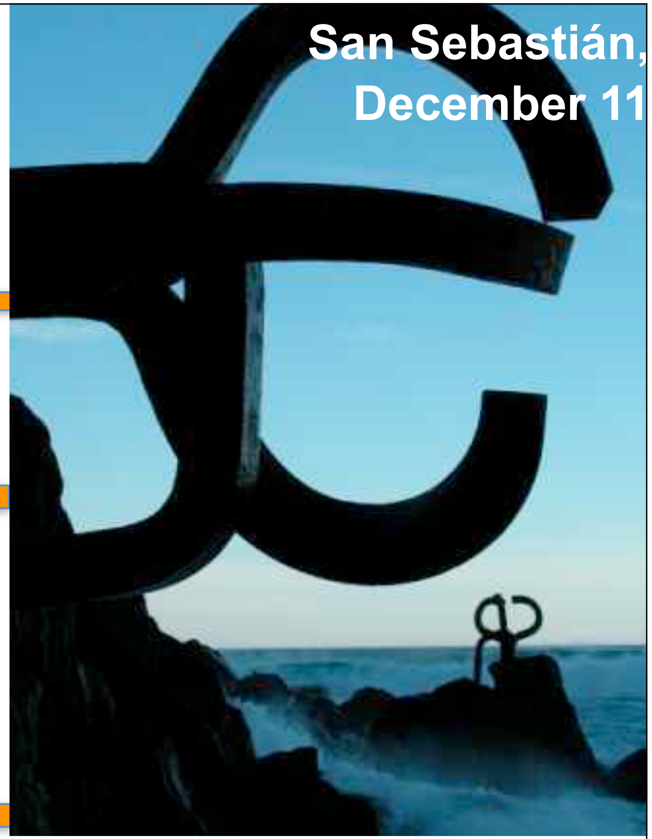


Learning to Predict One or More Ranks or Classes

Antonio Bahamonde

Artificial Intelligence Center. University of Oviedo at Gijón, Asturias, Spain

www.aic.uniovi.es

Spanish Association for Artificial Intelligence (AEPIA)

Two papers

Alonso, J., del Coz, J.J., Díez, J., Luaces, O., Bahamonde, A.: Learning to predict **one or more** ranks in ordinal regression tasks. ECML'08. pp. 39–54. No. LNAI 5211, Springer (2008)

- del Coz, J.J., Díez,J., Bahamonde,A.: Learning **nondeterministic** classifiers. Journal of Machine Learning Research 10, 2273–2293 (Oct 2009)

Outline

- Introduction to nondeterministic classifications
- Ordinal Classification (Regression)
  - The motivating application: profile of bovines
  - Formal framework
  - How to learn intervals of ranks
  - Experimental results
  - Conclusions
- Extension to multi-class Classification

Nondeterministic Classifications

- classification tasks in which the number of classes is higher than two
- most classification errors frequently occur between small **subsets of classes** that are somehow similar
- increase in reliability if classifiers were allowed to express their **doubts** whenever they were asked to classify some entries.

Nondeterministic Classifications

- Given that we learn classifiers with multiple outcomes, like nondeterministic automata; we shall call them **nondeterministic** classifiers.

- Since they return a set of values, these classifiers could be called **set-valued** classifiers.

Nondeterministic Classifiers

Example. Consider a screening for a set of medical diseases. For some inputs, a nondeterministic classifier would be able to predict more than one disease:

- will be provided to domain **experts**
- may **discard** some options and allow domain experts to make practical decisions.
- even when the nondeterministic classifier returns most of the available classes for an entry, we can read that the classifier is acknowledging its **ignorance**.

Nondeterministic Classifications

Definition 1. A nondeterministic hypothesis

$$h : \mathcal{X} \longrightarrow Intervals(\mathcal{Y})$$

- We interpret the output h(x) as an imprecise answer to a query about the class of x
- Nondeterministic classifiers can be seen as a kind of **Information Retrieval** task for each x

Ordinal Classification (Regression)

- The aim is to find hypotheses able to predict classes (ranks) that belong to a finite ordered set.

- Applications include
  - Information Retrieval,
  - Natural Language Processing,
  - Collaborative Filtering,
  - Finances
  - User Preferences

Introduction. Nondeterministic predictions

- New kind of predictions
- Hypotheses that try to predict the true rank, but when the classification is uncertain, they predict an interval of ranks, a set of consecutive ranks
  - a set as small as possible,
  - while still containing the true rank
- As we shall learn hypotheses with multiple outcomes, like nondeterministic automata, we shall call them nondeterministic ordinal regressors

Introduction. Similar approaches

- Confidence machines
  - Given an error rate $\in$, they make conformal predictions: a set of labels containing the true class with probability greater than 1-$\in$
- Hierarchical organization of biological objects
- Classification with reject option

The motivating application: profile of bovines

- The assessment of muscle proportion in carcasses of beef cattle

- Important in cattle breeding since determines:
  - the prices to be obtained by carcasses
  - the genetic value of animals to **select** studs for the next generation

Beef cattle assessment for selection purposes

Evaluate the merits of beef cattle as meat producers

Zoometric measurements
now

?

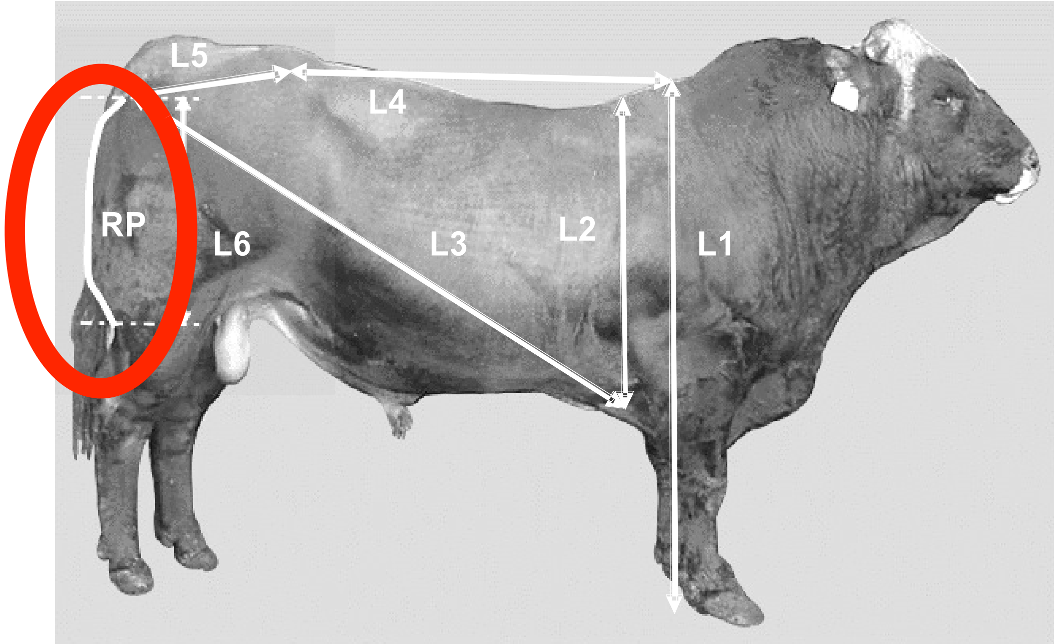
Beef cattle assessment

- ASEAVA:
  Asociación de Criadores de la Raza Asturiana de los Valles

- More than **60,000** animals

Beef cattle assessment

Zoometric measurements include the RP (round profile)



The motivating application: profile of bovines

The motivating application: profile of bovines

- It is assessed by visual appreciations of experts. But visual leads:
  - subjectivity ,
  - not repeatable, and
  - expensive

- Thus, a new learning task arises: to estimate this rank from repeatable live animal descriptions

The benefits of nondeterministic predictors

- Reliability of predictions is higher
  - Beef profiles: 76% (77%) raise to 85% (84%) with 1.28 (1.21) ranks

- When the hypothesis predicts only one rank, the estimation of the rank is very probably the true one.

The benefits of nondeterministic predictors

- When the prediction is an interval of more than one rank
- Appeal to a more expensive procedure to decide one class:
  - we may turn to an actual expert, or
  - we can wait until the natural growth of the animal make the classification more clear.

The benefits of nondeterministic predictors

- When the prediction is an interval of more than one rank, it still may be useful

- to discard an animal as a stud for the next generation
  - a prediction of [1, 2] must imply a poor genetic value as meat producer

- to compute the price of the carcass
  - averaging the values in the interval

Formal framework. Definition

- Definition 1. A nondeterministic hypothesis

$$h : \mathcal{X} \longrightarrow Intervals(\mathcal{Y})$$

- We interpret the output h(x) as an imprecise answer to a query about the rank of x
- Nondeterministic ordinal regression can be seen as a kind of **Information Retrieval** task for each x

Precision, Recall, F-measures

$$
\begin{array}{r|cc}
 & y=+1 & y=-1 \\
\hline
+1 = h(\vec{x}) & a & b \\
-1 = h(\vec{x}) & c & d
\end{array}
$$

$$R(h(\vec{x}), y) = \frac{a}{a+c}$$

$$P(h(\vec{x}), y) = \frac{a}{a+b}$$

$$LossF_\beta(a,b,c) = 1 - \frac{(1+\beta^2)PR}{\beta^2 P + R} = 1 - \frac{(1+\beta^2)a}{(1+\beta^2)a + b + \beta^2 c}$$

Formal framework. Loss functions in nd

- Recall: proportion of relevant documents found by a search

$$R(h(\mathbf{x}), y) = 1_{y \in h(\mathbf{x})}$$

- Precision: proportion of retrieved documents that are relevant

$$P(h(\mathbf{x}), y) = \frac{1_{y \in h(\mathbf{x})}}{|h(\mathbf{x})|}$$

Formal framework. Loss functions

- The harmonic average of these two amounts: F-measures

$$F_\beta(h(\mathbf{x}), y) = \frac{1+\beta^2}{\beta^2 + |h(\mathbf{x}'_j)|} 1_{y'_j \in h(\mathbf{x}'_j)}$$

Formal framework. Loss functions

- For a nondeterministic predictor h, and a test set S'

$$
\begin{aligned}
R^{\Delta}(h, S') &= \frac{1}{n}\sum_{j=1}^{n} \Delta(h(\vec{x'_j}), y'_j) = \frac{1}{n}\sum_{j=1}^{n}\Big(1 - F_\beta(h(\vec{x'_j}), y'_j)\Big) \\
&= \frac{1}{n}\sum_{j=1}^{n}\left(1 - \frac{1+\beta^2}{\beta^2 + |h(\vec{x'_j})|} 1_{y'_j \in h(\vec{x'_j})}\right)
\end{aligned}
$$

Formal framework. Loss functions. Remarks

- For a deterministic hypothesis F1, F2 and Recall are the proportion of successful classifications: the accuracy

- The nondeterministic Recall is a generalization of the deterministic accuracy. On a test set, it is the proportion of times that

$$y' \in h(\mathbf{x}')$$

How to learn intervals of ranks

- Let us assume known the posterior probabilities of ranks

$$Pr(rank = j|\mathbf{x}), \quad \forall j \in \{1, \dots, k\}$$

- Algorithm

$$h(\mathbf{x}) = argmin\{\Delta_{\mathbf{x}}(Z) : Z \in Intervals\{1, \dots, k\}\}$$

$$\Delta_{\mathbf{x}}(Z) = \sum_{y \in \mathcal{Y}} \Delta(Z, y) Pr(y|x) = \sum_{y \in \mathcal{Y}} (1 - F_\beta(Z, y))\, Pr(y|x)$$

How to learn intervals of ranks

- Proposition (Correctness)
- If posterior probabilities are known, the previous Algorithm returns the nondeterministic prediction that minimizes the risk given by the loss 1-Fβ

How to learn intervals of ranks. The role of β

- In practice, posterior probabilities are not known: they are learned probabilities as discriminant values instead of thorough descriptions of the distribution of classes
- β is a parameter that fixes the thresholds to decide the number of ranks to predict
- It should be tuned in order to achieve optimal results. For instance, to reach the highest F1 scores, it might be necessary to use a value of β different from 1.

Experimental results

- To evaluate the nondeterministic learners proposed, we compared:

- The F1 scores of well known deterministic learners and their nondeterministic counterparts

- Recall and size of predictions attained by nondeterministic learners

Experimental results. Deterministic learners

- The estimation of posterior probabilities of ranks
- Multiclass SVM (classes are not ordered)
  - libsvm (Wu, Lin, Weng, jmlr, 2004)
- Gaussian processes devised for ordinal regression tasks
  - MAP, (Chu, Ghahramani, jmlr, 2005)
- Nondeterministic counterparts
- nd_SVM
- nd_MAP

Experimental results. Datasets

- beef cattle profiles

- 12 benchmarks of metric regression (Luis Torgo's repository)
  - 5 and 10 bins with the same frequency of training examples
  - Therefore, we have 24 benchmark learning tasks

Experimental results. Comparisons

- To compare the performance of different approaches, we randomly split each data set into training/test partitions

- The scores compared are the averages over 20 independent trials

Experimental results. Dataset sizes

Dataset	#Attributes	#Train	#Test
pyrimidines	27	50	24
triazines	60	100	86
Wisconsin bc	32	150	44
machine cpu	6	150	59
auto mpg	7	200	192
stock	9	300	650
Boston	13	300	206
abalone	8	300	3877
bank	32	300	7892
computer	21	300	7892
California	8	300	20340
census	16	300	22484
Profiles 500	8	500	391
Profiles 300	8	300	591

Experimental results. In benchmarks

- nd_MAP >> nd_SVM
- In F1 and Recall

- Nondeterministic versions >> deterministic counterparts
- In F1
- Recall (differences about 0.25 – 0.30)

- Sizes: < 2 in 5 bins, around 3 in 10 bins

- Significant differences

Experimental results. In beef profiles

- In F1

Dataset	nd_MAP	(si.)	nd_SVM	MAP	(si.)	SVM	(si.)
Profiles 500	0.78	‡	**0.79**	0.76	‡	**0.77**	‡
Profiles 300	0.77	‡	**0.78**	0.76	‡	**0.77**	‡

- Recall and size of predictions

	Recall			aver. $\lvert h(\mathbf{x})\rvert$		
Dataset	nd_MAP	(si)	nd_SVM	nd_MAP	(si.)	nd_SVM
Profiles 500	**0.85**		0.84	1.28	‡	**1.21**
Profiles 300	**0.85**		0.84	1.30	‡	**1.22**

Conclusions

- Nondeterministic classifiers address the problem of deciding what to predict when it is possible to envision that the label returned by a learning algorithm is uncertain

- The utility of these predictions was illustrated in the context of a real world application: the assessment of muscle proportion in beef cattle carcasses

Conclusions

- The job of nondeterministic classifiers is a kind of Information Retrieval
- We derived an algorithm to optimize Fβ measures provided known posterior probabilities
- The algorithm used to estimate these probabilities is very important in the overall performance

Conclusions

- The main advantage of nondeterministic ordinal regressors over their deterministic counterpart
- A dramatic improvement in the proportion of predictions that include the true rank
- The price to be paid for that increase is usually a tiny proportion of predictions with more than one rank

Extension to multi-class Classification

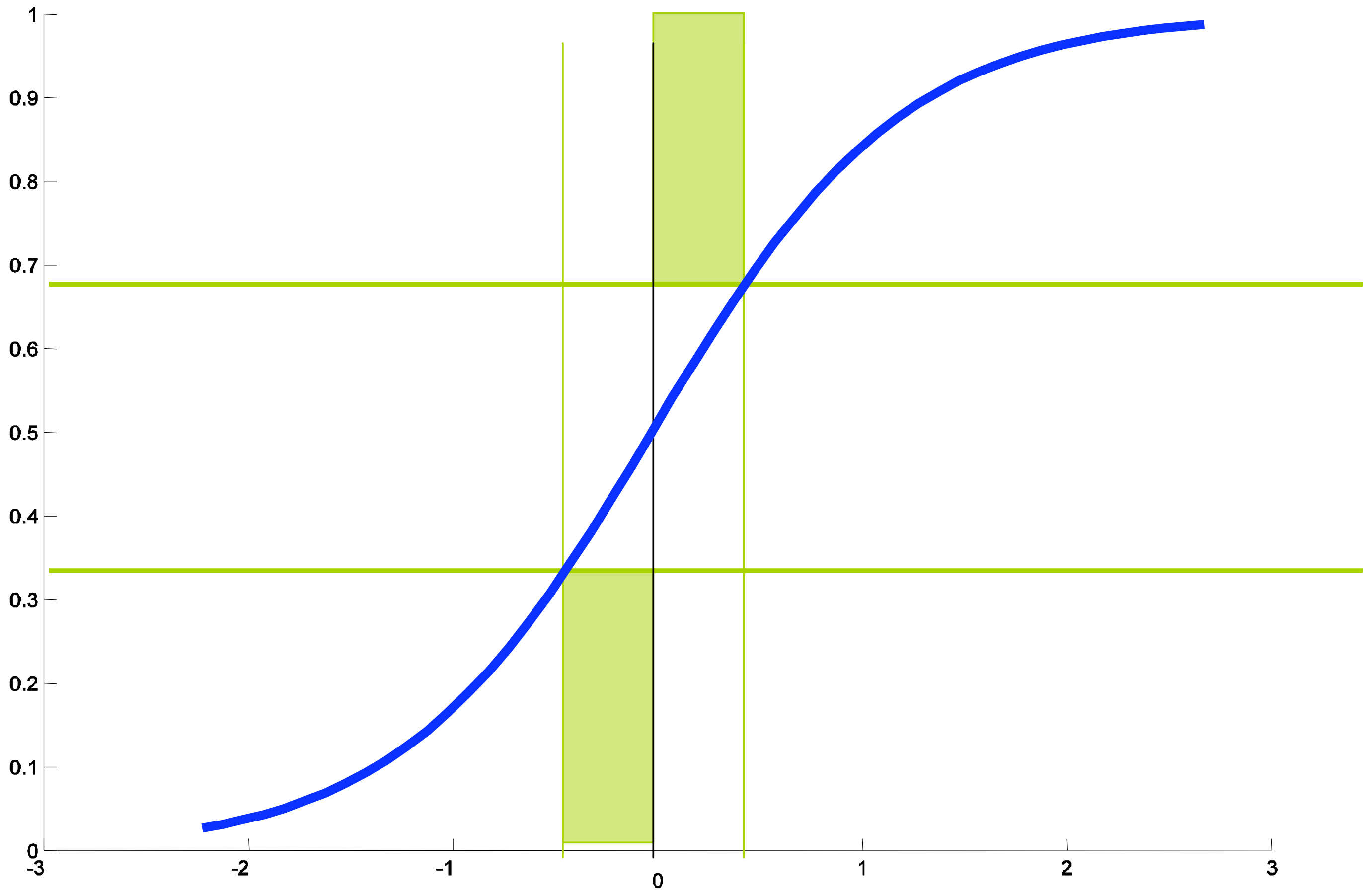
If the true class of x is 1, (y = 1)

$h(\boldsymbol{x})$	$Precision$	$Recall$	F_1	F_2
$[1, 2, 3]$	0.33	1	0.50	0.71
$[1, 2]$	0.50	1	0.67	0.83
$[1]$	1	1	1	1
$[2, 3, 4]$	0	0	0	0

Then the optimal F_1 for a binary classification

$$h_{ND}(\boldsymbol{x}) = \begin{cases} \{-1\} & if \;\; \eta(\boldsymbol{x}) < 1/3 \\ \{-1, +1\} & if \;\; 1/3 \leq \eta(\boldsymbol{x}) < 2/3 \\ \{+1\} & if \;\; 2/3 \leq \eta(\boldsymbol{x}), \end{cases}$$

Extension to multi-class Classification



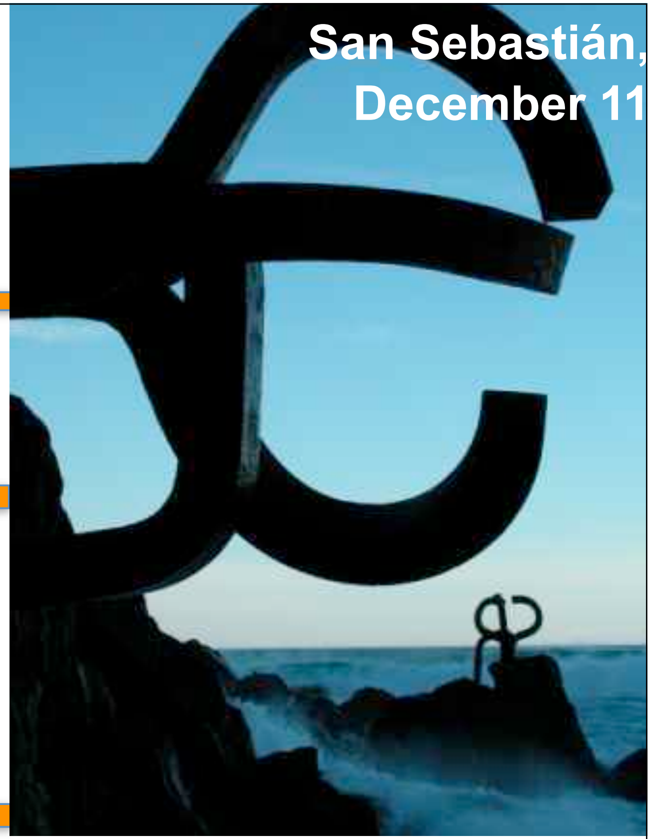
Extension to multi-class Classification

Assuming that:

- we know the posterior probabilities of classes
- classes are ordered according to these probabilities

The optimal hypothesis is given by:

$$h_{ND}(\vec{x}) = \left\{ C_1, \ldots, C_r : \sum_{j=1}^{r} Pr(C_j|\vec{x}) \geq (\beta^2 + r) Pr(C_{r+1}|\vec{x}) \right\}$$



Learning to Predict One or More Ranks or Classes

Antonio Bahamonde

Artificial Intelligence Center. University of Oviedo at Gijón, Asturias, Spain

www.aic.uniovi.es

Spanish Association for Artificial Intelligence (AEPIA)