Non-linear correspondence analysis in text retrieval: a kernel view

Davide Picca, Benoît Curdy, François Bavaud

University of Lausanne - CH-1015 Lausanne - Switzerland

davide.picca@unil.ch, benoit.curdy@unil.ch, francois.bavaud@unil.ch

Abstract

Classical factorial treatments applied on words-documents counts matrices (such as Correspondence Analysis (FCA), Latent Semantic Indexing (LSI), as well as non-linear generalizations of FCA (NLCA)) can be described in the framework of kernels associated to Support Vector Machines (SVM). This paper exposes the relationships between those formalisms, and demonstrates how textual pre-processing by a “power kernel” can improve (with respect to the classical FCA kernel) the documents classification in the Reuters-21578 corpus.

Resumé

Les traitements factoriels classiques de la matrice d’occurences termes-documents (tels que l’analyse des correspondances (AFC) et le “latent semantic indexing”, ainsi que la généralisation non-linéaire de l’AFC (NLCA)) peuvent être décrits dans le cadre du formalisme des noyaux intervenant dans les SVM (Support Vector Machines). Ce travail précise les liens entre ces formalismes, et montre comment le pré-traitement des données textuelles par un “noyau puissance” permet d’améliorer (relativement au noyau classique de l’AFC) la classification semi-automatique des documents dans le corpus Reuters-21578.

Keywords: Correspondence analysis, Information retrieval, Latent semantic indexing, Kernel methods, Support vector machines, Reuters-21578 corpus

Introduction

Support vector machines (SVM) approaches to pattern recognition, supervised and unsupervised classification, and information retrieval have been vigorously expanding during the last two decades within the computer scientific community, following Vapnik’s work on separability and generalizability in neural network architectures (Vapnik, 1998).

In a nutshell, SVM binary classification consists in learning the *maximum margin hyperplane* that separates the two classes, instead than using classical statistical decision rules, such as the Fisher’s linear or quadratic discriminant rules. When the classes are non-separable, the problem can be cast in terms of “soft” maximum margin separation, allowing for training errors; it also can be embedded (generally in conjunction with the soft approach) in a higher dimensional space endowed with better separability properties through a non-linear mapping $x \mapsto \Phi(x)$ transforming the initial objects’ profiles x into features $\Phi(x)$.

Kernels are scalar products $K_{jj'} :=< \Phi(x_j), \Phi(x_{j'}) >$ between features associated to objects j and j', and constitute a measure of their similarity. By construction, associated dissimilarities $D_{jj'} := K_{jj} + K_{j'j'} - 2K_{jj'}$ constitute squared Euclidean distances in the feature space. Much of the importance of the kernels lies in the "kernel trick" observation (see e.g. Schölkopf et al., 1998) that most classical data-analytic algorithms (factor analysis, multidimensional scaling, discriminant analysis, clustering), as well as the problem of the maximum margin determination itself, make an explicit use of $K_{jj'}$ only, without requiring the knowledge of the (possibly involved) transformation $\Phi(x)$.

Bavaud (Bavaud, 2004) has recently proposed that classical correspondence analysis can be generalized by performing the spectral decomposition of a matrix of scalar products between non-linearly transformed row profiles while preserving aggregation-invariance (aka "équivalence distributionnelle"), thus providing well-founded alternatives to the chi-square dissimilarity measure $D^{\chi}_{jj'}$ between objects. The next section recalls the details of the approach, its relationship to latent semantic indexing, and shows how it fits into a profiles-to-features transformation view $x \to \Phi(x)$ with associated kernel $< \Phi(x_j), \Phi(x_{j'}) >$, the case $\Phi(x) = x$ being correspondence analysis in its usual form. The section *Evaluation* demonstrates that such non-linear transformations can indeed improve semi-supervised classification, as shown on the Reuters-21578 corpus example.

Linear and non-linear correspondence analysis

Let $N = (n_{ij})$ denote the $(v \times p)$ the *word* × *document* matrix, counting the number of occurences of word i in document j. Define the word weights as $\pi_i = n_{i\bullet}/n$ and the document weights as $\rho_j = n_{\bullet j}/n$, where "•" denotes summation and $n = n_{\bullet\bullet}$ is the total size of the corpus. Define the representation of document j in the v-dimensional words space by the vector $\Phi(x_j) = (\varphi_{ij})$ whose i-th component ($i = 1, \ldots, v$) reads $\varphi_{ij} := q_{ij} - 1$, where $q_{ij} := (n \cdot n_{ij})/(n_{i\bullet} \cdot n_{\bullet j})$ is the *independence quotient* associated to component (ij), that is the ratio of the observed count over its expected value under independence.

By construction $\varphi_{ij} > 0$ iff word i appears more often in document j than expected under independence. The weighted scalar product $K_{jj'} := \sum_i \pi_i \varphi_{ij} \varphi_{ij'}$ constitutes a measure of similarity between documents j and j'. It turns out that the associated (squared) euclidean distance $K_{jj} + K_{j'j'} - 2K_{jj'}$ is nothing but the well known chi-square dissimilarity D^{χ} (Bavaud, 2004). Low-dimensional projection of the total inertia $\Delta := \frac{1}{2} \sum_{jj'} \rho_j \rho_{j'} D_{jj'}$ leads to define the $(p \times p)$ matrix $\tilde{K} = (\tilde{K}_{jj'})$ with components $\tilde{K}_{jj'} = \sqrt{\rho_j}\sqrt{\rho_{j'}} K_{jj'}$, whose spectral decomposition $\tilde{K} = U \Lambda U'$ (where $U = (u_{j\alpha})$ is diagonal and $\Lambda = \text{diag}(\lambda)$ diagonal) permits to define the usual objects coordinates $x_{j\alpha} := \sqrt{\lambda_\alpha} u_{j\alpha}/\sqrt{\rho_j}$ of Correspondence Analysis (FCA), obeying $\sum_\alpha (x_{j\alpha} - x_{j'\alpha})^2 = D^{\chi}_{jj'}$.

As observed by Cristianini et al. (Cristianini et al., 2002), Latent Semantic Indexing (LSI) can also be approached from a kernel perspective. In its original set-up (G. and C.Buckley, 1988), LSI uses a document-document similarity index $s_{jj'}$ defined as simply as $S = (s_{jj'}) = N'N$. Later developments led to modify this so-called *basic vector space model* by considering transformations of the form $N \to \Phi(N) = PNQ$ (where Q is a square matrix), yielding document similarities or kernels of the form $S := \Phi'(N)\Phi(N) = Q'N'P'PNQ$. Comparison with the FCA kernel $\tilde{K}$ above shows the spectral decompositions of S and $\tilde{K}$ to coincide (up to the trivial eigenvector $\sqrt{\rho_j}$ associated to the eigenvalue zero of $\tilde{K}$) provided P is $(v \times v)$ with components $p_{ii'} = \delta_{ii'}/\sqrt{n_{i\bullet}}$ and Q is $(p \times p)$ with components $q_{jj'} = \delta_{jj'}/\sqrt{n_{\bullet j}}$. The choices $p_{ii'} = \delta_{ii'}(p/p_i)$ or $p_{ii'} = \delta_{ii'} \log(p/p_i)$ (where p is the total number of documents and p_i the number of documents in which word i occurs) yield the so-called inverse document frequency (idf) weighting (see e.g. Manning et al., 1999 and references therein). The choice $P = N$ (or more generally choices such that $p_{ij} = 0$ whenever $n_{ij} = 0$) lead to the so-called co-occurences approaches (see e.g. Besançon,et al., 1999 ,Bavaud et

al., 2005).
Non-Linear Correspondence Analysis (NLCA) (Bavaud, 2004) can be introduced within the present formalism by definining features as $\varphi_{ij} := f(q_{ij})$, where $f(q)$ represents *any* increasing function with $f(1) = 0$, with associated kernel $K_{jj'} = \sum_i \pi_i \varphi_{ij} \varphi_{ij'}$. For instance, $f(q) = I(q > 0)$ defines the *presence-absence* kernel, and $f(q) = (q^\beta - 1)/\beta$ (for some $\beta > 0$) defines the *power* kernel, where $\beta = 1$ yields ordinary correspondence analysis and $\beta \to 0$ yields the logarithmic kernel $f(q) = \ln q$ investigated in (Aitchison et al., 2002). The elementary fact that $f(q_{ij})$ depends on the quotients q_{ij} only suffices to establish the aggregation-invariance of the resulting method (Bavaud, 2004).

Evaluation

Data

In order to assess the possible pragmatic virtues of NLCA, we have chosen a text categorization task, a supervised technique which attempts to classify documents belonging to pre-defined number of categories. Somehow arbitrarily, we decided to focus on the power kernel $f(q) = \frac{1}{\beta}(q^\beta - 1)$ and to compare three different values $\beta = 1.25$, $\beta = 1$ (which yields ordinary FCA) and $\beta = 0.75$. For purposes of comparison with other publications in Natural Language Processing, we have used the well-known *Reuters-21578* corpus. It includes 12'902 documents categorized into 90 topics. We choosed to use the training set of the widely used *ModApte split* of the *Reuters* corpus[1], retaining the 10 more frequent document categories only, resulting in an unbalanced size distribution shown on table 1. To limit the dimensionality v of documents profiles, rare words (such that $n_{i\bullet} < 3$) have been eliminated; also, very common words (stop-words) have also been removed from the data. The resulting words-by document matrix N consists of $v = 7'517$ words and $p = 5'918$ documents.

category	number of documents	relative frequency
earn	2689	0.45
acq	1462	0.25
Money-fx	330	0.06
crude	320	0.05
trade	307	0.05
interest	265	0.04
ship	168	0.03
wheat	141	0.02
corn	129	0.02
grain	107	0.02
total	5918	1

Table 1: *distribution of the ten most frequent categories within a set of 5'918 documents from the Reuters-21578 corpus*

Implementation details

The data preprocessing is the core of our experiment. Our aim is to obtain a good features representation of the documents corpus in which the document classes are well separated.
The first step was to transform the counts n_{ij} of the words-document matrix into quotients $q_{ij} = (n_{ij} n)/(n_{i\bullet} n_{\bullet j})$. Then quotients profiles were non-linearly transformed by constructing the $(v \times p)$ features matrix $\tilde{\Phi} = (\tilde{\varphi}_{ij})$ where $\tilde{\varphi}_{ij} := \sqrt{\pi_i}\sqrt{\rho_j} f(q_{ij})$ and $f(q) = \frac{1}{\beta}(q^\beta - 1)$. A Singular Value Decomposition $\tilde{\Phi} = V \Gamma U'$ was then performed, yielding words coordinates as $y_{i\alpha} := \gamma_\alpha v_{i\alpha}/\sqrt{\pi_i}$

[1] Compiled by David D. Lewis

and document coordinates as $x_{j\alpha} := \gamma_\alpha u_{j\alpha}/\sqrt{\rho_j}$. Of course, the latter coincide with the expressions given in the introduction, in view of the identity $\tilde{K} = \tilde{\Phi}'\tilde{\Phi} = U\Gamma V'V\Gamma U' = U\Lambda U'$ with $\Lambda = \Gamma^2$.

In the LSI tradition, authors keep usually between 50 and 400 factors (Landauer et al., 1998); we retained the first 50 factors only (out of 5'918). The documents can now be represented in a fifty-dimensional features space by the values $x_{j\alpha}$, the coordinates of document j in dimension $\alpha = 1, \ldots, 50$. However, in order to retain the spirit of LSI where each document is projected in a vector space such that each dimension is associated to a word, we choosed the alternative document representation $x^*_{j\alpha} := \sum_{i=1}^{v}(n_{ij}/n_{\bullet j})y_{i\alpha}$, that is each document lies at the center of gravity of the words it contains.

Starting from a randomly selected learning sample of documents containing 66% of the total, maximum margin hyperplanes best separating the ten classes of documents in the x^* space were then computed by means of a SVM. Specifically, we worked with the SMO algorithm (Platt et al., 1998) implemented in the software `Weka` (Witten, et al.,2005), which assigns each test document to a single category. Also, we used the "polynomial kernel of degree 1" option, that is a documents dissimilarity measure defined as $\sum_\alpha x^*_{j\alpha}x^*_{j'\alpha}$, although other kernels (such as the gaussian kernels $\exp(-\sum_\alpha(x^*_{j\alpha} - x^*_{j'\alpha})^2/\sigma^2)$) could have been used as well, resulting in a supplementary non-linear transformation of document features. The results of the categorisation are shown in the next subsection.

As the document distribution is fairly unbalanced, an undiscriminating classification rule systematically assigning all documents to the most frequent category would yield a of correct classification proportion as high as .45. To assess the real quality of the classification and substract the part of correct classification due to random attribution, we used the "Cohen's kappa" index defined as

$$\kappa := \frac{\sum_l c_{ll} - \sum_l c_{ll}^{\text{random}}}{\sum_{ll'} c_{ll'} - \sum_j c_{ll}^{\text{random}}} \qquad\qquad c_{ll}^{\text{random}} := \frac{c_{l\bullet}c_{\bullet l}}{c_{\bullet\bullet}}$$

where the components $c_{ll'}$ of the *confusion matrix* C count the number of documents belonging to category l and classified as l'. By construction, $\kappa = 1$ indicates a perfect classification, and $\kappa = 0$ indicates a classification performing not better than a random classification.

Others well-known measures have been considered like the *recall*, the *precision* and the F*-measure*. Recall is TN/(TN+TP) and precision is TP/(TP+FP), where TN is the number of true negative, TP the number of true positive and FP the number of false positive. The F-measure is the harmonic mean of recall and precision, that is $F = (2 \times \text{precision} \times \text{recall})/(\text{precision} + \text{recall})$.

Results

Tables 2 to 7 show the classification results obtained for various values of β, and clearly demonstrate that the best classifications are obtained when using a small β. For instance, Cohen's kappa is 0.28 for $\beta = 1.25$, 0.81 for $\beta = 1$ and 0.86 for $\beta = 0.75$.

When β is large, the most frequent categories are clearly systematically preferred (see 2): no document is classified outside the first two categories. When β gets smaller, the confusion matrix evolves and categories of smaller size start being chosen, resulting in better classification results as demonstrated form the various measures defined in the previous subsection.

classified as →	a	b	c	d	e	f	i	j	k	l
a = earn	836	70	0	0	0	0	0	0	0	0
b = acq	176	320	0	0	0	0	0	0	0	0
c = Money-fx	122	2	0	0	0	0	0	0	0	0
d = crude	106	3	0	0	0	0	0	0	0	0
e = trade	98	0	0	0	0	0	0	0	0	0
f = interest	91	0	0	0	0	0	0	0	0	0
g = ship	56	2	0	0	0	0	0	0	0	0
h = wheat	45	2	0	0	0	0	0	0	0	0
i = corn	49	0	0	0	0	0	0	0	0	0
j = grain	35	0	0	0	0	0	0	0	0	0

Table 2: *confusion matrix for* $\beta = 1.25$*, with* $\kappa = 0.28$

Conclusion

Despite marked differences in their temporal, geographical and disciplinary origins, both SVM and "Analyse des Données" methodologies share the same the data-driven perspective (as opposed to the probabilistic modelling approach) as well as a strong geometric flavour. It hence comes as no surprise that both formalisms possess in common a number of identical or similar features.

From an algebraic point of view, kernel dissimilarities with $\beta < 1$ attenuate the effect of high quotients $q_{ij} > 1$ (denoting that word i is typical of document j), and increase the effect of low quotients $q_{ij} < 1$ (denoting that word i seldom occurs in document j). From this perspective, low-β kernels tend to *regularize* the cloud of documents in the features space, in the sense that atypical representatives in the classical FCA approach tend to come closer to the origin. This being said, the crucial mechanism explaining why this circumstance obviously helps in separating document classes has yet to be fully identified. Also, the way in which the different options (choice of the transformation $f(q)$; choice of the "x-space" instead of the "x^*-space" for representing documents; choice of a further kernel in the SVM classification procedure) influence (or not) the quality of the classification is barely addressed yet, and would require further investigations.

Appendix

classified as →	a	b	c	d	e	f	i	j	k	l
a = earn	866	39	0	0	0	1	0	0	0	0
b = acq	16	471	0	8	1	0	0	0	0	0
c = Money-fx	2	1	69	0	13	39	0	0	0	0
d = crude	5	8	3	85	3	0	5	0	0	0
e = trade	4	0	5	0	88	0	0	0	1	0
f = interest	2	2	12	0	0	75	0	0	0	0
g = ship	0	6	1	5	4	1	36	5	0	0
h = wheat	1	1	0	0	1	0	0	36	4	4
i = corn	0	3	1	0	5	0	0	30	10	0
j = grain	1	0	0	0	5	0	2	19	5	3

Table 3: *confusion matrix for* $\beta = 1$*, with* $\kappa = 0.81$

classified as →	a	b	c	d	e	f	i	j	k	l
a = earn	866	19	0	1	0	0	0	0	0	0
b = acq	16	472	0	7	1	0	0	0	0	0
c = Money-fx	3	2	95	0	10	14	0	0	0	0
d = crude	3	9	3	85	2	0	7	0	0	0
e = trade	1	0	5	1	90	0	0	0	1	0
f = interest	2	1	13	0	0	75	0	0	0	0
g = ship	0	5	0	5	0	0	43	2	0	3
h = wheat	1	0	0	0	1	0	1	34	3	7
i = corn	0	2	1	0	2	0	0	13	26	5
j = grain	0	0	0	0	2	0	4	14	5	10

Table 4: *confusion matrix for* $\beta = 0.75$*, with* $\kappa = 0.86$

category	recall	precision	F-measure
earn	0.965	0.956	0.961
acq	0.802	0.645	0.715
Money-fx	0	0	0
crude	0	0	0
trade	0	0	0
interest	0	0	0
ship	0	0	0
wheat	0	0	0
corn	0	0	0
grain	0	0	0

Table 5: *precision and recall for* $\beta = 1.25$

category	recall	precision	F-measure
earn	0.965	0.956	0.961
acq	0.887	0.95	0.917
Money-fx	0.758	0.556	0.642
crude	0.867	0.78	0.821
trade	0.733	0.898	0.807
interest	0.647	0.824	0.725
ship	0.837	0.621	0.713
wheat	0.4	0.766	0.526
corn	0.5	0.204	0.29
grain	0.429	0.086	0.143

Table 6: *precision and recall for* $\beta = 1$

category	recall	precision	F-measure
earn	0.971	0.978	0.975
acq	0.925	0.952	0.938
Money-fx	0.812	0.766	0.788
crude	0.859	0.78	0.817
trade	0.833	0.918	0.874
interest	0.843	0.824	0.833
ship	0.782	0.741	0.761
wheat	0.54	0.723	0.618
corn	0.743	0.531	0.619
grain	0.4	0.286	0.333

Table 7: *precision and recall for* $\beta = 0.75$

References

Aitchison J. and Geenacre M. (2002). Biplots for compositional data. in *Applied Statistics 51* : 375-382.

Bavaud F. (2004). Generalized factor analyses for contingency table. in *Classification, Clustering and Data Mining Applications*, D.Banks et al. Eds. : 597-606.

Bavaud F. and Xanthos A. (2005). Markov associativities. in *Journal of Quantitative Linguistics 12* : 123-137.

Besançon R., Rajman M. and Chappelier J.-C. (1999). Textual Similarities based on a Distributional Approach. in *Proceedings of the Tenth International Workshop on Database and Expert Systems Applications (DEXA99)* : 180-184.

Buckley G. and Buckley C. (1988). Term weighting approaches in automatic text retrieval, in *Information Processing and Management 24* : 513-523.

Cristianini N. , Shawe-Taylor J. and Lodhi H. (2001). Latent Semantic Kernels. in *Proceedings of ICML-01, 18th International Conference on Machine Learning* : pp. 66-73.

Gliozzo A.M. and Strapparava C. (2005). Domain Kernels for Text Categorization. in *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*. Ann Arbor, Michigan, 2005 : pp. 56-63, .

Landauer T.K. and Foltz P.W. and Laham D. (1998). An introduction to Latent Semantic Analysis. in *Discourse Processes 25* : 259-284.

Manning C. and Schütze H. (1999). Foundations of Statistical Natural Language Processing, MIT Press.

Platt J. (1998). Fast training of support vector machines References using sequential minimal optimization. Advances in Kernel Methods-Support Vector Learning. Cambridge, MA:MIT Press.

Schölkopf B., Smola A. and Müller K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. in *Neural Computation 10* : 1299-1319.

Vapnik V. (1998). Statistical Learning Theory, Wiley.

Witten I.H. and Frank E. (2005). Data Mining: Practical machine learning tools and techniques. 2nd Edition. Morgan Kaufmann.