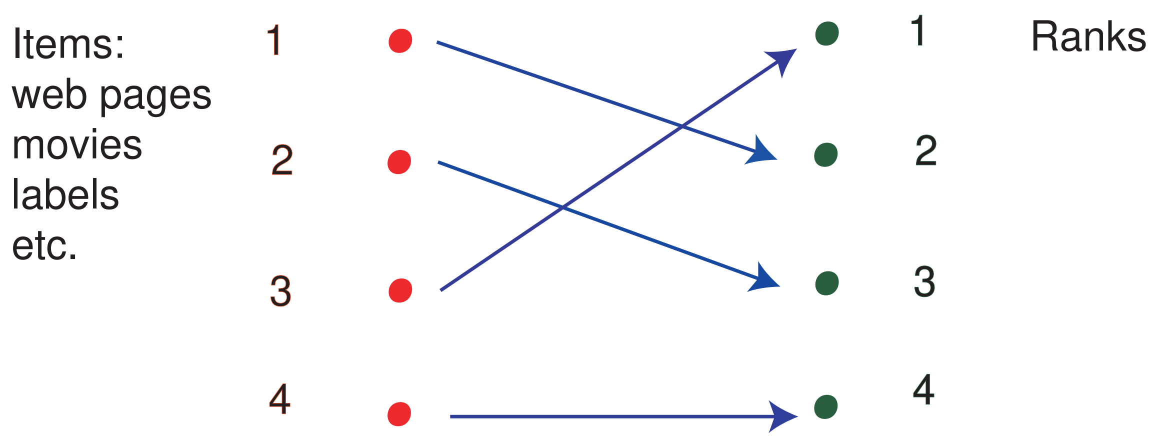
- A conditional model for ranked data predicts preference information over items based on covariates that encode preference information over the same items.

- It is a natural setting for applications such as search engine fusion and collaborative filtering, but may also be used to handle various classification data.

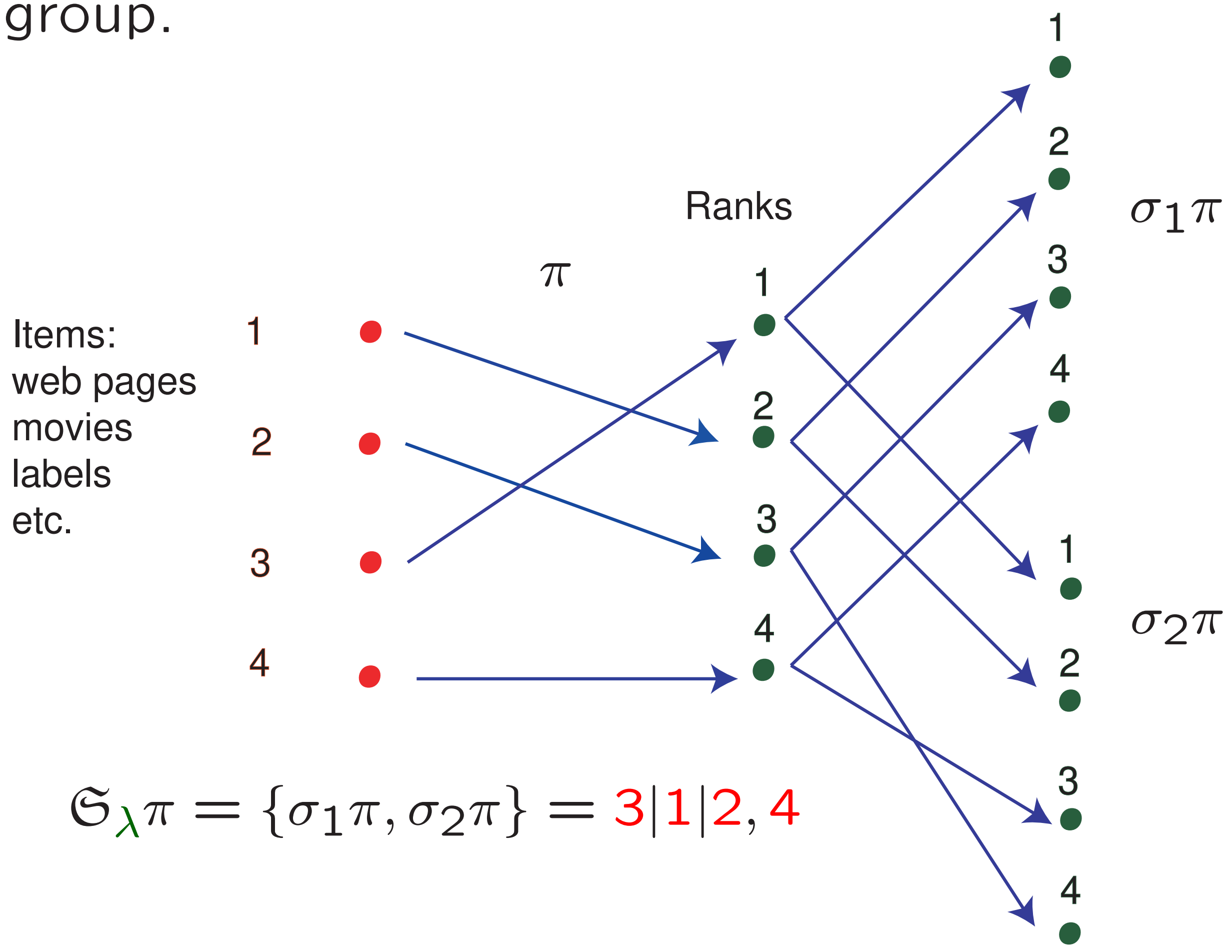
- There has been little work on this topic in machine learning. Statistics literature typically focuses on non-conditional models for permutations.

Full Ranking

A full ranking of the items $\{1, \ldots, n\}$ corresponds to an element of the symmetric group $\mathfrak{S}_n$.



A partial ranking of type λ over the items $\{1, \ldots, n\}$ corresponds to a coset in the space $\mathfrak{S}_n / \mathfrak{S}_\lambda$, where $\mathfrak{S}_\lambda$ is a subgroup of the symmetric group.



$$\mathfrak{S}_\lambda \pi = \{\sigma_1 \pi, \sigma_2 \pi\} = 3|1|2, 4$$

- Full orderings $\pi \in \mathfrak{S}_n$ correspond to permutations e.g. $3|2|4|1$.

- Cosets $\mathfrak{S}_n/\mathfrak{S}_{n-1}$ correspond to classifications e.g. $3|1,2,4$.

- Cosets $\mathfrak{S}_n/\mathfrak{S}_{n-k}$ correspond to top k ordering e.g. $1|3|2,4$.

- Cosets $\mathfrak{S}_n/\mathfrak{S}_{\lambda}$ corresponds to partial ranking e.g. $1,2,4|3$.

Conditional Models for Ranked Data

- The Task: Learn a model

$$p(\pi|\sigma_1, \ldots, \sigma_k\,; \theta) \qquad \begin{aligned} &\pi \in \mathfrak{S}_n/\mathfrak{S}_\lambda \\ &\sigma_1 \in \mathfrak{S}_n/\mathfrak{S}_{\eta_1} \\ &\vdots \\ &\sigma_k \in \mathfrak{S}_n/\mathfrak{S}_{\eta_k} \end{aligned}$$

based on data

$$D = \left\{ (\pi^{(1)}, \boldsymbol{\sigma}^{(1)}), (\pi^{(2)}, \boldsymbol{\sigma}^{(2)}), \ldots, (\pi^{(N)}, \boldsymbol{\sigma}^{(N)}) \right\}$$

Examples

- Ensemble multiclass classification

$$\pi = 2|1,3,4 \quad \sigma_i = 1|2,3,4$$

- Ensemble multilabel classification

$$\pi = 1,3|2,4 \quad \sigma_i = 1,2|3,4$$

- Meta-search $\pi = 1|3|2,4 \quad \sigma_i = 3|2|1,4$

- Collaborative filtering

$$\pi = 1|3|2|4 \quad \sigma_i = 1,2|3,4$$

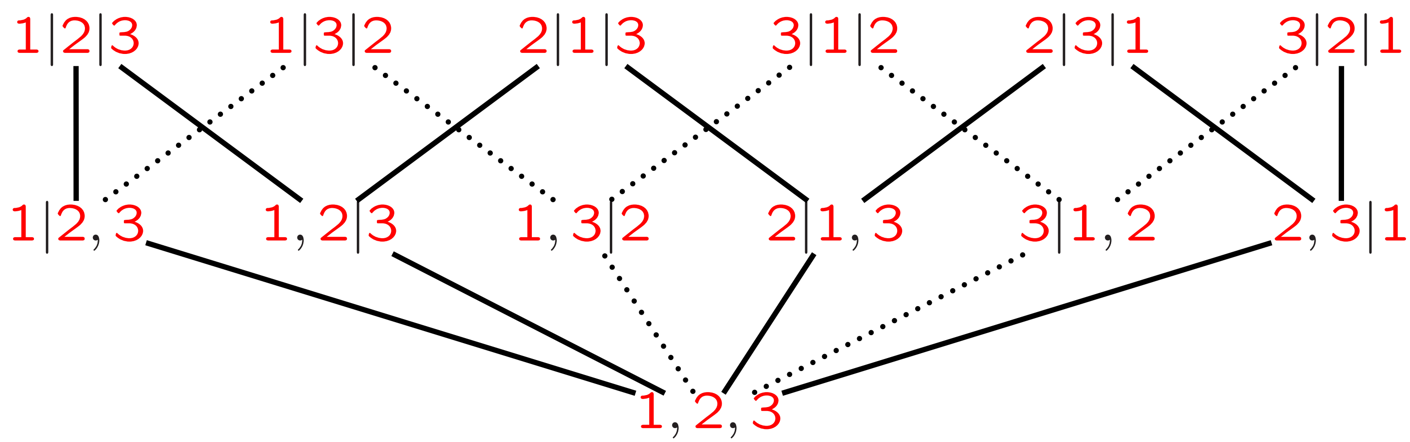
Model Design

- What should be the parametric form of $p(\pi|\sigma_1, \ldots, \sigma_k\,; \theta)$?
- How can we express a connection between the predicted variable π and the covariates $\sigma_1, \ldots, \sigma_k$?
- $\Rightarrow$ We define a combinatorial structure, the ranking poset, that facilitates measuring relations between partial rankings and exploiting the algebraic structure of rankings.

The Ranking Poset

- The set of all partial rankings $\{\mathfrak{S}_\lambda \pi\}_{\pi,\lambda}$ with a partial order of refinement $\preceq$ forms a partially ordered set $\mathfrak{W}_n$.
  E.g. $1|3|2|4 \preceq 1|3,2|4 \preceq 1|3,2,4$.

- The Hasse diagram of the ranking poset over three items



Distances on Permutations

A distance on permutations is a function $d : \mathfrak{S}_n \times \mathfrak{S}_n \to \mathbb{R}$ that satisfies

$$\begin{aligned} d(\pi,\pi) &= 0, & \forall \pi \in \mathfrak{S}_n \\ d(\pi,\sigma) &= d(\sigma,\pi) > 0 & \forall \pi \neq \sigma \in \mathfrak{S}_n \\ d(\pi,\sigma) &\leq d(\pi,\tau) + d(\tau,\sigma), & \forall \pi,\sigma,\tau \in \mathfrak{S}_n \end{aligned}$$

To obtain symmetric treatment of all items it is necessary to require right-invariance to arbitrary relabeling of $y_1, \ldots, y_n$

$$d(\pi,\sigma) = d(\pi \cdot \tau, \sigma \cdot \tau) \quad \forall \pi,\sigma,\tau \in \mathfrak{S}_n.$$

Symmetric Group's Action on the Ranking Poset

The symmetric group acts on the ranking poset

$$\{\tau_{i_1}\}_{i_1}|\cdots|\{\tau_{i_k}\}_{i_k} \cdot \pi \mapsto \{\pi(\tau_{i_1})\}_{i_1}|\cdots\{\pi(\tau_{i_k})\}_{i_k}$$

Up $\nearrow$ and down $\searrow$ moves on the Hasse diagram commute with the above action

$$\begin{tikzcd} \mathfrak{W}_n \arrow[r, "\mathfrak{S}_n"] \arrow[d, "\searrow"'] & \mathfrak{W}_n \arrow[d, "\searrow"] \\ \mathfrak{W}_n \arrow[r, "\mathfrak{S}_n"] & \mathfrak{W}_n \end{tikzcd} \qquad \begin{tikzcd} \mathfrak{W}_n \arrow[r, "\mathfrak{S}_n"] \arrow[d, "\nearrow"'] & \mathfrak{W}_n \arrow[d, "\nearrow"] \\ \mathfrak{W}_n \arrow[r, "\mathfrak{S}_n"] & \mathfrak{W}_n \end{tikzcd}$$

Distances on the Ranking Poset

A distance function $d : \Pi \times \Sigma \to \mathbb{R}_+$, $\Pi, \Sigma \subset \mathfrak{W}_n$ is right invariant if

$$d(\pi, \sigma) = d(\pi \cdot \tau, \sigma \cdot \tau) \quad \forall \tau \in \mathfrak{S}_n \quad \forall \pi \in \Pi \quad \forall \sigma \in \Sigma$$

Since $\nearrow, \searrow$ moves commutes with the action of the symmetric group, distances that are based on $\nearrow, \searrow$ moves are right invariant.

Conditional Models on the Ranking Poset

$$p(\pi|\boldsymbol{\sigma};\boldsymbol{\theta}) = \frac{1}{Z} e^{\sum_{j=1}^{k} \theta_j d(\pi,\sigma_j)} \qquad \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^k \qquad \pi \in \Pi \subset \mathfrak{W}_n \qquad \sigma_j \in \Sigma \subset \mathfrak{W}_n$$

- By selecting Θ, Π, Σ and d appropriately we show that the following models are special cases:

  Mallows model $p(\pi;\sigma,\theta) = \exp(\theta T(\pi,\sigma) - \psi(\theta)), \quad \pi,\sigma \in \mathfrak{S}_n$
  Cranking $p(\pi|\sigma,\theta) = \exp(\sum_i \theta_i T(\pi,\sigma_i) - \psi(\theta,\sigma)), \quad \pi,\sigma_i \in \mathfrak{S}_n$
  ECOC, Discrete Logistic Regression and AdaBoost.M2.

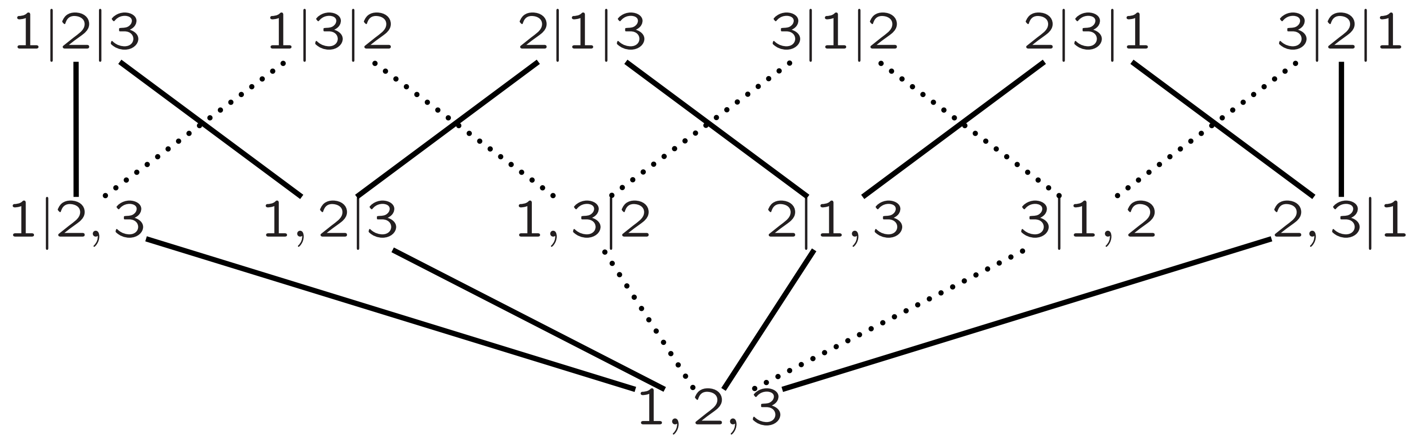
- Cranking and Mallows Model
  $\Theta = \mathbb{R}^k, \Pi = \Sigma = \mathfrak{S}_n$ and $d(\pi, \sigma)$ be the minimum number of $\searrow\nearrow$ moves on the Hasse diagram needed to bring π to σ.

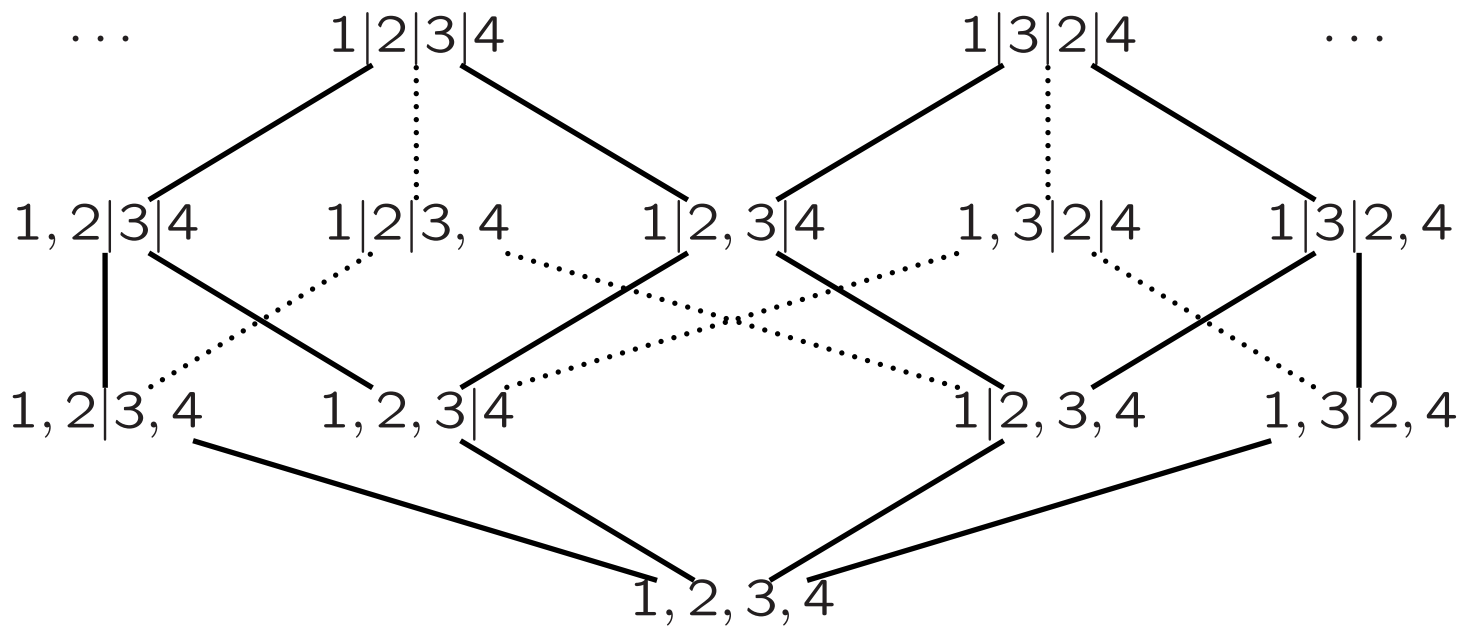
- Error Correcting Output Codes

  $$\Theta = \{\boldsymbol{\theta} \in \mathbb{R}^k | \theta_1 = \theta_2 = \cdots = \theta_k,\ \forall i\ \theta_i < 0\}, \quad \Pi = \mathfrak{S}_n / \mathfrak{S}_{n-1}$$

  $\Sigma = \mathfrak{W}_{n|1} \setminus \Pi$ and $d(\pi, \sigma)$ be the minimal number of $\nearrow\searrow$ moves needed to bring π to σ.

- Discrete Logistic Regression and AdaBoost.M2
  $\Theta = \mathbb{R}^k\ \Pi = \Sigma = \mathfrak{S}_n / \mathfrak{S}_{n-1}$ and $d(\pi, \sigma)$ be the minimum number of $\nearrow\searrow$ moves.


1|2|3
1|3|2
2|1|3
3|1|2
2|3|1
3|2|1
1|2,3
1,2|3
1,3|2
2|1,3
3|1,2
2,3|1
1,2,3


· · ·
1|2|3|4
1|3|2|4
· · ·
1,2|3|4
1|2|3,4
1|2,3|4
1,3|2|4
1|3|2,4
1,2|3,4
1,2,3|4
1|2,3,4
1,3|2,4
1,2,3,4

Bayesian Interpretation

Proposition 1 *If d is right invariant, Σ is invariant under the action of $\mathfrak{S}_n$, and $\mathfrak{S}_n$ acts transitively on Π, then the model*

$$p_\theta(\pi|\sigma_1, \ldots, \sigma_k) = \frac{1}{Z} q_0(\pi) e^{\sum_{j=1}^k \theta_j d(\pi, \sigma_j)} \quad \pi \in \Pi \subset \mathfrak{W}_n \quad \sigma_j \in \Sigma \subset \mathfrak{W}_n$$

is the posterior under independent sampling of generalized Mallows models,

$$\sigma_j \sim e^{\theta_j d(\pi, \sigma_j) - \psi(\theta_j)}$$

with prior $\pi \sim q_0$.

The conditions of this proposition are satisfied, for example, when $\Pi = \mathfrak{S}_n / \mathfrak{S}_\lambda$ and $\Sigma = \cup_i \mathfrak{S}_n / \mathfrak{S}_{\lambda_i}$

Cranking: A Conditional Model for Combining Rankings

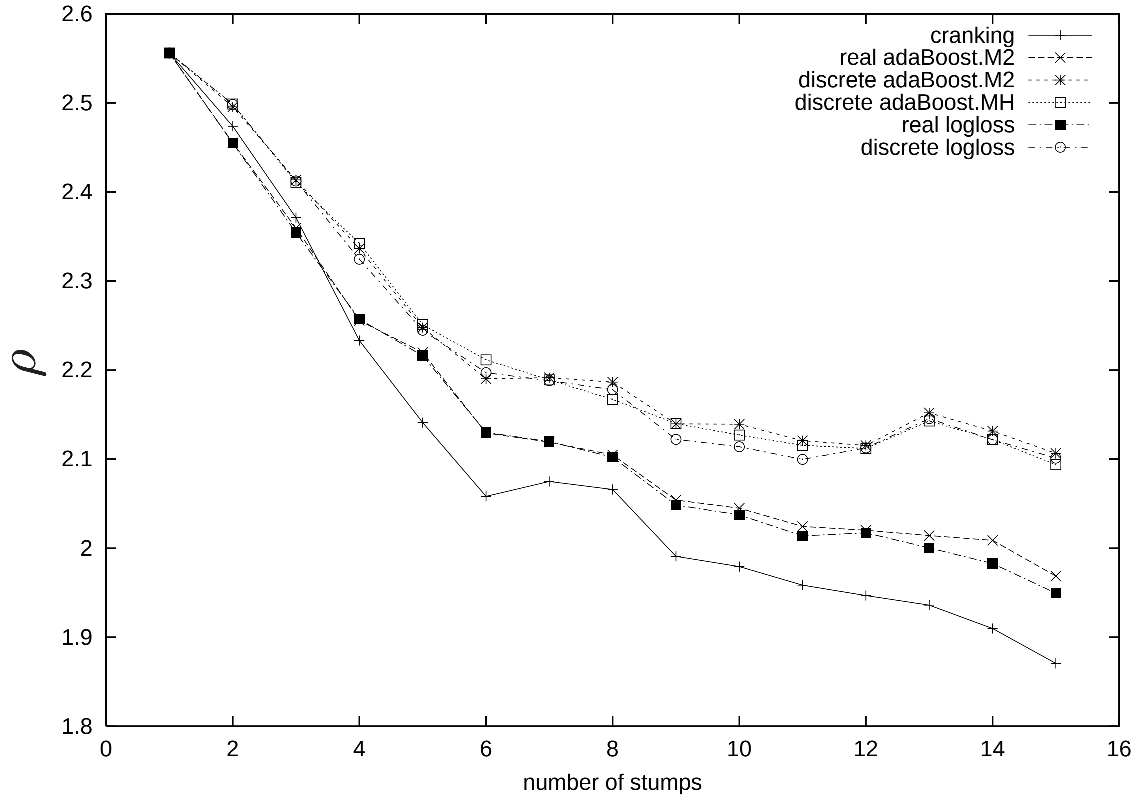
We extend Mallows model to the following *conditional* model (see also Feigin 1993)

$$p(\pi \mid \boldsymbol{\sigma}; \boldsymbol{\theta}) \;=\; \frac{1}{Z(\boldsymbol{\theta}, \boldsymbol{\sigma})}\, e^{\sum_j \theta_j d(\pi, \sigma_j)} \qquad \pi, \sigma_i \in \mathfrak{S}_n$$

$\boldsymbol{\theta}$ are estimated using *conditional* maximum likelihood over the data $D = \{(\pi^{(i)}, \boldsymbol{\sigma}^{(i)})\}_i$

$$\boldsymbol{\theta} = \arg\max_{\boldsymbol{\theta}} \; \log \prod_i p(\pi^{(i)} | \boldsymbol{\sigma}^{(i)}; \boldsymbol{\theta})$$

Classification Experiment: Mixture of 5 Gaussians



Test rank rate vs. number of (random) decision stumps

Metasearch Experiment

Data: Queries with names of ML researchers, expanded in different ways and used to retrieve Web pages (Cohen, Schapire & Singer, 1999)

- 14 input rankers
- Unique correct answer
- Conditional model is trained to maximize marginal likelihood, using a Metropolis-Hastings algorithm.

Coset size	*Bayes Fuse* (Laplace)	*Bayes Fuse* (Kernel)	*Cranking*	*Best ranker*
3	1.5	1.7	2.5	3.3
7	3.6	3.5	3.0	4.8
10	3.6	3.1	2.7	6.0

Permutation model gives better results when ranking a larger number of pages. More computationally expensive.

New Directions

- Learning and inference in permutation models require approximation and are often not efficient.

- The ranking poset allows defining more efficient models directly on cosets.

- Extensions of the ranking poset to

  1. Censored ranking $5|1, 4|3 \quad [2, 6, 7...]$

  2. Categorical ranking $|3|\,|1, 4|2$

Acknowledgments

Work described was done in collaboration with John Lafferty.

Thanks to

William Cohen, Douglas Critchlow and Joseph Verducci.