Three Approaches to Ordinal Classification

Krzysztof Dembczyński, Wojciech Kotłowski

Institute of Computing Science
Poznań University of Technology

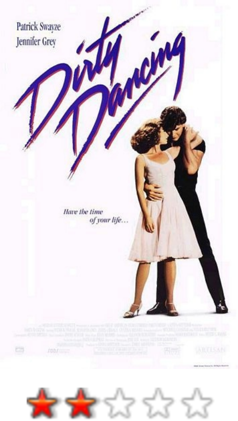
EURO 2009, Bonn, July 8, 2009

Ordinal classification consists in **predicting** a **label** taken from a **finite** and **ordered set** for an **object** described by some **attributes**.

This problem shares some characteristics of **multi-class classification** and **regression**, but:

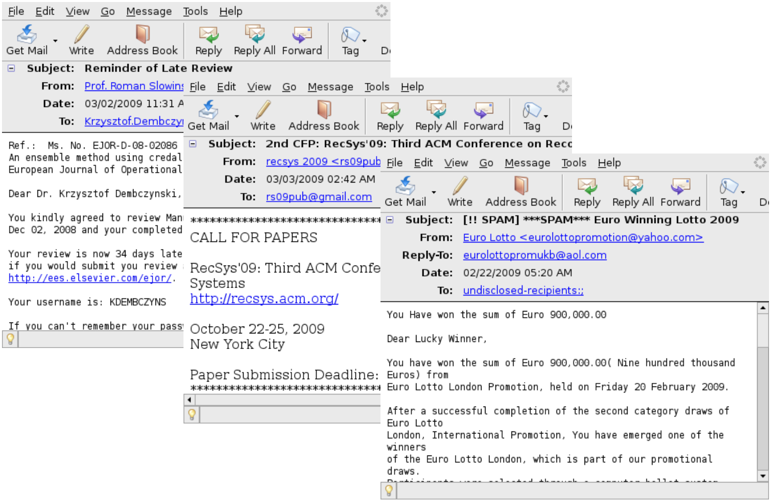
- the **order** between class labels **cannot** be **neglected**,
- the **scale** of the decision attribute is **not cardinal**.

Recommender system predicting a rating of a movie for a given user.



???

Email filtering to ordered groups like: important, normal, later, or spam.



Nature of ordinal classification:

- Classification with ordered class labels?
- Degenerate ranking problem?

Denotation:

- K – number of classes
- y – actual label
- $\hat{y}$ – predicted label
- $\mathbf{x}$ – attributes
- $f(\mathbf{x})$ – prediction (ranking or utility) function
- $L(\cdot)$ – loss function
- $[\![\cdot]\!]$ – Boolean test

Ordinal Classification – Probability Estimation:

- Prediction risk is defined by a **loss matrix**:

$$\mathbf{L}(y, \hat{y}) = (l_{y,\hat{y}})_{K \times K}$$

with **v-shaped rows** and zeros on diagonal.

$$\mathbf{L}(y, \hat{y}) = \begin{pmatrix} 0 & 1 & 2 & 3 \\ 1 & 0 & 1 & 2 \\ 2 & 1 & 0 & 1 \\ 3 & 2 & 1 & 0 \end{pmatrix}$$

Ordinal Classification – Probability Estimation:

- **Bayes decision** for the loss matrix $\mathbf{L}(y, \hat{y})$ is given by:

$$\hat{y}^* = \arg\min_{\hat{y}} \sum_{k=1}^{K} \Pr(y = k|\mathbf{x})\mathbf{L}(k, \hat{y}).$$

- To solve the problem, we need to **estimate** conditional probabilities $\Pr(y = k|\mathbf{x})$ – a lot of algorithms ...
- We can **decompose** the problem to $K - 1$ binary problems by utilizing the order of labels y: the result then are estimates of $\Pr(y > k|\mathbf{x})$, $k = 1, \ldots, K - 1$.
- To satisfy **monotonicity** of $\Pr(y > k|\mathbf{x})$, $k = 1, \ldots, K - 1$, we use **isotonic regression**.
- Other possibilities allowed ...

Ordinal Classification – Probability Estimation:

- Given $\Pr(y = k|\mathbf{x})$, $k = 1, \ldots, K$, the **optimal prediction** is:

$$\hat{y}^* = \begin{cases} \arg\max_k \Pr(y = k|\mathbf{x}), & \text{for} \;\; l_{y\hat{y}} = [\![y \neq \hat{y}]\!], \\ \\ \text{median}(y|\mathbf{x}), & \text{for} \;\; l_{y\hat{y}} = |y - \hat{y}|, \\ \\ E(y|\mathbf{x}), & \text{for} \;\; l_{y\hat{y}} = (y - \hat{y})^2. \end{cases}$$

- **Absolute-error** loss seems to be the most natural since its Bayes decision is **median** that does not depend on scale of labels.
- **Any** function of the probability distribution can be used for **object ranking**.

Ordinal Classification – Degenerate Ranking:

- Prediction risk is defined by a **rank loss** computed over pairs of objects:

$$L\left(y_{\circ\bullet}, f(\mathbf{x}_{\circ}), f(\mathbf{x}_{\bullet})\right) = [\![y_{\circ\bullet}(f(\mathbf{x}_{\circ}) - f(\mathbf{x}_{\bullet})) \leqslant 0]\!],$$

where

$$y_{\circ\bullet} = \operatorname{sgn}(y_{\circ} - y_{\bullet}),$$

and $f(\mathbf{x})$ is a **ranking** (or **utility**) function.

$$\begin{array}{ccccccccccc} y_{i_1} & > & \color{red}{y_{i_2}} & > & \color{red}{y_{i_3}} & > \ldots > & y_{i_{N-1}} & > & y_{i_N} \\ f(\mathbf{x}_{i_1}) & > & \color{red}{f(\mathbf{x}_{i_3})} & > & \color{red}{f(\mathbf{x}_{i_2})} & > \ldots > & f(\mathbf{x}_{i_{N-1}}) & > & f(\mathbf{x}_{i_N}) \end{array}$$

Ordinal Classification – Degenerate Ranking:

- This approach **ranks** the objects.
- To **assign class labels**, one has to compute **thresholds** on a range of the ranking function with respect to a given **loss matrix**.
- Rank loss minimization is strictly connected with maximization of **AUC criterion** used in binary classification.
- Minimization of rank loss on training set has **quadratic complexity** with respect to number of object, however, in the case of K ordered classes, the algorithm can work in **linear** time.

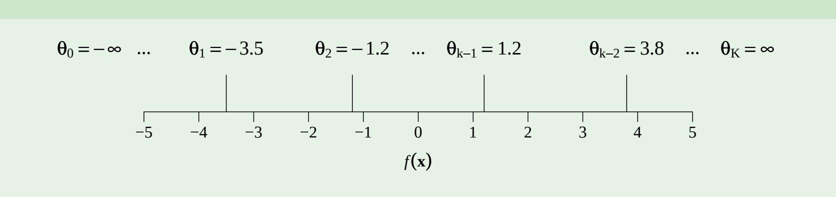
Ordinal Classification – Threshold Loss:

- Prediction risk is defined by **threshold loss**:

$$L(y, f(\mathbf{x}), \boldsymbol{\theta}) = \sum_{k=1}^{K-1} [\![y_k (f(\mathbf{x}) - \theta_k) \geqslant 0]\!],$$

where $\boldsymbol{\theta} = (\theta_0, \ldots, \theta_K)$ are **consecutive thresholds** to be computed **simultaneously** with $f(\mathbf{x})$, and

$$y_k = 1, \text{ if } y > k, \text{ and } y_k = -1, \text{ otherwise } y \leqslant k.$$



Ordinal Classification – Threshold Loss:

- This approach **shares** characteristics of the previous two.
- Comparison of an object to **thresholds** instead to **all other training objects** – lower complexity, but linear algorithms exist for rank loss minimization in ordinal classification settings.
- **Joint** solution for all $K - 1$ binary problems – **no need** of **isotonization** of conditional probabilities, but the result is a **single** value.
- Weighted threshold loss can **approximate** any loss matrix.

Boosting-like Algorithms for Three Approaches:

- Prediction function is an **ensemble of decision rules**:

$$f(\mathbf{x}) = \alpha_0 + \sum_{m=1}^{M} r_m(\mathbf{x}).$$

- We used **boosting** approach to learn $f(\mathbf{x})$: in each iteration, a single rule is generated by concentrating on examples which were hardest to classify correctly by previous rules with respect to a given **loss function**.

Boosting-like Algorithms for Three Approaches:

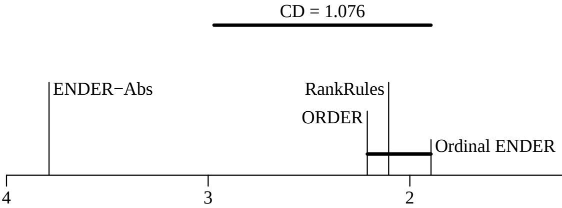
- **Ordinal ENDER** – decomposes the problem into a sequence of binary problems for estimating $\Pr(y > k|\mathbf{x})$; uses isotonic regression for isotonization of the estimates; final prediction is median over computed class distribution.
- **RankRules** – minimizes (exponential) rank loss; parameterized to minimize absolute-error.
- **ORDER** – minimizes (exponential) threshold loss; parameterized to minimize absolute-error.

- **ENDER-Abs** – reference algorithm constructing ensemble of decision rules by direct minimization of absolute-error.

All the algorithms work in **linear time** with respect to number of training example (plus log-linear time for sorting used once in preprocessing phase).

Experimental Results:

- Comparison of Ordinal ENDER, RankRules, RankRules and ENDER AE.
- 19 benchmark sets taken from Luis Torgo repository – transformed from regression to ordinal classification settings.
- Average ranks are computed with repect to mean absolute error obtained on each data set.
- Critical difference in average ranks is CD $= 1.076$.



Experimental Results:

- There is almost **no quantitative difference** in performance and time consumption: RankRules is slightly slower.
- **Qualitative differences**: Ordinal ENDER is related to probability estimation, but RankRules to AUC maximization.
- Ensemble of decision rules are **competitive** to: RankBoost AE, ORBoost-All, SVM-IMC.

Ordinal Matrix Factorization:

- Given **sparse** matrix $\mathbf{Y}$ of observed values build a model based on **matrix factorization**:

$$\mathbf{Y} \simeq \hat{\mathbf{Y}} = \mathbf{U}\mathbf{V}^T$$

where $\mathbf{U}$ is an $I \times M$ and $\mathbf{V}^T$ is a $M \times J$ matrix.

- The **prediction** is then defined by:

$$\hat{y}_{ij} = \sum_{m=1}^{M} u_{im} v_{jm}.$$

- **Example**: I is the number of users, J is the number of movies in the movie recommender system, and M is number of features describing users and movies.
- For learning we use **gradient descent** applied alternately to $\mathbf{U}$ and $\mathbf{V}$ matrices with respect to a given **loss function**.

Ordinal Matrix Factorization for Three Approaches:

- **Decomposition schema** for probability estimation.
- Minimization of **rank loss**.
- Minimization of **threshold loss**.
- **Hypothesis**: all the approaches perform **similarly**.
- For all three approaches **linear** algorithms exists: minimization of (exponential) rank loss, however, is the most demanding.
- No satisfactory results yet :(
- Work in progress ...

Conclusions:

- Nature of ordinal classification?
- Three approaches to ordinal classification.
- Boosting-like algorithm: rather qualitative than quantitative differences between these approaches.
- Ordinal Matrix factorization: in progress ...