

Learning to rank from medical imaging data

Fabian Pedregosa, Elodie Cauvet, Gaël Varoquaux, Christophe Pallier,
Bertrand Thirion, Alexandre Gramfort

▶ **To cite this version:**

Fabian Pedregosa, Elodie Cauvet, Gaël Varoquaux, Christophe Pallier, Bertrand Thirion, et al.. Learning to rank from medical imaging data. Third International Workshop on Machine Learning in Medical Imaging - MLMI 2012, INRIA, Oct 2012, Nice, France. hal-00717990v2

HAL Id: hal-00717990
https://inria.hal.science/hal-00717990v2

Submitted on 30 Sep 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Learning to rank from medical imaging data

Fabian Pedregosa[1,2,4], Elodie Cauvet[3,2], Gaël Varoquaux[1,2], Christophe Pallier[3,1,2], Bertrand Thirion[1,2], and Alexandre Gramfort[1,2]

[1] Parietal Team, INRIA Saclay-Île-de-France, Saclay, France
fabian.pedregosa@inria.fr,
[2] CEA, DSV, I^{2}BM, Neurospin bât 145, 91191 Gif-Sur-Yvette, France
[3] Inserm, U992, Neurospin bât 145, 91191 Gif-Sur-Yvette, France
[4] SIERRA Team, INRIA Paris - Rocquencourt, Paris, France

Abstract. Medical images can be used to predict a clinical score coding for the severity of a disease, a pain level or the complexity of a cognitive task. In all these cases, the predicted variable has a natural order. While a standard classifier discards this information, we would like to take it into account in order to improve prediction performance. A standard linear regression does model such information, however the linearity assumption is likely not be satisfied when predicting from pixel intensities in an image. In this paper we address these modeling challenges with a supervised learning procedure where the model aims to order or rank images. We use a linear model for its robustness in high dimension and its possible interpretation. We show on simulations and two fMRI datasets that this approach is able to predict the correct ordering on pairs of images, yielding higher prediction accuracy than standard regression and multiclass classification techniques.

Keywords: fMRI, supervised learning, decoding, ranking

1 Introduction

Statistical machine learning has recently gained interest in the field of medical image analysis. It is particularly useful for instance to learn imaging biomarkers for computer aided diagnosis or as a way to obtain additional therapeutic indications. These techniques are particularly relevant in brain imaging [4], where the complexity of the data renders visual inspection or univariate statistics unreliable. A spectacular application of learning from medical images is the prediction of behavior from functional MRI activations [7]. Such a supervised learning problem can be based on regression [11] or classification tasks [6].

In a classification setting, *e.g.* using support vector machines (SVM) [12], class labels are treated as an unordered set. However, it is often the case that labels corresponding to physical quantities can be naturally ordered: clinical scores, pain levels, the intensity of a stimulus or the complexity of a cognitive task are examples of such naturally ordered quantities. Because classification models treat these as a set of classes, the intrinsic order is ignored leading to suboptimal results. On the other hand, in the case of linear regression models

such as Lasso [13] or Elastic Net [2], the explained variable is obtained by a linear combination of the variables, the pixel intensities in the present case, which benefits to the interpretability of the model [2, 14]. The limitation of regression models is that they assume linear relationship between the data and the predicted variable. This assumption is a strong limitation in practical cases. For instance, in stroke studies, light disabilities are not well captured by the standard NIHSS score. For this reason most studies use a classification approach, forgoing the quantitative assessment of stroke severity and splitting patients in a small number of different classes. A challenge is therefore to work with a model that is able to learn a non-linear relationship between the data and the target variable.

In this paper, we propose to use a *ranking* strategy to learn from medical imaging data. Ranking is a type of supervised machine learning problem that has been widely used in web search and information retrieval [1, 16] and whose goal is to automatically construct an order from the training data. We first detail how the ranking problem can be solved using binary classifiers applied to pairs of images and then provide empirical evidence on simulated data that our approach outperforms standard regression techniques. Finally, we provide results on two fMRI datasets.

Notations We write vectors in bold, $\mathbf{a} \in \mathbb{R}^n$, matrices with capital bold letters, $\mathbf{A} \in \mathbb{R}^{n \times n}$. The dot product between two vectors is denoted $\langle \mathbf{a}, \mathbf{b} \rangle$. We denote by $\|\mathbf{a}\| = \sqrt{\langle \mathbf{a}, \mathbf{a} \rangle}$ the ℓ_2 norm of a vector.

2 Method: Learning to rank with a linear model

A dataset consists of n images (resp. volumes) containing p pixels (resp. voxels). The matrix formed by all images is denoted $\mathbf{X} \in \mathbb{R}^{n \times p}$.

In the supervised learning setting we want to estimate a function f that predicts a target variable from an image, $f : \mathbb{R}^p \to \mathcal{Y}$. For a classification task, $\mathcal{Y} = \{1, 2, 3, ..., k\}$ is a discrete unordered set of labels. The classification error is then given by the number of misclassified images (0-1 loss). On the other hand, for a regression task, $\mathcal{Y}$ is a metric space, typically $\mathbb{R}$, and the loss function can take into account the full metric structure, *e.g.* using the mean squared error.

Here, we consider a problem which shares properties of both cases. As in the classification setting, the class labels form a finite set and as in the regression setting there exists a natural ordering among its elements. One option is to ignore this order and classify each data point into one of the classes. However, this approach ignores valuable structure in the data, which together with the high number of classes and the limited number of images, leads in practice to poor performance. In order to exploit the order of labels, we will use an approach known as *ranking* or *ordinal regression.*

Ranking with binary classifiers Suppose now that our output space $\mathcal{Y} = \{r_1, r_2, ..., r_k\}$ verifies the ordering $r_1 \leq r_2 \leq .. \leq r_k$ and that $f : \mathbb{R} \to \mathcal{Y}$ is our prediction

function. As in [8], we introduce an increasing function $\theta : \mathbb{R} \to \mathbb{R}$ and a linear function $g(\mathbf{x}) = \langle \mathbf{x}, \mathbf{w} \rangle$ which is related to f by

$$f(\mathbf{x}) = r_i \iff g(\mathbf{x}) \in [\theta(r_{i-1}), \theta(r_i)[\tag{1}$$

Given two images $(\mathbf{x}_i, \mathbf{x}_j)$ and their associated labels (y_i, y_j) $(y_i \neq y_j)$ we form a new image $\mathbf{x}_i - \mathbf{x}_j$ with label $\text{sign}(y_i - y_j)$. Because of the linearity of g, predicting the correct ordering of these two images, is equivalent to predicting the sign of $g(\mathbf{x}_i) - g(\mathbf{x}_j) = \langle \mathbf{x}_i - \mathbf{x}_j, \mathbf{w} \rangle$ [8].

The learning problem is now cast into a binary classification task that can be solved using standard supervised classification techniques. If the classifier used in this task is a Support Vector Machine Classifier, the model is also known as *RankSVM*. One of the possible drawbacks of this method is that it requires to consider all possible pairs of images. This scales quadratically with the number of training samples, and the problem soon becomes intractable as the number of samples increases. However, specialized algorithms exist with better asymptotic properties [10]. For our study, we used the Support Vector Machine algorithms proposed by scikit-learn [15].

The main benefit of this approach is that it outputs a linear model even when the function θ is non-linear, and thus ranking approaches are applicable to a wider set of problems than linear regression. Compared to multi-label classification task, where the number of coefficient vectors increase with the number of labels (ranks), the number of coefficients to learn in the pairwise ranking is constant, yielding better-conditioned problems as the number of unique labels increases.

Performance evaluation Using the linear model previously introduced, we denote the estimated coefficients as $\hat{\mathbf{w}} \in \mathbb{R}^p$. In this case, the prediction function corresponds to the sign of $\langle \mathbf{x}_i - \mathbf{x}_j, \hat{\mathbf{w}} \rangle$. This means that the larger $\langle \mathbf{x}_i, \hat{\mathbf{w}} \rangle$, the more likely the label associated to $\mathbf{x}_i$ is to be high. Because the function θ is non-decreasing, one can project along the vector $\hat{\mathbf{w}}$ to order a sequence of images. The function θ is generally unknown, so this procedure does not directly give the class labels. However, under special circumstances (as is the case in our empirical study), for example when there is a fixed number of samples per class this can be used to recover the target values.

Since our ranking model operates naturally on pairs of images, we will define an evaluation measure as the mean number of label inversions. Formally, let $(\mathbf{x}_i, y_i)_{i=1,\ldots,n}$ denote the validation dataset and $\mathcal{P} = \{(i,j) \text{ s.t. } y_i \neq y_j\}$ the set of pairs with different labels. The prediction accuracy is defined as the percentage of incorrect orderings for pairs of images. When working with such a performance metric, the chance level is at 50%:

$$\text{Error} = \#\left\{(i,j) \in \mathcal{P} \text{ s.t. } (y_i - y_j)(f(\mathbf{x}_j) - f(\mathbf{x}_i))\rangle < 0\right\} / \#\mathcal{P} \ . \tag{2}$$

Model selection The Ranking SVM model has one regularization parameter denoted C, which we set by nested cross-validation on a grid of 50 geometrically-spaced values between 10^{-3} and 10^3. We use 5-folds splitting of the data: 60% of

the data is used for training, 20% for parameter selection and 20% for validation. To establish significant differences between methods we perform 20 such random data splits.

3 Results

We present results on simulated fMRI data and two fMRI datasets.

3.1 Simulation study

Data generation The simulated data $\mathbf{X}$ contains $n = 300$ volumes (size $7 \times 7 \times 7$), each one consisting of Gaussian white noise smoothed by a Gaussian kernel with standard deviation of 2 voxels. This mimics the spatial correlation structure observed in real fMRI data. The simulated vector of coefficients $\mathbf{w}$ has a support restricted to four cubic Regions of Interest (ROIs) of size $(2 \times 2 \times 2)$. The values of $\mathbf{w}$ restricted to these ROIs are $\{5, 5, -5, -5\}$.

We define the target value $\mathbf{y} \in \mathbb{R}^n$ as a logistic function of $\mathbf{Xw}$:

$$\mathbf{y} = \frac{1}{1 + \exp(-\mathbf{Xw})} + \epsilon \tag{3}$$

where $\epsilon \in \mathbb{R}^n$ is a Gaussian noise with standard deviation $\gamma > 0$ chosen such that the signal-to-noise ratio verifies $\|\epsilon\|/\|\sqrt{\mathbf{Xw}}\| = 10\%$. Finally, we split the 300 generated images into a training set of 240 images and a validation set of other 60 images.

Results We compare the ranking framework presented previously with standard approaches. Ridge regression was chosen for its widespread use as a regression technique applied to fMRI data. Due to the non-linear relationship between the data and the target values, we also selected a non-linear regression model: support vector regression (SVR) with a Gaussian kernel [5]. Finally, we also considered classification models such as multi-class support vector machines. However, due to the large number of classes and the limited number of training samples, these methods were not competitive against its regression counterpart and were not included in the final comparison.

One issue when comparing different models is the qualitatively different variables they estimate: in the regression case it is a continuous variable whereas in the ranking settings it is a discrete set of class labels. To make both comparable, a score function that is applicable to both models must be used. In this case, we used as performance measure the percentage of incorrect orderings for pairs as defined in (2).

Figure 1-a describes the performance error of the different models mentioned earlier as a function of number of images in the training data. We considered a validation set of 60 images and varied the number of samples in the training set from 40 to 240. With a black dashed line we denote the optimal prediction error, i.e. the performance of an oracle with perfect knowledge of $\mathbf{w}$. This error is

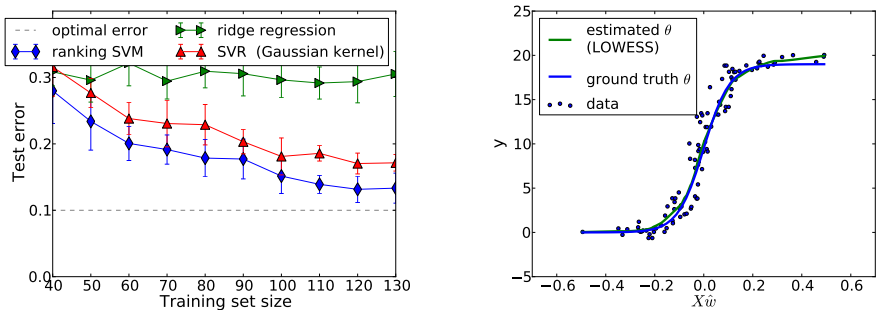


Fig. 1. a) Prediction accuracy as the training size increases. Ranking SVM performs consistently better than the alternative methods, converging faster to the empirical optimal error. b) Estimation of the θ function using non-parametric local regression and ranking SVM. As expected, we recover the logistic function introduced in the data generation section.

non zero due to noise. All of the model parameters were set by cross-validation on the training set. The figure not only shows that Ranking SVM converges to an optimal model (statistical consistency for prediction), but also that it converges faster, *i.e.* has lower sample complexity than alternative approaches, thus it should be able to detect statistical effects with less data. Gaussian kernel SVR performs better than ridge regression and also reaches the optimal error. However, the underlying model of SVR is not linear and is therefore not well-suited for interpretation [2]; moreover, it is less stable to high-dimensional data.

As stated previously, the function θ can be estimated from the data. In Fig. 1-b we use the knowledge of target values from the validation dataset to estimate this function. Specifically, we display class labels as a function of $\mathbf{X}\hat{\mathbf{w}}$ and regularize the result using a local regression (LOWESS). Both estimated function and ground truth overlap for most part of the domain.

3.2 Results on two functional MRI datasets

To assess the performance of ranking strategy on real data, we investigate two fMRI datasets. The first dataset, described in [3], consists of 34 healthy volunteers scanned while listening to 16 words sentences with five different levels of complexity. These were 1 word constituent phrases (the simplest), 2 words, 4 words, 8 words and 16 words respectively, corresponding to 5 levels of complexity which was used as class label in our experiments. To clarify, a sentence with 16 words using 2 words constituents is formed by a series of 8 pairs of words. Words in each pair have a common meaning but there is meaning between each pair. A sentence has therefore the highest complexity when all the 16 words form a meaningful sentence.

The second dataset is described in [17] and further studied in [9] is a gambling task where each of the 17 subjects was asked to accept or reject gambles that offered a 50/50 chance of gaining or losing money. The magnitude of the potential gain and loss was independently varied across 16 levels between trials. No outcomes of these gambles were presented during scanning, but after the scan three gambles were selected at random and played for real money. Each gamble has an amount that can be used as class label. In this experiment, we only considered gain levels, yielding 8 different class labels. This dataset is publicly available from `http://openfmri.org` as the *mixed-gambles task* dataset. The features used in both experiments are SPM β-maps, *a.k.a.* GLM regression coefficients.

Both of these datasets can be investigated with ranking, multi-label classification or regression. We compared the approaches of regression and ranking to test if the added flexibility of the ranking model translates into greater statistical power. Due to the high number of classes and limited number of samples, we found out that multi-label classification did not perform significantly better than chance and thus was not further considered.

To minimize the effects that are non-specific to the task we only consider pairs of images from the same subject.

The first dataset contains four manually labeled regions of interest: Anterior Superior Temporal Sulcus (aSTS), Temporal Pole (TP), Inferior Frontal Gyrus Orbitalis (IFGorb) and Inferior Frontal Gyrus triangularis (IFG tri). We then compare ranking, ridge regression and Gaussian kernel SVR models on each ROI separately. Those results appear in the first three rows of Table 1 and are denoted as language complexity with its corresponding ROI in parenthesis. We observe that the ranking model obtains a significant advantage on 3 of the 4 ROIs. This could be explained by a relatively linear effect in IFGorb or a higher noise level.The last row concerns the second dataset, denoted gambling, where we selected the gain experiment (8 class labels). In this case, since we were not presented manually labeled regions of interest, we performed univariate dimensionality reduction to 500 voxels using ANOVA before fitting the learning models. It can be seen that the prediction accuracy are lower compared to the first dataset. Ranking SVM however still outperforms alternative methods. Locations in the brain of the ROIs for the first dataset, with a color coding for their predictive power are presented in Fig. 2-a. As shown previously in the simulated dataset, Fig. 2-b shows the validation data projected along the coefficients of the linear model and regularized using LOWESS local regression for each one of the four highest ranked regions of interest. Results show that the link function between $\mathbf{X}\hat{\mathbf{w}}$ and the target variable $\mathbf{y}$ (denoted θ in the methods section) varies in shape across ROIs, suggesting that the BOLD response is not unique over the brain.

	RankSVM	Ridge	SVR	P-val
lang. comp (aSTS)	**0.706**	0.661	0.625	2e-3***
lang. comp. (TP)	**0.687**	0.645	0.618	7e-4***
lang. comp. (IFGorb)	**0.619**	0.609	0.539	0.3
lang. comp. (IFG tri)	**0.585**	0.566	0.533	5e-2*
gambling	**0.58**	0.56	0.53	1e-2**

Table 1. Prediction accuracy of the ranking strategy on two real fMRI datasets: language complexity (lang. comp.) in 3 ROIs and gambles. As a comparison, scores obtained with alternative regression techniques (ridge regression and SVR using a non-linear Gaussian kernel) are presented. As confirmed by a Wilcoxon paired test between errors obtained for each fold using Ranking SVM and ridge regression, Ranking SVM leads to significantly better scores than other approaches on 4 of the 5 experiments.

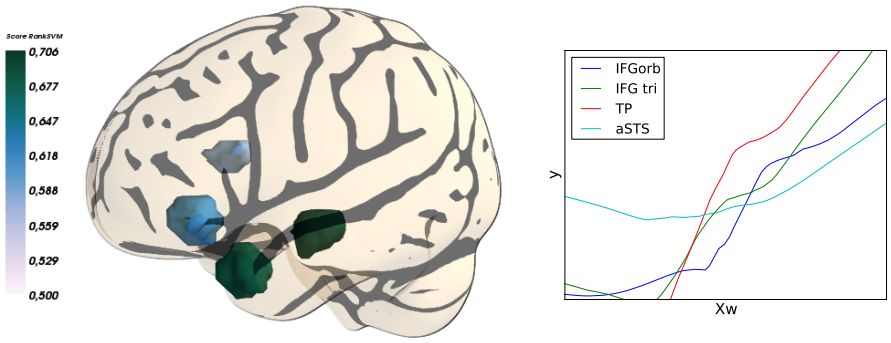


Fig. 2. a) Scores obtained with the Ranking SVM on the 4 different ROIs. The regions with the best predictive power are the temporal pole the anterior superior temporal sulcus. b) The target variable $\mathbf{y}$ as a function of $\mathbf{X}\hat{\mathbf{w}}$ for the four regions of interest. We observe that the shape of the curves varies across brain regions.

4 Discussion and conclusion

In this paper, we describe a ranking strategy that addresses a common use case in the statistical analysis of medical images, which is the prediction of an ordered target variable. Our contribution is to formulate the variable quantification problem as a ranking problem. We present a formulation of the ranking problem that transforms the task into a binary classification problem over pairs of images. This approach makes it possible to use efficient linear classifiers while coping with non-linearities in the data. By doing so we retain the interpretability and favorable behavior in high-dimension of linear models.

From a statistical standpoint, mining medical images is challenging due to the high dimensionality of the data, often thousands of variables, while the number of images available for training is small, typically a few hundreds. In this regard, the benefit of our approach is to retain a linear model with few parameters. It is thus better suited to medical images than multi-class classification.

On simulations we have shown that our problem formulation leads to a better prediction accuracy and lower sample complexity than alternative approaches such as regression or classification techniques. We apply this method to two fMRI datasets and discuss practical considerations when dealing with fMRI data. We confirm the superior prediction accuracy compared to standard regression techniques.

References

1. Burges, C.J.C.: From RankNet to LambdaRank to LambdaMART : An overview. Learning 11(MSR-TR-2010-82), 23–581 (2010)
2. Carroll, M.K., Cecchi, G.A., Rish, I., Garg, R., Rao, A.R.: Prediction and interpretation of distributed neural activity with sparse models. NeuroImage 44(1), 112–122 (2009)
3. Cauvet, E.: Traitement des Structures Syntaxiques dans le langage et dans la musique. Ph.D. thesis, Ecole doctorale n158, Cerveau - Cognition - Comportement (2012)
4. Cuingnet, R., Rosso, C., Chupin, M., Lehéricy, S., Dormont, D., Benali, H., Samson, Y., Colliot, O.: Spatial regularization of SVM for the detection of diffusion alterations associated with stroke outcome. Medical Image Analysis (2011)
5. Drucker, H., Burges, C.J.C., Kaufman, L., Smola, A.J., Vapnik, V.: Support vector regression machines. In: NIPS. pp. 155–161 (1996)
6. Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., Pietrini, P.: Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex. Science 293(5539), 2425–2430 (2001)
7. Haynes, J.D., Rees, G.: Decoding mental states from brain activity in humans. Nat. Rev. Neurosci. 7, 523 (2006)
8. Herbrich, R., Graepel, T., Obermayer, K.: Large margin rank boundaries for ordinal regression, vol. 88, pp. 115–132. MIT Press, Cambridge, MA (2000)
9. Jimura, K., Poldrack, R.A.: Analyses of regional-average activation and multivoxel pattern information tell complementary stories. Neuropsychologia pp. 1–9 (2011)
10. Joachims, T.: Training linear SVMs in linear time. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 217–226. KDD '06, ACM, New York, NY, USA (2006)
11. Kay, K.N., Naselaris, T., Prenger, R.J., Gallant, J.L.: Identifying natural images from human brain activity. Nature 452, 352–355 (2008)
12. LaConte, S., Strother, S., Cherkassky, V., Anderson, J., Hu, X.: Support vector machines for temporal classification of block design fMRI data. NeuroImage 26(2), 317 – 329 (2005)
13. Liu, H., Palatucci, M., Zhang, J.: Blockwise coordinate descent procedures for the multi-task lasso, with applications to neural semantic basis discovery. In: Proceedings of the 26th Annual International Conference on Machine Learning. pp. 649–656. ICML '09, ACM, New York, NY, USA (2009)
14. Michel, V., Gramfort, A., Varoquaux, G., Eger, E., Thirion, B.: Total variation regularization for fMRI-based prediction of behaviour. IEEE Transactions on Medical Imaging 30(7), 1328 – 1340 (Feb 2011)
15. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine Learning in Python . Journal of Machine Learning Research 12, 2825–2830 (2011)

16. Richardson, M., Prakash, A., Brill, E.: Beyond PageRank: machine learning for static ranking, pp. 707–715. WWW '06, ACM, New York, NY, USA (2006)
17. Tom, S.M., Fox, C.R., Trepel, C., Poldrack, R.A.: The neural basis of loss aversion in decision-making under risk. Science 315(5811), 515–518 (2007)