Ranking Problems

9.520 Class 09, 08 March 2006

Giorgos Zacharia

Supervised Ranking Problems

- Preference Modeling:
  - Given a set of possible product configurations x_1, x_2,…x_d predict the most preferred one; predict the rating
- Information Retrieval:
  - Given a query q, and set of candidate matches x_1, x_2,…x_d predict the best answer
- Information Extraction:
  - Given a set of possible part of speech tagging choices, x_1, x_2,…x_d predict the most correct tag boundaries
    - E.g "The_day_they_shot_John_Lennon/WE at the Dogherty_Arts_Center/WE"
- Multiclass classification:
  - Given a set of possible class labels y_1, y_2,…y_d and confindense scores c_1, c_2,…c_d, predict the correct label

Types of information available

- Preference modeling:
  - Metric based:
    - User rated configuration x_i with y_i=U (x_i)
  - Choice based:
    - Given choices x_1, x_2,…x_d, the user chose x_f
  - Prior information about the features:
    - Cheaper is better
    - Faster is better
    - etc

Types of information available

- Information Retrieval:
  - Metric based:
    - Users clicked on link x_i with a frequency y_i=U (x_i)
  - Choice based:
    - Given choices x_1, x_2,…x_d, the user clicked on x_f
  - Prior information about the features:
    - Keyword matches (the more the better)
    - Unsupervised similarity scores (TFIDF)
    - etc

Types of information available

- Information Extraction:
  - Choice based:
    - Given tagging choices x_1, x_2,…x_d, the hand labeling chose x_f
  - Prior information about the features:
    - Unsupervised scores
- Multiclass:
  - Choice based:
    - Given vectors the confidence scores c_1, c_2,…c_d for class labels 1,2,…d the correct label was $y_{f..}$. The confidence scores may be coming from set of weak classifiers, and/or OVA comparisons.
  - Prior information about the features:
    - The higher the confidence score the more likely to represent the correct label.

(Semi-)Unsupervised Ranking Problems

- Learn relationships of the form:
  - Class A is closer to B, than it is to C
- We are given a set of l labeled comparisons for a user, and a set of u seemingly-unrelated comparisons from other users.
  - How do we incorporate the seemingly-unrelated information from the u instances
  - How do we measure similarity

Rank Correlation Kendall's τ

$$\tau = \frac{P-Q}{P+Q} = 1 - \frac{2Q}{\binom{n}{2}} = \frac{2P}{\binom{n}{2}} - 1$$

- P is the number of concordant pairs
- Q is the number of discordant pairs
- Value ranges from -1 for reverse rankings to +1 for same rankings.
- 0 implies independence

Example

Person	A	B	C	D	E	F	G	H
Rank by Height	1	2	3	4	5	6	7	8
Rank by Weight	3	4	1	2	5	7	8	6

- $P = 5 + 4 + 5 + 4 + 3 + 1 + 0 + 0 = 22$

$$\tau = \frac{2P}{\binom{n}{2}} - 1 = \frac{44}{22} - 1 = 0.57$$

Minimizing discordant pairs

maximize

$$Kendall's\ \tau = 1 - \frac{2Q}{\binom{n}{2}}$$

Equivalent to satisfying all constraints:

$$\forall\ r(x_i) \geq r(x_j):\ w\Phi(x_i) \geq w\Phi(x_j)$$

Familiar problem

accounting for noise:

$\forall\ r(x_i) \geq r(x_j): w\Phi(x_i) \geq w\Phi(x_j)+1-\xi_{ij}$

$\xi_{ij} \geq 0$

rearranging :

$w\left(\Phi(x_i) - \Phi(x_j)\right) \geq 1-\xi_{ij}$

equivalent to classification of pairwise difference vectors

Regularized Ranking

$$\min_{f \in H_K} \sum_{j,i=1}^{l} V\left(y_i - y_j, f\left(x_i - x_j\right)\right) + \gamma \| f \|_K^2$$

Notes:
$V(.)$ can be any relevant loss function
We could use any binary classifier; RLSC, SVM, Boosted Trees, etc
The framework for classifying vectors of differences is general enough to apply to both metric, and choice based problems

Bound on Mean Average Precision

Minimizing Q, works for other IR metrics as well.
Consider Mean Average Precision:

$$Mean\left(AvgPrec\right) = \frac{1}{n}\sum_{i=1}^{n}\frac{i}{p_i}$$

$$p_i = rank\ of\ sorted\ retrieved\ item\ i$$

$$n = number\ of\ ranked\ retrieved\ items$$

$$\sum_{i=1}^{n} p_i = Q + n\left(n+1\right)/2$$

$$Q = number\ of\ discordant\ items$$

$$\min \frac{1}{n}\sum_{i=1}^{n}\frac{i}{p_i}$$

$$subject\ to\ p_i < p_j \in \mathbb{N} \forall i < j$$

Bound on Mean Average Precision

$$use\ Lagrange\ multipliers:$$

$$\min L = \frac{1}{n}\sum_{i=1}^{n}\frac{i}{p_i} + \mu\left[\sum_{i=1}^{n} p_i - Q - n(n+1)/2\right]$$

$$\frac{\partial L}{\partial p_i} = -\frac{i}{n} p_i^{-2} + \mu = 0 \Rightarrow p_i = \sqrt{\frac{i}{n\mu}}$$

$$L = \frac{1}{n}\sum_{i=1}^{n}\frac{i}{\sqrt{\frac{i}{n\mu}}} + \mu\left[\sum_{i=1}^{n}\sqrt{\frac{i}{n\mu}} - Q - n(n+1)/2\right] = 2\sqrt{\frac{\mu}{n}}\sum_{i=1}^{n}\sqrt{i} - \mu\left[Q + n(n+1)/2\right]$$

$$\frac{\partial L}{\partial \mu} = \sqrt{\frac{1}{n\mu}}\sum_{i=1}^{n}\sqrt{i} - \left[Q + n(n+1)/2\right] = 0 \Rightarrow \mu = \frac{1}{n}\left[\sum_{i=1}^{n}\sqrt{i}\Big/\left[Q + n(n+1)/2\right]\right]^2$$

$$\Rightarrow Mean(AvgPrec) \geq \frac{1}{n}\left(\sum_{i=1}^{n}\sqrt{i}\right)^2\left[Q + n(n+1)/2\right]^{-1}$$

Prior Information

- Ranking problems come with a lot of prior knowledge
  - Positivity constraints
    - For a pairwise comparison, where all attributes are equal, except one, the instance with the highest (lowest) value is preferred.
  - If A is better than B, then B is worse than A

Prior information

Positivity constraints

Assume linear SVM case:

$$\min_{w_1,\ldots,w_m,\xi_i} \sum_{i=1}^{n} \xi_i + \lambda \sum_{f=1\ldots m} w_f^2$$

$$\forall i \in \{1,\ldots,n\}$$

$$w_f \geq 1-\xi_f, \forall f = 1, \ldots m$$

The problem becomes:

$$\min_{w_1,\ldots,w_m,\xi_i} \sum_{i=1}^{n} \xi_i + C\sum_{f=1}^{m} \xi_f + \lambda \sum_{f=1\ldots m} w_f^2$$

Symmetric comparisons

$$if$$

$$f\left(x_i - x_j\right) = +1$$

$$then$$

$$f\left(x_j - x_i\right) = -1$$

Constructing the training set from examples

- Sometimes the comparisons are not explicit:
  - Information Retrieval (Learn from clickthrough data)
    - "Winning" instances are the ones clicked most often
    - Features are other ranking scores (similarity of query with title, or text segments in emphasis etc). This also implies positivity constraints
  - Supervised summarization
    - "Winning" words are they ones that show up in the summary
    - Features are other content-word predictors (TFIDF score, distance from beginning of text, etc). We can again incorporate positivity constraints

Semi-unsupervised Ranking

- Learn distance metrics from comparisons of the form:
  - A is closer to B, than C
- Examples from WEBKB (Schultz&Joachims):
  - Webpages from the same university are closer than ones from different schools
  - Webpages about the same topic (faculty, student, project, and course) are closer than pages from different ones
  - Webpages about same topic are close. If from different topics, but one of them a student page, and one a faculty page, then they are closer than other different topic pages.

Learning weighted distances

$$d_{\Phi,W}\left(\phi(x),\phi(y)\right)=\sqrt{\left(\phi(x)-\phi(y)\right)^{T}\Phi W\Phi^{\mathrm{T}}\left(\phi(x)-\phi(y)\right)}$$

$$=\sqrt{\sum_{i=1}^{n}W_{ij}\left(K(x,x_i)-K(y,x_i)\right)^2}$$

$this\, leads\, to:$

$$\min\frac{1}{2}\left\|AWA^{T}\right\|^{2}+C\sum_{i,j,k}\xi_{ijk}$$

$$s.t.(i,j,k)\in P_{train}:\left(x_i-x_k\right)^{T}AWA^{T}\left(x_i-x_k\right)-\left(x_i-x_j\right)^{T}AWA^{T}\left(x_i-x_j\right)\geq 1-\xi_{ijk}$$

$or\, we\, can\, write\, it\, as:$

$$\min\frac{1}{2}w^{T}Lw+C\sum_{i,j,k}\xi_{ijk}$$

$$with\, A=\Phi,\; L=(A^{T}A)(A^{T}A)\quad s.t.\quad \left\|AWA^{T}\right\|^{2}=w^{T}Lw$$

Learning distance metrics

Experiments (Schultz&Joachims)

	Learned	Binary	TFIDF
University Distance	98.43%	67.88%	80.72%
Topic Distance	75.40%	61.82%	55.57%
Topic+FacultyStudent Distance	79.67%	63.08%	55.06%

Note: Schultz&Joachims report that they got the best results with a linear kernel where A=I. They do not regularize the complexity of their weighted distance metric (Remember Regularized Manifolds from previous class)

Learning from seemingly-unrelated comparisons

(Evgeniou&Pontil; Chappelle&Harchaoui)
Given l comparisons from the same user
and u comparisons from seemingly-unrelated users:

$$\min_{f \in H_K} \sum_{i=1}^{l} V\left(y_i - f\left(x_i\right)\right) + \mu^2 \sum_{i=l+1}^{l+u} V\left(y_i - f\left(x_i\right)\right) + \gamma \| f \|_K^2$$

$$0 \leq \mu \leq 1$$

$$where\ y_i = y_j - y_k\ and\ \ x_i = x_j - x_k, \forall j \neq k$$

Results of RLSC experiments with $l=10$ comparisons per user, with u instances of seemingly-unrelated comparisons, and weight μ on loss contributed by the seemingly-unrelated data.

	u=10	u=20	u=30	u=50	u=100
μ=0	18.141 %	18.090 %	18.380 %	18.040%	18.430 %
μ=0.000001	18.268 %	18.117 %	17.847 %	18.152%	18.009 %
μ=0.00001	17.897 %	18.123 %	18.217 %	18.182%	18.164 %
μ=0.0001	17.999 %	18.135 %	18.067 %	18.089 %	18.036 %
μ=0.001	18.182 %	17.835 %	18.092 %	18.140 %	18.135 %
μ=0.01	17.986 %	17.905 %	18.043 %	18.023 %	18.174 %
μ=0.1	17.132 %	16.508 %	16.225 %	15.636 %	**15.242%**
μ=0.2	16.133 %	**15.520 %**	**15.157 %**	**15.323 %**	15.276 %
μ=0.3	**15.998 %**	15.602 %	15.918 %	16.304 %	17.055 %
μ=0.4	16.581 %	16.786 %	17.162 %	17.812 %	19.494 %
μ=0.5	17.455 %	17.810 %	18.676 %	19.838 %	22.090 %
μ=0.6	18.748 %	19.589 %	20.440 %	22.355 %	25.258 %

Ranking learning with seemingly-unrelated data

- More seemingly-unrelated comparisons in the training set improve results
- There is no measure of similarity of the seemingly-unrelated data (recall Schultz&Joachims)

Regularized Manifolds

$$f^* = \operatorname{argmin}_{f \in H_K} \frac{1}{l} \sum_{i=1}^{l} V\left(x_i, y_i, f\right) + \gamma_A \left\| f \right\|_K^2 + \frac{\gamma_I}{\left(u+l\right)^2} \sum_{i,j=1}^{l} V\left(f\left(x_i\right) - f\left(x_j\right) \right)^2 W_{ij}$$

$$= \operatorname{argmin}_{f \in H_K} \frac{1}{l} \sum_{i=1}^{l} V\left(x_i, y_i, f\right) + \gamma_A \left\| f \right\|_K^2 + \frac{\gamma_I}{\left(u+l\right)^2} f^T L f$$

Laplacian $L = D - W$

Laplacian RLSC:

$$\min_{f \in H_K} \frac{1}{l} \sum_{i=1}^{l} \left(y_i - f\left(x_i\right) \right)^2 + \gamma_A \left\| f \right\|_K^2 + \frac{\gamma_I}{\left(u+l\right)^2} f^T L f$$

Laplacian RLSC for ranking with seemingly-unrelated data

$$\min_{f \in H_K} \frac{1}{l}\sum_{i=1}^{l}\left(y_i - f\left(x_i\right)\right)^2 + \frac{\mu^2}{u}\sum_{i=l+1}^{l+u}\left(y_i - f\left(x_i\right)\right)^2 + \gamma_A \left\| f \right\|_K^2 + \frac{\gamma_I}{\left(u+l\right)^2} f^T L f$$

This is equivalent to the following minimization:

$$\min_{f \in H_K} \frac{1}{l}\sum_{i=1}^{l+u}\left(y_i^{\mu} - f\left(x_i^{\mu}\right)\right)^2 + \gamma_A \left\| f \right\|_K^2 + \frac{\gamma_I}{\left(u+l\right)^2} f^T L f$$

Laplacian RLSC for ranking with seemingly-unrelated data

$$f^*(x) = \sum_{i=1}^{l+u} \alpha_i^* K^\mu(x, x_i)$$

$$y_i^\mu = y_i \,, x_i^\mu = x_i \;\text{ for } i \le l$$

$$y_i^\mu = \mu' y_i \,, x_i^\mu = \mu' x_i \;\text{ for } l < i \le l + u$$

$$\mu' = \frac{\mu l}{u}$$

K^μ is the $(l+u)\times(l+u)$ gram matrix $K_{ij}^\mu = K(x_i^\mu, x_j^\mu)$

$Y^\mu = \left[y_1 \ldots y_l \,, \mu y_{l+1} \ldots \mu y_{l+u} \right]$

Replace $f(x)$, take partial derivatives and solve for α^*

$$\alpha^* = \left(K^\mu + \gamma_A l I + \frac{\gamma_I l}{(u+l)^2} L K^\mu \right)^{-1} Y^\mu$$

Results of Laplacian RLSC experiments with l=10 comparisons per user, with u instances of seemingly-unrelated data, and μ weight on loss contributed by the seemingly-unrelated comparisons.

	u=10	u=20	u=30	u=50	u=100
μ=0	17.50%	18.50%	18.38%	18.20%	17.54%
μ=0.000001	17.34 %	19.46 %	17.52 %	18.11 %	20.10 %
μ=0.00001	18.30 %	18.20 %	17.54 %	18.46 %	18.10 %
μ=0.0001	18.56 %	18.76 %	18.02 %	17.73 %	17.90 %
μ=0.001	17.20 %	18.12 %	18.28 %	17.87 %	18.00 %
μ=0.01	16.92 %	17.52 %	17.98 %	17.70 %	18.15 %
μ=0.1	16.86 %	16.68 %	16.04 %	15.58 %	16.30 %
μ=0.2	**14.80 %**	**14.68 %**	**14.86 %**	**14.89 %**	**14.30 %**
μ=0.3	16.22 %	16.76 %	16.74 %	16.57 %	18.60 %
μ=0.4	15.94 %	16.54 %	17.94 %	17.93 %	20.75 %
μ=0.5	17.90 %	16.64 %	18.74 %	19.48 %	20.60 %
μ=0.6	17.74 %	20.20 %	20.60 %	22.38 %	25.35 %

Observations

- Optimal μ (estimated by CV) gives better performance, than without the Manifold setting
- More seemingly-unrelated data, do not affect performance significantly
- Seemingly-unrelated examples have impact that depends on the manifold transformation:
  - *The intrinsic penalty term accounts for examples that are neighboring on the manifold, and have opposite labels.*