On the Consistency of Preference Learning

Sauro Menchetti
DSI, Dipartimento di Sistemi e Informatica
Università di Firenze, Italy
menchett@dsi.unifi.it
http://www.dsi.unifi.it/~menchett/

Last update on February 3, 2006

Technical Report RT 1/2006

Abstract

In this paper, we give a novel theoretical analysis which explains why a setwise loss function exhibits a better performance of a pairwise loss function based on an utility function. We introduce a model of preference and ranking problems based on the concept of partial order relation and we provide three different approaches for carrying out this model. For understanding what is the approach with smaller generalization error, we evaluate the Bayes risk of realizing the preference and ranking model by each one of the three approaches. We will understand that the direct approach exhibits better performance than the utility function approach and than a model based on a function that works directly on pairs. Finally, we show how the ranking and preference generalization error depends on the size of set of alternatives.
Keywords: Bayes Function, Bayes Risk, Preference and Ranking Learning, Partial Order Model, Utility Function, Dependency on Alternative Set Size.

1 Introduction

Work on learning theory has mostly concentrated on classification and regression. However, there are many applications in which it is desirable to choose the best element in a set of alternatives (preference problem) or to order a collection of objects (ranking problem). There are many works in literature that try to propose a solution for these problems. Herbrich et al. (2000) investigates the problem of ordinal regression and uses a large margin algorithm based on a mapping from objects to scalar utility values for classifying pairs of objects. Herbirch et al. (1998) deals with the task of learning a preference relation from a given set of ranked documents. The problem is reformulated as a classification problem on pairs of documents, where each document is mapped to a scalar utility value. Crammer and Singer (2002b) discusses the problem of ranking instances. They describe an efficient online algorithm similar to perceptron algorithm that projects the instances into sub–intervals of the reals: each interval

is associated with a distinct rank. Also Cohen et al. (1999) considers the problem of ranking instances. It describes a two–stages approach: before a binary preference function indicating if a instance is better than another is learned, and then new instances are ordered with the purpose of maximizing the agreement with the learned preference function. In Crammer and Singer (2002a) and Elisseeff and Weston (2002) is described the problem of multi–labelled documents. Both Crammer and Singer (2002a) and Elisseeff and Weston (2002) maintains a set of prototypes associated with topics. Elisseeff and Weston (2002) reduce the multi–label problem into multiple binary problems by comparing all pairs of labels. Crammer and Singer (2002a) suggests an online algorithm similar to perceptron algorithm that updates the prototypes only if the predicted ranking is not perfect. In Joachims (2002b) is described a method to rerank the results of a search engine, adapting them to a particular group of users: it uses a SVM classifier on pairs of examples.

Menchetti et al. (2003, 2005) show an experimental analysis comparing recursive neural networks and voted perceptron for solving preference problems. Results indicate that both RNNs and the kernel VP are effective to solve the proposed problems. The experiments also indicate that the choice of a pairwise or global loss function plays an import role. In particular, it appears that the pairwise loss is not well suited to train an RNN and it remains to be investigated if this is also the case for kernel methods. Interesting, previous works with kernels Herbrich et al. (2000); Joachims (2002a); Collins and Duffy (2001) focus exclusively on pairwise loss functions. The development of global loss function for preference tasks may lead to more effective solutions. So it is interesting to theoretically investigate why a pairwise loss function behaves worse than a global loss function based on all the elements of the set of alternatives.

The remainder of the paper is organized as follows. In Section 2 we introduce some useful results on the Bayes function for regression and binary and multiclass classification problems. In Section 3 we derive a new model of preference and ranking problems based on the idea that a binary partial order relation can model the constraints of preference and ranking problems. We describe three possible approaches for the partial order model based on a 0–1 loss function. Section 4 we compare the three approaches described in Section 3, showing which is the best methods. Finally, in Section 5 we describe how the ranking and preference errors depend on the size of set of alternatives.

2 The Bayes Function

The Bayes function is the minimizer of expected risk

$$\mathrm{err}_\rho(f) \doteq \int_{\mathcal{X}\times\mathcal{Y}} V(f(\boldsymbol{x}), y)\rho(\boldsymbol{x}, y)d\boldsymbol{x}dy \tag{1}$$

and depends on the distribution ρ (Devroye et al., 1996; Cucker and Smale, 2001; Duda et al., 2001). If ρ was known, then it would be possible to compute directly the Bayes function f_ρ using its definition

$$f_\rho \doteq \arg\min_{f\in\mathcal{T}} \mathrm{err}_\rho(f)$$

In the following, the Bayes function is derived for regression and for binary and multiclass classification problems. We start from the Bayes function for

regression (Cucker and Smale, 2001).

Theorem 2.1 (Bayes Function for Regression) *In the regression problem where $f : \mathcal{X} \mapsto \mathbb{R}$, for a quadratic loss function $V(f(\boldsymbol{x}), y) = (y - f(\boldsymbol{x}))^2$, the Bayes function (also called regression function) is:*

$$f_\rho(\boldsymbol{x}) = \int_{\mathcal{Y}} y\rho(y|\boldsymbol{x})dy = \mathbf{E}_{\mathcal{Y}|\boldsymbol{x}}\{\mathcal{Y}|\boldsymbol{x}\} \tag{2}$$

So the Bayes function $f_\rho(\boldsymbol{x})$ is the expected value of random variable $\mathcal{Y}$ given $\mathcal{X} = \boldsymbol{x}$.

Proof In the case of regression, it is useful to write the expected risk (1) as follows:

$$\mathrm{err}_\rho(f) = \int_{\mathcal{X}} \int_{\mathcal{Y}} (y - f(\boldsymbol{x}))^2 \rho(y|\boldsymbol{x})dy\rho(\boldsymbol{x})d\boldsymbol{x} \tag{3}$$

Setting its first derivative with respect to $f(\boldsymbol{x})$ to zero, it follows that

$$\begin{aligned}
\frac{\partial \mathrm{err}_\rho(f)}{\partial f(\boldsymbol{x})} &= -\int_{\mathcal{X}} \int_{\mathcal{Y}} 2(y - f(\boldsymbol{x}))\rho(y|\boldsymbol{x})dy\rho(\boldsymbol{x})d\boldsymbol{x} = 0 \\
\int_{\mathcal{X}} \int_{\mathcal{Y}} y\rho(y|\boldsymbol{x})dy\rho(\boldsymbol{x})d\boldsymbol{x} &= \int_{\mathcal{X}} \int_{\mathcal{Y}} f(\boldsymbol{x})\rho(y|\boldsymbol{x})dy\rho(\boldsymbol{x})d\boldsymbol{x} \\
\int_{\mathcal{X}} \int_{\mathcal{Y}} y\rho(y|\boldsymbol{x})dy\rho(\boldsymbol{x})d\boldsymbol{x} &= \int_{\mathcal{X}} \int_{\mathcal{Y}} \rho(y|\boldsymbol{x})dyf(\boldsymbol{x})\rho(\boldsymbol{x})d\boldsymbol{x}
\end{aligned}$$

Since $\int_{\mathcal{Y}} \rho(y|\boldsymbol{x})dy = 1$, then

$$\int_{\mathcal{X}} \int_{\mathcal{Y}} y\rho(y|\boldsymbol{x})dy\rho(\boldsymbol{x})d\boldsymbol{x} = \int_{\mathcal{X}} f(\boldsymbol{x})\rho(\boldsymbol{x})d\boldsymbol{x}$$

Comparing the two members of previous equation, the regression function is derived

$$f_\rho(\boldsymbol{x}) = \int_{\mathcal{Y}} y\rho(y|\boldsymbol{x})dy$$

□

It is interesting to compute the error associated with the regression function which represents a lower bound on the error that depends only on the intrinsic difficulty of the problem.

Proposition 2.1 (Regression Bayes Function Error) *In the regression problem where $f : \mathcal{X} \mapsto \mathbb{R}$, for a quadratic loss function $V(f(\boldsymbol{x}), y) = (y - f(\boldsymbol{x}))^2$, the error of Bayes function $f_\rho(\boldsymbol{x})$ is:*

$$\mathrm{err}_\rho(f_\rho(\boldsymbol{x})) = \mathbf{E}_{\mathcal{X}}\big\{\mathbf{Var}_{\mathcal{Y}|\boldsymbol{x}}\{\mathcal{Y}|\boldsymbol{x}\}\big\} \tag{4}$$

Proof Substituting Equation (2) in Equation (3), if follows that

$$\begin{aligned}
\mathrm{err}_\rho(f_\rho(\boldsymbol{x})) &= \int_{\mathcal{X}} \int_{\mathcal{Y}} (y - \mathbf{E}_{\mathcal{Y}|\boldsymbol{x}}\{\mathcal{Y}|\boldsymbol{x}\})^2 \rho(y|\boldsymbol{x})dy\rho(\boldsymbol{x})d\boldsymbol{x} \\
&= \int_{\mathcal{X}} \mathbf{Var}_{\mathcal{Y}|\boldsymbol{x}}\{\mathcal{Y}|\boldsymbol{x}\}\rho(\boldsymbol{x})d\boldsymbol{x} \\
&= \mathbf{E}_{\mathcal{X}}\big\{\mathbf{Var}_{\mathcal{Y}|\boldsymbol{x}}\{\mathcal{Y}|\boldsymbol{x}\}\big\}
\end{aligned}$$

□

Remark If $\mathbf{Var}_{\mathcal{Y}|\boldsymbol{x}}\{\mathcal{Y}|\boldsymbol{x}\} = 0 \; \forall \, \boldsymbol{x} \in \mathcal{X}$, that is if exists only one possible y for each $\boldsymbol{x}$ (the relation between $\mathcal{X}$ and $\mathcal{Y}$ is deterministic and not probabilistic), then $\mathrm{err}_\rho(f_\rho(\boldsymbol{x})) = 0$ if we assume that the target space $\mathcal{T}$ contains the function which assigns to each $\boldsymbol{x}$ its output y.
After the regression task, the Bayes function is devised for binary and multiclass classification problems where the cost of each error is weighted by the value of the loss function (Lee et al., 2004; Tewari and Bartlett, 2005).

Theorem 2.2 (Bayes Function for Multiclass Classification) *In the multiclass classification problem where $f : \mathcal{X} \mapsto \{1, \ldots, c\}$, if we use a loss function $V(f(\boldsymbol{x}), y)$, the Bayes function is:*

$$f_\rho(\boldsymbol{x}) = \arg \min_{f(\boldsymbol{x}) \in \{1,\ldots,c\}} \mathbf{E}_{\mathcal{Y}|\boldsymbol{x}}\{V(f(\boldsymbol{x}), y)\} \tag{5}$$

So the Bayes function $f_\rho(\boldsymbol{x})$ predicts the class label $\hat{y} = f(\boldsymbol{x})$ which minimizes the expected value of $V(\hat{y}, y)$ over $\rho(y|\boldsymbol{x})$.

Proof The expected risk $\mathrm{err}_\rho(f)$ for a multiclass classification problem is:

$$\mathrm{err}_\rho(f) = \int_{\mathcal{X}} \sum_{y=1, y \neq f(\boldsymbol{x})}^{c} V(f(\boldsymbol{x}), y) \rho(y|\boldsymbol{x}) \rho(\boldsymbol{x}) d\boldsymbol{x}$$

So the task is to minimize the value of integral on the input space $\mathcal{X}$:

$$f_\rho(\boldsymbol{x}) = \arg \min_{f \in \mathcal{T}} \sum_{y=1, y \neq f(\boldsymbol{x})}^{c} V(f(\boldsymbol{x}), y) \rho(y|\boldsymbol{x}) \tag{6}$$

$$= \arg \min_{f \in \mathcal{T}} \sum_{y=1}^{c} V(f(\boldsymbol{x}), y) \rho(y|\boldsymbol{x}) \tag{7}$$

since $V(f(\boldsymbol{x}), f(\boldsymbol{x}))\rho(f(\boldsymbol{x})|\boldsymbol{x}) = 0$. If we define the expected value of the loss function $V(f(\boldsymbol{x}), y)$ over $\rho(y|\boldsymbol{x})$ as

$$\mathbf{E}_{\mathcal{Y}|\boldsymbol{x}}\{V(f(\boldsymbol{x}), y)\} = \sum_{y=1}^{c} V(f(\boldsymbol{x}), y) \rho(y|\boldsymbol{x}) \tag{8}$$

then

$$f_\rho(\boldsymbol{x}) = \arg \min_{f \in \mathcal{T}} \mathbf{E}_{\mathcal{Y}|\boldsymbol{x}}\{V(f(\boldsymbol{x}), y)\} \tag{9}$$

The Bayes function chooses the class label $f(\boldsymbol{x})$ which minimizes the expected value of $V(f(\boldsymbol{x}), y)$ over $\rho(y|\boldsymbol{x})$.

□

Remark For a binary classification problem where $f : \mathcal{X} \mapsto \{+1, -1\}$, Equation (5) reduces to

$$f_\rho(\boldsymbol{x}) = \arg \min_{f(\boldsymbol{x}) \in \{+1,-1\}} \{V(f(\boldsymbol{x}), +1)\rho(+1, \boldsymbol{x}) + V(f(\boldsymbol{x}), -1)\rho(-1, \boldsymbol{x})\}$$

or, equivalently:

$$f_\rho(\boldsymbol{x}) = \begin{cases} +1 & V(-1,+1)\rho(+1,\boldsymbol{x}) \geq V(+1,-1)\rho(-1,\boldsymbol{x}) \\ -1 & \text{otherwise} \end{cases} \tag{10}$$

If the classification loss functions treats all the errors in the same way as in the case of the 0–1 loss function, the previous results can be simplified.

Proposition 2.2 (Bayes Function for Multiclass Classification) *In the multiclass classification problem where $f : \mathcal{X} \mapsto \{1, \ldots, c\}$, if we use a loss function $V(f(\boldsymbol{x}), y) = \mathcal{I}(y \neq f(\boldsymbol{x}))$ where $\mathcal{I}$ is the indicator function, the Bayes function is:*

$$f_\rho(\boldsymbol{x}) = \arg \max_{y=1,\ldots,c} \rho(y|\boldsymbol{x}) \tag{11}$$

So the Bayes function $f_\rho(\boldsymbol{x})$ assigns to each $\boldsymbol{x}$ its maximal probability output.

Proof First of all, we compute the expected risk $\mathrm{err}_\rho(f)$ for multiclass classification problem:

$$\begin{aligned} \mathrm{err}_\rho(f) &= \int_{\mathcal{X}} \int_{\mathcal{Y}} \mathcal{I}(y \neq f(\boldsymbol{x}))\rho(y|\boldsymbol{x})dy\rho(\boldsymbol{x})d\boldsymbol{x} \\ &= \int_{\mathcal{X}} \sum_{y=1, y\neq f(\boldsymbol{x})}^{c} \rho(y|\boldsymbol{x})\rho(\boldsymbol{x})d\boldsymbol{x} \\ &= \int_{\mathcal{X}} \left(1 - \rho(f(\boldsymbol{x})|\boldsymbol{x})\right) \rho(\boldsymbol{x})d\boldsymbol{x} \end{aligned} \tag{12}$$

where we used the normalization property that

$$\sum_{y=1}^{c} \rho(y|\boldsymbol{x}) = 1$$

So we are looking for a function that minimizes the previous error (12)

$$f_\rho(\boldsymbol{x}) = \arg \min_{f \in \mathcal{T}} \left(1 - \rho(f(\boldsymbol{x})|\boldsymbol{x})\right) = \arg \max_{y=1,\ldots,c} \rho(y|\boldsymbol{x})$$

□

Remark For a binary classification problem where $\mathcal{Y} = \{+1, -1\}$, Equation (11) reduces to

$$f_\rho(\boldsymbol{x}) = \begin{cases} +1 & \rho(+1|\boldsymbol{x}) \geq \rho(-1|\boldsymbol{x}) \\ -1 & \text{otherwise} \end{cases} \tag{13}$$

Now we compute the error for the multiclass classification Bayes function in the case of a 0–1 loss function.

Proposition 2.3 (Multiclass Classification Bayes Function Error) *In the multiclass classification problem where $f : \mathcal{X} \mapsto \{1, \ldots, c\}$, if we use a loss function $V(f(\boldsymbol{x}), y) = \mathcal{I}(y \neq f(\boldsymbol{x}))$, the error of Bayes function $f_\rho(\boldsymbol{x})$ is:*

$$\begin{aligned} \mathrm{err}_\rho(f_\rho(\boldsymbol{x})) &= \mathbf{E}_{\mathcal{X}} \left\{ 1 - \max_{y=1,\ldots,c} \rho(y|\boldsymbol{x}) \right\} \\ &= \int_{\mathcal{X}} \left(1 - \max_{y=1,\ldots,c} \rho(y|\boldsymbol{x}) \right) \rho(\boldsymbol{x})d\boldsymbol{x} \end{aligned} \tag{14}$$

where $(1 - \max_{y=1,\ldots,c} \rho(y|\boldsymbol{x}))$ *is the probability that* $\boldsymbol{x}$ *is not classified in the most probable class.*

Proof Substituting Equation (11) in Equation (12) yields Equation (14).

□

3 A New Model of Preference and Ranking

In this section, we introduce a new model for preference and ranking problems. We have to model a framework in which we are given a set of i.i.d. pairs $\mathcal{D}_m = \{(\boldsymbol{X}_i, R_i)\}_{i=1}^m$ where $\boldsymbol{X}_i = \{\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{ik_i}\} \subseteq \mathcal{X}$, $\boldsymbol{x}_{ij} \in \mathcal{X}$ and R_i is a relation between the elements of each subset. For example, R_i can be the ranking of $\{\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{ik_i}\}$ or a preference relation which chooses the best element of $\{\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{ik_i}\}$. Before introducing the model for preference and ranking problems, we recall the definitions of binary relation, partial order relation and total order relation.

Definition 3.1 *A binary relation* R *is a subset of Cartesian product of two sets* $\mathcal{A}$ *and* $\mathcal{B}$*:* $R \subseteq \mathcal{A} \times \mathcal{B}$*,* aRb *is an ordered pair* (a, b)*.*

Definition 3.2 *A partial order* $\preceq$ *on a set* $\mathcal{A}$ *is a binary relation* $\preceq \subseteq \mathcal{A} \times \mathcal{A}$ *that satisfies the following three properties:*

1. *Reflexivity:* $a \preceq a \ \forall a \in \mathcal{A}$
2. *Antisymmetry: if* $a \preceq b$ *and* $b \preceq a$*, then* $a = b \ \forall a, b \in \mathcal{A}$
3. *Transitivity: if* $a \preceq b$ *and* $b \preceq c$*, then* $a \preceq \forall a, b, c \in \mathcal{A}$

Definition 3.3 *A total order* $\leq$ *on a set* $\mathcal{A}$ *is a partial order that satisfies the following property:*

4. *Comparability:* $\forall a, b \in \mathcal{A}$*, either* $a \leq b$ *or* $b \leq a$

3.1 The Partial Order Model

The partial order model of preference and ranking is based on the idea that a binary partial order relation can model the constraints of a preference and ranking problem. Let $\mathcal{R}_\mathcal{X}$ be the set of all the partial order relations on $\mathcal{X}$

$$\mathcal{R}_\mathcal{X} = \{R : R \subseteq \mathcal{X} \times \mathcal{X}, R \text{ is a partial order on } \mathcal{X}\} \subseteq 2^{\mathcal{X} \times \mathcal{X}} \tag{15}$$

We can model $\mathcal{D}_m = \{(\boldsymbol{X}_i = \{\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{ik_i}\}, R_i)\}_{i=1}^m$ where $\boldsymbol{x}_{ij} \in \mathcal{X}$ as a set of i.i.d. pairs $(\boldsymbol{X}_i, R_i)$ drawn from a fixed but unknown distribution ρ on $2^\mathcal{X} \times \mathcal{R}_\mathcal{X}$, where $2^\mathcal{X}$ is the set of all the subsets of $\mathcal{X}$ and $\mathcal{R}_\mathcal{X}$ is the set of all the partial order relations on $\mathcal{X}$. The goal is to learn a function $f \in \mathcal{H}$

$$f : 2^\mathcal{X} \mapsto \mathcal{R}_\mathcal{X} \tag{16}$$

such that

$$f(\boldsymbol{X}) \approx R \tag{17}$$

which models the probabilistic relation between $2^{\mathcal{X}}$ and $\mathcal{R}_{\mathcal{X}}$. If we decompose ρ as

$$\rho(2^{\mathcal{X}}, \mathcal{R}_{\mathcal{X}}) = \rho(\mathcal{R}_{\mathcal{X}}|2^{\mathcal{X}})\rho(2^{\mathcal{X}}) \tag{18}$$

each pair of the dataset $\mathcal{D}_m$ can be obtained by a two steps process:

1. first we get a subset $\boldsymbol{X}$ of $\mathcal{X}$ in according to $\rho(2^{\mathcal{X}})$;
2. then we get a partial order R on $\boldsymbol{X}$ from $\rho(\mathcal{R}_{\mathcal{X}}|2^{\mathcal{X}})$.

This two steps process expressed by Equation (18) is able to model the noise that can corrupt the function from the input to the target space: for example, different users or the same user can sort in a different way the same set $\boldsymbol{X}$ of instances. Note that for a given $\boldsymbol{X}$, the relation on $\boldsymbol{X}$ is consistent in the sense that the transitivity between the elements of $\boldsymbol{X}$ holds. Given this model, it is interesting

- defining a loss function to measure how good is a function on a given collection of data;
- finding the Bayes function $f_\rho : 2^{\mathcal{X}} \mapsto \mathcal{R}_{\mathcal{X}}$ assuming that ρ is known;
- computing its expected risk $\mathrm{err}_\rho(f_\rho)$.

3.2 The 0–1 Loss Function

The simplest loss is 0–1 loss function defined as

$$V(f(\boldsymbol{X}), R) = \mathcal{I}(f(\boldsymbol{X}) \neq R) = \begin{cases} 1 & \text{if } R \neq f(\boldsymbol{X}) \\ 0 & \text{if } R = f(\boldsymbol{X}) \end{cases} \tag{19}$$

It counts an error when $f(\boldsymbol{X}) \neq R$, without evaluating if $f(\boldsymbol{X})$ and R are similar or very different. It behaves as the misclassification loss for classification defined as

$$V(f(\boldsymbol{x}), y) = \theta(-yf(\boldsymbol{x})) \tag{20}$$

where $\theta(x)$ is heaviside function which equals 1 if $x \geq 0$, else 0.

3.3 Three Approaches for the Partial Order Model

Given the framework described in Section 3.1, we can compute the Bayes function and its expected risk of this model. Then we can learn f using several models and then compute the expected risk of these different models. At this point, the expected risk of Bayes function can be compared with the error of other models to find the model with smaller Bayes error. We investigate three models for the function $f : 2^{\mathcal{X}} \mapsto \mathcal{R}_{\mathcal{X}}$.

1. We can directly model the probabilistic relation between $2^{\mathcal{X}}$ and $\mathcal{R}_{\mathcal{X}}$ using a function

   $$D : 2^{\mathcal{X}} \mapsto \mathcal{R}_{\mathcal{X}} \tag{21}$$

   This is the more expressive method of modelling f. For all practical purposes, this approach is not simple to realize because the target space $\mathcal{R}_{\mathcal{X}}$ is a complex output space: so we are looking for a function with a simpler target space. We will call this approach the direct model.

2. We can map each object into a real number which measures its importance by a function

$$U : \mathcal{X} \mapsto \mathbb{R} \tag{22}$$

then we can sort the alternatives by this score using

$$\pi_{ij} = \sum_{r=1}^{k_i} \theta(U(\boldsymbol{x}_{ir}) - U(\boldsymbol{x}_{ij})) \tag{23}$$

and

$$\pi_i = \arg \max_{j=1,\ldots,k_i} U(\boldsymbol{x}_{ij}) \tag{24}$$

for ranking and preference respectively (Herbirch et al., 1998; Herbrich et al., 2000; Crammer and Singer, 2002b). This is a very simple approach which assumes that exists a function that maps each object into a real number whereby we can sort the objects, hypothesis which is not always valid. It is employed in the utility function approach described in Menchetti et al. (2003, 2005). We will call this approach the utility function model.

3. Finally we can use a function which works on pairs of objects assigning a score or a label (Cohen et al., 1999)

$$P : \mathcal{X} \times \mathcal{X} \mapsto \{+1, -1\} \quad \text{or} \quad P : \mathcal{X} \times \mathcal{X} \mapsto [0, 1] \tag{25}$$

In this way, we can sort pairs of objects based on their scores or labels but we have to guarantee the transitivity property and to resolve possible inconsistencies. For example, a score greater than 0.5 or a label +1 means that the first object has to be ranked first than the other one. We will call this approach the pairwise model.

We use f_D, f_U and f_P for indicating the ranking and preference function f modelled by D, U and P respectively.

4 A Comparison of the Three Approaches

After defining in Section 3.3 three different models for the problems of ranking and preference, we show a new approach for comparing these three models. We first compute the Bayes function of preference and ranking problem under our framework with its corresponding risk (this corresponds to modelling f using D) and then compare this value with the expected risk of the other two models. We start computing the Bayes function of the problem defined in Section 3.1 which corresponds to directly model the probabilistic relation between $2^{\mathcal{X}}$ and $\mathcal{R}_{\mathcal{X}}$.

4.1 The Direct Model

In the direct model, we model f by a function D that has its same behaviour, so we can compute the output as

$$f(\boldsymbol{X}) = f_D(\boldsymbol{X}) = D(\boldsymbol{X}) \tag{26}$$

The following theorem compute the Bayes function of preference and ranking problem under the model described in Section 3.1: it is based on the assumption that if $|\boldsymbol{X}|$ is finite, then also the cardinality $|\mathcal{R}_{\mathcal{X}}|$ of set of all the partial order relations on $\mathcal{X}$ is finite. This observation permits to deriving the Bayes function using the results on multiclass classification.

Theorem 4.1 (Preference and Ranking Bayes Function) *In a preference and ranking problem in which $f : 2^{\mathcal{X}} \mapsto \mathcal{R}_{\mathcal{X}}$, if we use a 0–1 loss function $V(f(\boldsymbol{X}), R) = \mathcal{I}(f(\boldsymbol{X}) \neq R)$, the Bayes function $f_\rho(\boldsymbol{X})$ is:*

$$f_\rho(\boldsymbol{X}) = \arg \max_{R \in \mathcal{R}_{\mathcal{X}}} \rho(R|\boldsymbol{X}) \tag{27}$$

Proof If we assume that $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\} \subseteq \mathcal{X}$ is a finite set, then also $|\mathcal{R}_{\mathcal{X}}|$, the set of all the partial order relations on $\mathcal{X}$, is finite. The expected risk of f becomes:

$$\begin{aligned} \mathrm{err}_\rho(f) &= \int_{2^{\mathcal{X}}} \int_{\mathcal{R}_{\mathcal{X}}} V(f(\boldsymbol{X}), R)\rho(R|\boldsymbol{X})dR\rho(\boldsymbol{X})d\boldsymbol{X} \\ &= \int_{2^{\mathcal{X}}} \sum_{R \in \mathcal{R}_{\mathcal{X}}} \mathcal{I}(f(\boldsymbol{X}) \neq R)\rho(R|\boldsymbol{X})\rho(\boldsymbol{X})d\boldsymbol{X} \\ &= \int_{2^{\mathcal{X}}} [1 - \rho(f(\boldsymbol{X})|\boldsymbol{X})]\rho(\boldsymbol{X})d\boldsymbol{X} \end{aligned}$$

So we have cast the preference and ranking problems to a multiclass classification problem where the categories are the elements of $\mathcal{R}_{\mathcal{X}}$. The direct application of Equation (11) leads to Equation (27) which proves the theorem.

□

As in the case of multiclass classification, the Bayes function (27) assigns to each set of alternatives its maximal probability relation. Computing the expected risk of the Bayes function is a direct consequence of Proposition 2.3.

Theorem 4.2 (Bayes Risk for Preference and Ranking) *In a preference and ranking problems in which $f : 2^{\mathcal{X}} \mapsto \mathcal{R}_{\mathcal{X}}$, if we use a 0–1 loss function $V(f(\boldsymbol{X}), R) = \mathcal{I}(f(\boldsymbol{X}) \neq R)$, the error of Bayes function $f_\rho(\boldsymbol{X})$ is:*

$$\begin{aligned} \mathrm{err}_\rho(f_\rho(\boldsymbol{X})) &= \mathbf{E}_{2^{\mathcal{X}}}\left\{1 - \max_{R \in \mathcal{R}_{\mathcal{X}}} \rho(R|\boldsymbol{X})\right\} \\ &= \int_{2^{\mathcal{X}}} [1 - \max_{R \in \mathcal{R}_{\mathcal{X}}} \rho(R|\boldsymbol{X})]\rho(\boldsymbol{X})d\boldsymbol{X} \end{aligned} \tag{28}$$

Proof Applying Proposition 2.3 for multiclass classification to Equation (27) leads to Equation (28).

□

If we model f by $D : 2^{\mathcal{X}} \mapsto \mathcal{R}_{\mathcal{X}}$, the expected risk of the Bayes function f_ρ corresponds to expected risk of f_D

$$\begin{aligned} \mathrm{err}_\rho(f_D(\boldsymbol{X})) &= \mathbf{E}_{2^{\mathcal{X}}}\left\{1 - \max_{R \in \mathcal{R}_{\mathcal{X}}} \rho(R|\boldsymbol{X})\right\} \\ &= \int_{2^{\mathcal{X}}} [1 - \max_{R \in \mathcal{R}_{\mathcal{X}}} \rho(R|\boldsymbol{X})]\rho(\boldsymbol{X})d\boldsymbol{X} \end{aligned} \tag{29}$$

The next step involves the computation of the expected risk when we model the ranking and preference function by an utility function U and by a function P that works on pairs of objects.

4.2 The Utility Function Model

Now we model the ranking and preference function f by an utility function $U : \mathcal{X} \mapsto \mathbb{R}$ which assigns to each object a score proportional to its importance. The prediction $f(\boldsymbol{X}) = f_U(\boldsymbol{X})$ can be reconstruct using the utility function U in the following way:

$$f_U(\boldsymbol{X}) = \{(\boldsymbol{x}, \boldsymbol{z}) \in \mathcal{X} \times \mathcal{X}, \boldsymbol{x}, \boldsymbol{z} \in \boldsymbol{X} : U(\boldsymbol{x}) \geq U(\boldsymbol{z})\} \tag{30}$$

Note that if $U(\boldsymbol{x}) \neq U(\boldsymbol{z}) \, \forall \, \boldsymbol{x}, \boldsymbol{z} \in \mathcal{X} \times \mathcal{X}$, then U induces a total order on $\mathcal{X}$ and the cardinality of $f_U(\boldsymbol{X})$ is equal to the number of simple combinations

$$C_{n,k} = \frac{D_{n,k}}{P_k} = \frac{n!}{k!(n-k)!}$$

where $k = 2$ is the size of the subsets (in our case we pick pairs) and n is the cardinality of $\mathcal{X}$

$$|f_U(\boldsymbol{X})| = C_{|\mathcal{X}|,2} = \frac{|\mathcal{X}|(|\mathcal{X}| - 1)}{2} \tag{31}$$

But if $\exists \, \boldsymbol{x}, \boldsymbol{z} \in \mathcal{X} \times \mathcal{X} : U(\boldsymbol{x}) = U(\boldsymbol{z})$, we can model ties by two elements of the relation as $\boldsymbol{x} \preceq \boldsymbol{z}$ and $\boldsymbol{z} \preceq \boldsymbol{x} \Rightarrow \boldsymbol{x} = \boldsymbol{z}$. Then the maximum value of $|f_U(\boldsymbol{X})|$ is the number simple arrangements

$$D_{n,k} = \frac{P_n}{P_{n-k}} = \frac{n!}{(n-k)!}$$

where $k = 2$ is the size of the subsets and n is the cardinality of $\mathcal{X}$. So $|f_U(\boldsymbol{X})|$ ranges from $C_{|\mathcal{X}|,2}$ to $D_{|\mathcal{X}|,2}$ depending on the number of ties:

$$\frac{|\mathcal{X}|(|\mathcal{X}| - 1)}{2} \leq |f_U(\boldsymbol{X})| \leq |\mathcal{X}|(|\mathcal{X}| - 1) \tag{32}$$

The upper bound represent the situation in which all the alternatives get the same score. The following theorem compares the expected risk of the Bayes function f_ρ modelled by the direct model $D : 2^{\mathcal{X}} \mapsto \mathcal{R}_{\mathcal{X}}$ and by the utility function model $U : \mathcal{X} \mapsto \mathbb{R}$.

Theorem 4.3 (Direct Model vs Utility Function Model) *The expec–ted risk of the ranking and preference function $f : 2^{\mathcal{X}} \mapsto \mathcal{R}_{\mathcal{X}}$ modelled by a direct approach $D : 2^{\mathcal{X}} \mapsto \mathcal{R}_{\mathcal{X}}$ is less than or equal to the expected risk of modelling f by an utility function $U : \mathcal{X} \mapsto \mathbb{R}$ such that*

$$f_U(\boldsymbol{X}) = \{(\boldsymbol{x}, \boldsymbol{z}) \in \mathcal{X} \times \mathcal{X}, \boldsymbol{x}, \boldsymbol{z} \in \boldsymbol{X} : U(\boldsymbol{x}) \geq U(\boldsymbol{z})\} \tag{33}$$

In mathematical terms

$$\mathrm{err}_\rho(f_D(\boldsymbol{X})) \leq \mathrm{err}_\rho(f_U(\boldsymbol{X})) \tag{34}$$

Proof We start computing the expected risk of the utility function model, then we compare this value to the expected risk of direct model.

$$\begin{aligned}\mathrm{err}_\rho(f_U(\boldsymbol{X})) &= \int_{2^{\mathcal{X}}}\int_{\mathcal{R}_{\mathcal{X}}} V(f_U(\boldsymbol{X}),R)\rho(R|\boldsymbol{X})dR\rho(\boldsymbol{X})d\boldsymbol{X} \\ &= \int_{2^{\mathcal{X}}}\sum_{R\in\mathcal{R}_{\mathcal{X}}} \mathcal{I}(f_U(\boldsymbol{X})\neq R)\rho(R|\boldsymbol{X})\rho(\boldsymbol{X})d\boldsymbol{X} \\ &= \int_{2^{\mathcal{X}}}[1-\rho(f_U(\boldsymbol{X})|\boldsymbol{X})]\rho(\boldsymbol{X})d\boldsymbol{X} \qquad (35)\end{aligned}$$

Comparing Equations (28) and (35), we obtain Equation (34). If U is expressive enough such that

$$f_U(\boldsymbol{X}) = \max_{R\in\mathcal{R}_{\mathcal{X}}} \rho(R|\boldsymbol{X})$$

then

$$\mathrm{err}_\rho(f_D(\boldsymbol{X})) = \mathrm{err}_\rho(f_U(\boldsymbol{X}))$$

□

The Theorem 4.3 shows that modelling the ranking and preference function f by an utility function $U : \mathcal{X} \mapsto \mathbb{R}$ leads to a Bayes risk greater than or equal to the direct model. Only in the case that the utility function U leads to a f_U which behaves as the Bayes function, the two errors are the same. As a consequence, the utility function model by itself could induce a greater generalization error.

4.3 The Pairwise Model

The pairwise model is a more expressive model than the utility function one. Precisely, scoring the pairs and not single objects can lead to a more rich relation on the set of alternatives and the utility function approach can be obtained as a particular case of the pairwise model. It can be proved that exists some relations modelled by the pairwise approach which are not represented in the utility function one. The ranking and preference prediction function $f(\boldsymbol{X}) = f_P(\boldsymbol{X})$ can be reconstruct using the pairwise function P in the following way

$$f_P(\boldsymbol{X}) = \{(\boldsymbol{x},\boldsymbol{z}) \in \mathcal{X}\times\mathcal{X}, \boldsymbol{x},\boldsymbol{z}\in\boldsymbol{X} : P(\boldsymbol{x},\boldsymbol{z}) \geq 0.5\} \qquad (36)$$

if $P : \mathcal{X}\times\mathcal{X} \mapsto [0,1]$ is a probability score on pairs and as

$$f_P(\boldsymbol{X}) = \{(\boldsymbol{x},\boldsymbol{z}) \in \mathcal{X}\times\mathcal{X}, \boldsymbol{x},\boldsymbol{z}\in\boldsymbol{X} : P(\boldsymbol{x},\boldsymbol{z}) = +1\} \qquad (37)$$

if $P : \mathcal{X}\times\mathcal{X} \mapsto \{+1,-1\}$ is a binary classification function on pairs. The following theorem compares the expected risk of the Bayes function f_ρ modelled by the direct model $D : 2^{\mathcal{X}} \mapsto \mathcal{R}_{\mathcal{X}}$ and by the pairwise function model $P : \mathcal{X}\times\mathcal{X} \mapsto [0,1]$ or $P : \mathcal{X}\times\mathcal{X} \mapsto \{+1,-1\}$.

Theorem 4.4 (Direct Model vs Pairwise Model) *The expected risk of the ranking and preference function $f : 2^{\mathcal{X}} \mapsto \mathcal{R}_{\mathcal{X}}$ modelled by a direct approach $D : 2^{\mathcal{X}} \mapsto \mathcal{R}_{\mathcal{X}}$ is less than or equal to the expected risk of modelling f by a pairwise function $P : \mathcal{X}\times\mathcal{X} \mapsto [0,1]$ or $P : \mathcal{X}\times\mathcal{X} \mapsto \{+1,-1\}$ as described in Equations (36) and (37)*

$$\mathrm{err}_\rho(f_D(\boldsymbol{X})) \leq \mathrm{err}_\rho(f_P(\boldsymbol{X})) \qquad (38)$$

Proof The proof is the same of the utility function approach: we compute the expected risk of the pairwise model, then we compare this value to the expected risk of direct model.

$$\begin{aligned}
\mathrm{err}_\rho(f_P(\boldsymbol{X})) &= \int_{2^{\mathcal{X}}} \int_{\mathcal{R}_{\mathcal{X}}} V(f_P(\boldsymbol{X}), R)\rho(R|\boldsymbol{X})dR\rho(\boldsymbol{X})d\boldsymbol{X} \\
&= \int_{2^{\mathcal{X}}} \sum_{R\in\mathcal{R}_{\mathcal{X}}} \mathcal{I}(f_P(\boldsymbol{X}) \neq R)\rho(R|\boldsymbol{X})\rho(\boldsymbol{X})d\boldsymbol{X} \\
&= \int_{2^{\mathcal{X}}} [1-\rho(f_P(\boldsymbol{X})|\boldsymbol{X})]\rho(\boldsymbol{X})d\boldsymbol{X} \qquad (39)
\end{aligned}$$

Comparing Equations (28) and (39), we obtain Equation (38). Note that if P is expressive enough such that

$$f_P(\boldsymbol{X}) = \max_{R\in\mathcal{R}_{\mathcal{X}}} \rho(R|\boldsymbol{X})$$

then

$$\mathrm{err}_\rho(f_D(\boldsymbol{X})) = \mathrm{err}_\rho(f_P(\boldsymbol{X}))$$

□

As in the case of the utility function model, the Theorem 4.4 shows that modelling the ranking and preference function f by a pairwise function P leads to a Bayes risk greater than or equal to the direct model. Only in the case that the pairwise function P leads to a f_P which behaves as the Bayes function, the two errors are the same. As a consequence, pairwise function model by itself could induce a greater generalization error.

Finally, to conclude, we can show the relation between the expected risk of the direct, the utility function and the pairwise function models:

$$\mathrm{err}_\rho(f_D(\boldsymbol{X})) \leq \mathrm{err}_\rho(f_P(\boldsymbol{X})) \leq \mathrm{err}_\rho(f_U(\boldsymbol{X})) \qquad (40)$$

Modelling the ranking and preference function f by indirect approaches as the utility or pairwise function can lead to a greater generalization error than the direct one due to the inherent characteristics of the model which is unable to represent all the possible relations on the set of alternatives: the more expressive is the model, the smaller will be the prediction error.

5 Dependence on Size of Set of Alternatives

In this section, we describe a novel approach on how the ranking and preference errors depend on the size of set of alternatives. The larger is the size of the set of alternatives, the bigger is the probability of a ranking or preference error. But if the scores of the objects are well "separated", the probability of error can become arbitrarily small.

In the utility function approach, we learn a function $U : \mathcal{X} \mapsto \mathbb{R}$ that measures the importance of an object using a training set $\mathcal{D}_m$. Then to rank a set of alternatives, we sort the elements by their score; in the case of a preference problem, we select only the best element. Since U depends on $\mathcal{D}_m$ and since $\mathcal{D}_m$ is a set of i.i.d. pairs sampled from a probability distribution ρ on $\mathcal{X} \times \mathbb{N}$, it follows that U and furthermore also $U(\boldsymbol{x})$ are random variables depending on $\mathcal{D}_m$.

5.1 Ranking Two Alternatives

Let $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2\}$ be a set of alternatives that contains only two elements: in this case, ranking the two elements or choosing the best element are equivalent. We assume that $\boldsymbol{x}_2$ is ranked first than $\boldsymbol{x}_1$, that is $y_1 = 2$ and $y_2 = 1$. Let U_1 and U_2 be the random variables whose realizations u_1 and u_2 represent the scores associated to $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ by the utility function U and let p_{U_1}, P_{U_1}, p_{U_2} and P_{U_2} be the probability distribution functions and the cumulative distribution functions of U_1 and U_2 respectively.

Since $\boldsymbol{x}_2 \prec \boldsymbol{x}_1$, we expect that $U(\boldsymbol{x}_1) = u_1 < u_2 = U(\boldsymbol{x}_2)$. If $u_1 \geq u_2$, then we have a ranking error:

$$\mathbf{Pr}\{\text{Error}\} = \mathbf{Pr}\{u_2 \leq u_1\} = 1 - \mathbf{Pr}\{u_2 > u_1\} \tag{41}$$

Using the definition of cumulative distribution function

$$\mathbf{Pr}\{u_2 \leq c\} = P_{U_2}(c) \quad \forall\, c \in [0,1] \tag{42}$$

we obtain

$$\mathbf{Pr}\{u_2 \leq u_1\} = \int_{U_1} P_{U_2}(u_1) p_{U_1}(u_1) du_1 = \int_{U_1} P_{U_2}(u_1) P'_{U_1}(u_1) du_1 \tag{43}$$

where for definition $p_{U_1}(u_1) = P'_{U_1}(u_1) = \dfrac{dP_{U_1}(u_1)}{du_1}$, that is the probability distribution is the derivative of the cumulative distribution. If U_1 and U_2 have the same distribution probability but different expected values $\mathbf{E}_{U_1}\{u_1\}$ and $\mathbf{E}_{U_2}\{u_2\}$, we obtain

$$P_{U_2}(u) = P_{U_1}(u - \Delta) \tag{44}$$

where $\Delta = \mathbf{E}_{U_2}\{u_2\} - \mathbf{E}_{U_1}\{u_1\}$. The Equation (43) becomes

$$\mathbf{Pr}\{u_2 \leq u_1\} = \int_{U_1} P_{U_1}(u_1 - \Delta) P'_{U_1}(u_1) du_1 \tag{45}$$

In the case that $\Delta = 0 \Rightarrow \mathbf{E}_{U_1}\{u\} = \mathbf{E}_{U_2}\{u\}$, i.e. the two probability distributions of U on $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are the same, it follows that

$$\mathbf{Pr}\{u_2 \leq u_1\} = \int_{U_1} P_{U_1}(u_1) P'_{U_1}(u_1) du_1 = \left.\frac{P^2_{U_1}(u_1)}{2}\right|_{-\infty}^{+\infty} = \frac{1}{2} \tag{46}$$

where, for definition, $P_{U_1}(+\infty) = 1$ and $P_{U_1}(-\infty) = 0$. This means that if on average the utility function U maps both $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ into the same value, then we have the highest probability to make an error.

If we suppose that the cumulative distributions $P_{U_1}(u)$ and $P_{U_2}(u)$ are bell–shaped distributions, for example:

$$P_{U_1}(u) = \frac{1}{1 + e^{-u}} \Rightarrow p_{U_1}(u) = P_{U_1}(u)(1 - P_{U_1}(u)) \tag{47}$$

then the Equation (45) becomes

$$\begin{aligned} \mathbf{Pr}\{u_2 \leq u_1\} &= \int_{U_1} P_{U_1}(u_1 - \Delta) P'_{U_1}(u_1) du_1 \\ &= \int_{-\infty}^{+\infty} \frac{1}{1 + e^{-(u_1 - \Delta)}} \frac{1}{1 + e^{-u_1}} \frac{e^{-u_1}}{1 + e^{-u_1}} du_1 \\ &= \int_{-\infty}^{+\infty} \frac{e^{-u_1}}{(1 + e^{-u_1} e^{\Delta})(1 + e^{-u_1})^2} du_1 \end{aligned} \tag{48}$$

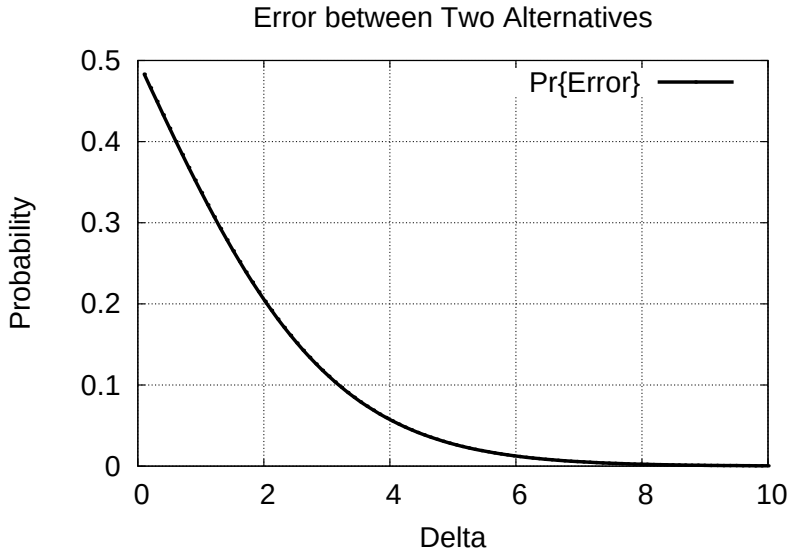

Figure 1: Ranking and preference error in function of the difference Δ between the two expected values of U on $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$.

After solving the integral , we obtain

$$\mathbf{Pr}\{\text{Error}\} = \frac{e^{\Delta}(\Delta - 1) + 1}{(e^{\Delta} - 1)^2} \approx \frac{\Delta}{e^{\Delta}} \tag{49}$$

Also in this case, if $\Delta = 0$, then $\mathbf{Pr}\{\text{Error}\} = 1/2$. The plot of $\mathbf{Pr}\{\text{Error}\}$ is shown in Figure 1. We see that the probability of error tends quickly to zero as Δ increments.

5.2 Ranking k Alternatives

Now we generalize above results to the case in which $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k\}$ is a set of k alternatives. We can define two different types of errors: the ranking error, that is probability of incorrectly ranking the set of alternatives and the preference error, that is the probability of not ranking first the best element of the set of alternatives. The ranking error is the probability of the joined event $u_1 < u_2 < \cdots < u_k$, where we supposed that $y_i = k - i + 1$, $i = 1, \ldots, k$

$$\mathbf{Pr}\{\text{RankingError}\} \quad = \quad 1 - \mathbf{Pr}\{u_1 < u_2 < \cdots < u_k\} \tag{50}$$

If we assume that the single events are independent, then we can express the ranking in function of pairs of elements

$$\begin{aligned} \mathbf{Pr}\{\text{RankingError}\} &= 1 - \mathbf{Pr}\{u_1 < u_2\}\mathbf{Pr}\{u_2 < u_3\} \cdots \mathbf{Pr}\{u_{k-1} < u_k\} \\ &= 1 - \prod_{i=1}^{k-1} \mathbf{Pr}\{u_i < u_{i+1}\} \\ &= 1 - \prod_{i=1}^{k-1} (1 - \mathbf{Pr}\{u_{i+1} \leq u_i\}) \end{aligned} \tag{51}$$

In a preference problem, the error is the probability of the joined event $u_1 < u_k, u_2 < u_k, \ldots, u_{k-1} < u_k$, where we suppose that $\boldsymbol{x}_k$ is the best element

$$\mathbf{Pr}\{\text{PreferenceError}\} = 1 - \mathbf{Pr}\{u_1 < u_k, u_2 < u_k, \ldots, u_{k-1} < u_k\} \quad (52)$$

Also for the preference task, if we suppose that the single events are independent, then we can express the preference in function of pairs of elements involving the best element

$$\begin{aligned}\mathbf{Pr}\{\text{PreferenceError}\} &= 1 - \mathbf{Pr}\{u_1 < u_k\}\mathbf{Pr}\{u_2 < u_k\}\cdots\mathbf{Pr}\{u_{k-1} < u_k\} \\ &= 1 - \prod_{i=1}^{k-1} \mathbf{Pr}\{u_i < u_k\} \\ &= 1 - \prod_{i=1}^{k-1} (1 - \mathbf{Pr}\{u_k \leq u_i\}) \end{aligned} \quad (53)$$

So the ranking and preference errors have been reduced to a production on error on pairs of elements and so we can use Equation (49) that express the probability of error of a pair of objects. Since

$$0 \leq \mathbf{Pr}\{u_i \leq u_j\} \leq \frac{1}{2} \Rightarrow \frac{1}{2} \leq 1 - \mathbf{Pr}\{u_i \leq u_j\} \leq 1$$

we can derive a lower and an upper bound for the probability of error for ranking and preference:

$$1 - \prod_{i=1}^{k-1} 1 \leq \mathbf{Pr}\{\text{Error}\} \leq 1 - \prod_{i=1}^{k-1} \frac{1}{2} \Rightarrow 0 \leq \mathbf{Pr}\{\text{Error}\} \leq 1 - \frac{1}{2^{k-1}} \quad (54)$$

where $\mathbf{Pr}\{\text{Error}\}$ is either $\mathbf{Pr}\{\text{RankingError}\}$ or $\mathbf{Pr}\{\text{PreferenceError}\}$. The curve of the upper bound of $\mathbf{Pr}\{\text{Error}\}$ is plotted in Figure 2, where we can see that the probability of error grows exponentially fast towards 1 in function of the cardinality of the set of alternatives. If we define

$$\Delta_{i,j} = \mathbf{E}_{U_i}\{u_i\} - \mathbf{E}_{U_j}\{u_j\}, \quad i, j = 1, \ldots, k : i > j \quad (55)$$

then we can express the probabilities of ranking and preference error in function of distance between the values of the utility function U on pairs

$$\begin{aligned}\mathbf{Pr}\{\text{RankingError}\} &= 1 - \prod_{i=1}^{k-1} (1 - \mathbf{Pr}\{u_{i+1} \leq u_i\}) \\ &= 1 - \prod_{i=1}^{k-1} \left(\frac{e^{\Delta_{i+1,i}}(e^{\Delta_{i+1,i}} - \Delta_{i+1,i} - 1)}{(e^{\Delta_{i+1,i}} - 1)^2} \right) \end{aligned}$$

for the ranking error and

$$\begin{aligned}\mathbf{Pr}\{\text{PreferenceError}\} &= 1 - \prod_{i=1}^{k-1} (1 - \mathbf{Pr}\{u_k \leq u_i\}) \\ &= 1 - \prod_{i=1}^{k-1} \left(\frac{e^{\Delta_{k,i}}(e^{\Delta_{k,i}} - \Delta_{k,i} - 1)}{(e^{\Delta_{k,i}} - 1)^2} \right) \end{aligned}$$

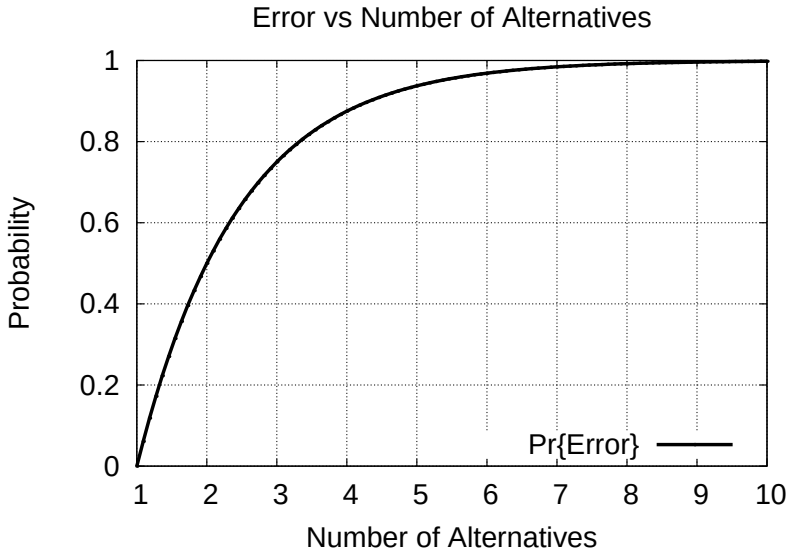

Figure 2: Upper bound of preference and ranking error in function of the number of alternatives.

for the preference error. Note that the ranking error depends on $\Delta_{i+1,i}$, $i = 1, \ldots, k-1$ while the preference error depends on $\Delta_{k,i}$, $i = 1, \ldots, k-1$, where $\boldsymbol{x}_k$ is the best element in both problems. If the utility function is able to map more similar elements into closer values, we see that the ranking problem is inherently more difficult then the preference one since $\Delta_{i+1,i} < \Delta_{k,i}$, $i = 1, \ldots, k-1$. But if the scores of the objects computed by U are well "separated", the probability of error can become arbitrarily small despite the size of the set of alternatives. Finally, note that similar results can be obtained using any other probability distribution for the scores assigned by the utility function to the elements in the set of alternatives.

6 Conclusions

We derived three approaches for a new partial order model of preference and ranking based on a 0–1 loss function exploiting the idea that a binary partial order relation can model the constraints of preference and ranking problems. We showed that modelling the ranking and preference function by indirect approaches as the utility or pairwise function could lead to a greater generalization error than the direct one due to the inherent characteristics of the model which is unable to represent all the possible relations on the set of alternatives.

Finally, we described a novel approach about how the ranking and preference errors depend on the size of set of alternatives. The larger is the size of the set of alternatives, the bigger is the probability of an error. But if the scores of the objects computed by the utility function were well separated, the probability of error could become arbitrarily small.

References

Cohen, W. W., Schapire, R. E., and Singer, Y. (1999). Learning to Order Things. *Journal of Artificial Intelligence Research*, 10:243–270.

Collins, M. and Duffy, N. (2001). Convolution Kernels for Natural Language. In Dietterich, T., Becker, S., and Ghahramani, Z., editors, *Advances in Neural Information Processing Systems 14*, pages 625–632, Cambridge, MA, USA. NIPS 14, MIT Press.

Crammer, K. and Singer, Y. (2002a). A New Family of Online Algorithms for Category Ranking. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 151–158, New York, NY, USA. SIGIR 2002, ACM Press.

Crammer, K. and Singer, Y. (2002b). Pranking with Ranking. In Dietterich, T. G., Becker, S., and Ghahramani, Z., editors, *Advances in Neural Information Processing Systems 14*, pages 641–647, Cambridge, MA, USA. NIPS 14, MIT Press.

Cucker, F. and Smale, S. (2001). On the Mathematical Foundations of Learning. *Bulletin of Americam Mathematical Society*, 39(1):1–49.

Devroye, L., Györfi, L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer–Verlag.

Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern Classification*. John Wiley and Sons, New York.

Elisseeff, A. and Weston, J. (2002). A Kernel Method for Multi–Labelled Classification. In Dietterich, T. G., Becker, S., and Ghahramani, Z., editors, *Advances in Neural Information Processing Systems 14*, pages 681–687, Cambridge, MA, USA. NIPS 14, MIT Press.

Herbirch, R., Graepel, T., Bollmann-Sdorra, P., and Obermayer, K. (1998). Learning Preference Relations for Information Retrieval. In *Proceedings Workshop Text Categorization and Machine Learning, International Conference on Machine Learning*, pages 80–84, Madison Wisconsin. ICML/AAAI–98 Workshop on Learning for Text Categorization, The AAAI Press.

Herbrich, R., Graepel, T., and Obermayer, K. (2000). Large Margin Rank Boundaries for Ordinal Regression. In Smola, A., Bartlett, P., Schölkopf, B., and Schuurmans, D., editors, *Advances in Large Margin Classifiers*. MIT Press.

Joachims, T. (2002a). Evaluating Retrieval Performance using Clickthrough Data. In *Proceedings of the SIGIR Workshop on Mathematical/Formal Methods in Information Retrieval*.

Joachims, T. (2002b). Optimizing Search Engines Using Clickthrough Data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2002)*, pages 133–142, New York, NY, USA. ACM, ACM Press.

Lee, Y., Lin, Y., and Wahba, G. (2004). Multicategory Support Vector Machines, Theory, and Application to the Classification of Microarray Data and Satellite Radiance Data. *Journal of the American Statistical Association*, 99:67–81.

Menchetti, S., Costa, F., Frasconi, P., and Pontil, M. (2003). Comparing Convolution Kernels and Recursive Neural Networks for Learning Preferences on Structured Data. In *Proceedings of IAPR – TC3 International Workshop on Artificial Neural Networks in Pattern Recognition (ANNPR 2003)*.

Menchetti, S., Costa, F., Frasconi, P., and Pontil, M. (2005). Wide Coverage Natural Language Processing using Kernel Methods and Neural Networks for Structured Data. *Pattern Recognition Letters, Special Issue on Artificial Neural Networks in Pattern Recognition*, 26(12):1896–1906. PATREC3670.

Tewari, A. and Bartlett, P. L. (2005). On the Consistency of Multiclass Classification Methods. In Auer, P. and Meir, R., editors, *Proceedings of* 18^{th} *Annual Conference on Learning Theory (COLT 2005), Bertinoro, Italy, June 27–30, 2005*, volume 3559 of *Lecture Notes in Computer Science*, pages 143–157. Springer.