CIS 700/4: Machine Learning and Econometrics **January 19, 2017**

Lecture 2: Models and Methods for Ranked Data

Professor Shivani Agarwal *Zach Schutzman*

NB: *These notes are from CIS700 at Penn.*

2.1

For large numbers of items, fully-ranked is usually not practical.

Definition 2.1 *A **permutation** on n items is a bijective mapping $\sigma : [n] \to [n]$.*
Also considered an element of the symmetric group on n elements, S_n.
$\sigma(i) :=$ the position of item i.
$\sigma^{-1}(j) :=$ the item at position j.

Suppose we have ranked data with full rankings $\sigma_1\sigma_2 \ldots \sigma_M$.
We may have a number of questions about this data.

Statistical learning to infer a distribution over the permutations (or some other property)

Rank aggregation to find a "consensus" ranking or 'winner'

Learn a voting rule to determine winning item(s)

Example: *Borda Voting Rule: $B(i) := \sum_{m=1}^{M}(n - \sigma_m(i))$. The Borda Winner is $w \in \arg\max_i B(i)$*

Definition 2.2 *The **Kendall Tau Distance** between permutations σ and σ' is:*
$d_K(\sigma, \sigma') = \sum_{i<j} \mathbb{1}((\sigma(i) - \sigma(j))(\sigma'(i) - \sigma'(j)) < 0)$

i.e. the number of pairs that are flipped between the two rankings

Definition 2.3 Kemeny Ranking:
$\hat{\sigma} \in \arg\min_{\sigma \in S_n} \sum(d_K(\sigma, \sigma'))$

i.e. find the ranking that minimizes the distance to the data

This requires $O(n!)$ many comparisons, NP-Hard

These will show up in web-search applications

Definition 2.4 *A distribution on permutations of n items is an assignment of probability $p(\sigma)$ to each permutation $\sigma \in S_n$.*

There are uncountably many such distributions (we need $n! - 1$ parameters from $[0, 1]$

We often use families of distributions to make this easier to work with.

Mallows Model

Definition 2.5 *The Mallows distribution is a probability distribution that looks like $\frac{1}{Z(\theta)} \exp -\theta d_K(\sigma, \sigma_0)$, where $\sigma_0 \in S_n$ is the modal permutation, $\theta \geq 0$ is the dispersion parameter. Z is a normalizer.*

This looks kind of Gaussian, as the distribution is a function of the distance from the mean/mode.

How do we estimate parameters? Maximum Likelihood! Recall the likelihood function $\mathcal{L}(\beta) = \prod p(\sigma_m; \beta)$

Given data $\sigma_1 \dots \sigma_M$, the log-likelihood function is $\ell(\sigma_0, \theta) = -M \ln Z(\theta) - \theta \sum_1^M d_K(\sigma_m, \sigma_0)$

But this is equivalent to the Kemeny Rank problem, so it also is NP-Hard. There is an efficient algorithm if most of the observations are clustered around σ_0 (i.e. θ is large).

Plackett-Luce Model (PL)

Parameters $w = (w_1, \dots, w_n)^T \in \mathbb{R}^+$.

We have $p(\sigma; w) = \frac{w_{i_1}}{w_{i_1} + \dots w_{i_n}} \frac{w_{i_2}}{w_{i_2} + \dots w_{i_n}} \dots$

We can also find a log-likelihood function, but there is no closed-form expression for the maximizing $\hat{w}$. Early algorithms used Newton-Raphson root-finding. Faster algorithms do minorization-maximization. Recent work has developed faster spectral methods. Other approaches include method-of-moments.

Random Utility Models

Random variables $X_1, \dots, X_n$ with some joint distribution.

$p(\sigma) = P(X_{\sigma^{-1}(1)} > X_{\sigma^{-1}(2)} > \dots)$

Special Cases:
Gumbel random variables ($X_i \sim e^{-e^{-(x-\mu_i)}}$). Plackett-Luce is a special case of this.
Thurstone Models: $X_i \sim N(\mu_i, 1)$

We can define a consensus ranking σ^* as the permutation that maximizes the probability $p(\sigma)$ for all $\sigma \in S_n$.

We can define a voting rule as the permutation with the smallest expected rank or the one that has the highest probability of being ranked in the top position.

2.1.0.1 Other types of input data

Top k-rankings, partial rankings, pairwise comparisons, marginals (i.e. $P(\sigma(i) = c)$