A LEARNING ALGORITHM WITH COMPRESSION-BASED REGULARIZATION

Matias Vera[⋆] *Leonardo Rey Vega*[⋆†] *Pablo Piantanida*[⋆]

[⋆] Facultad de Ingeniería - Universidad de Buenos Aires, Buenos Aires, Argentina
[†] CSC-CONICET, Buenos Aires, Argentina
[⋆] CentraleSupélec - CNRS - Université Paris-Sud, Gif-sur-Yvette, France

ABSTRACT

This paper investigates, from information theoretic principles, a learning problem based on the principle that any regularity in a given dataset can be exploited to extract compact features from data, in order to build meaningful representations of a relevant content. We begin by introducing the fundamental tradeoff between the average risk and the model complexity. Interestingly, our formulation allows an information theoretic formulation of the *multi-task learning* (MTL) problem. Then, we present an iterative algorithm for computing the optimal tradeoffs. Remarkably, empirical results illustrate that there exists an optimal information rate minimizing the *excess risk* which depends on the nature and the amount of available training data. An application to hierarchical text categorization is also investigated, extending previous works.

Index Terms— Information Bottleneck, Arimoto-Blahut Algorithm, Multi-Task Learning, Side Information

1. INTRODUCTION

The actual goal of learning is neither accurate estimation of model parameters; rather, we are interested in the generalization capabilities, i.e., its ability to successfully apply rules extracted from previously seen data to characterize unseen data. It is known that complex models tend to produce *overfitting*, i.e., represent the training data too accurately, therefore diminishing their ability to handle unseen data. To overcome this issue, regularization methods include parameter penalization, noise addition, and averaging over multiple models trained with different sample sets. Nevertheless, it is still not clear how to optimally control the model complexity making this problem an active research topic in machine learning.

Shannon [1] provides a function for measuring the distortion (or loss) between the original signal and its compressed representation. The rate-distortion function is related to a similarity measure in cluster analysis and has demonstrated substantial performance improvement over standard learning methods (see [2] and references therein). Tishby *et al.* [3] associated this information-theoretic setup to a learning problem with a specific loss function. The idea of the so-called

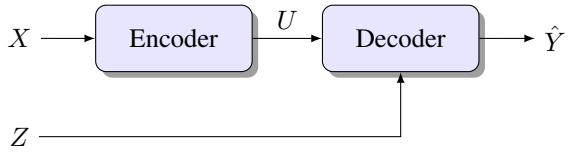

Fig. 1. Block diagram of IB problem with side information.

Information Bottleneck (IB) method is to identify relevant information from observed samples as being the information that those observations provide about another hidden signal. Since then, it was applied to derive several clustering algorithms for a wide variety of problems such as: text classification [4], speaker recognition [5], among others.

The algorithm for computing the classical rate-distortion problem was developed independently by Arimoto [6] and Blahut [7]. Although this algorithm can be applied to the IB criterion [3], we emphasize that conventional algorithms [6, 7] are only expected to converge to a local minimum since the IB is a non-convex problem. Chechik *et al.* [8] adapts a Blahut–Arimoto algorithm to include a restricted form of side information. In a different but related optimization problem, Kumar and Thangaraj [9] adapt the Blahut–Arimoto algorithm and analysis techniques provided in [10] to a non-convex problem. Yasui and Matsushima [11] extend the Kumar idea for computing information-rate regions.

MTL [12] is an approach to inductive transfer that improves generalization by using the domain information contained in the training signals of related tasks as an inductive bias. It does this by learning tasks in parallel while using a shared data representation. What is learned for each task can help other tasks to be learned better and thus can result in improved efficiency and prediction accuracy when compared to training the models separately [13]. An information-theoretic environment for MTL was introduced in [14].

In this paper, we first introduce an information-theoretic paradigm which provides the fundamental tradeoff between the *log-loss* (average risk) and the information rate of the features (statistical model complexity). We derive an iterative Arimoto-Blahut like algorithm to approach the IB problem when there is side information only at the decoder, as described in Fig. 1. It is important to mention that this IB

method formulation as a noisy source coding problem with side information at the decoder provides an information theoretic view of the MTL problem providing an interesting link between the areas of machine learning and Shannon theory. In our framework, the encoder aims at extracting relevant (common) information U from a data set X about labels $Y|Z=z$ which depends on several related tasks Z at the decoder. The representations are expected to summarize the data X in a compact way, where compactness of the model is measured in terms of the minimum Shannon entropy rate. An important property of our approach is that it provides a natural safeguard against overfitting by minimizing an average risk penalized by the model complexity. Remarkably, empirical results illustrate that there exists an optimal information rate minimizing the *excess risk* which depends on the nature and the amount of available training data. We evaluate the performance of this algorithm on hierarchical text categorization of documents and numerical results demonstrates its merits in terms of the classification performance. Major mathematical details of the results of this papers can be found in [15].

2. PROBLEM DEFINITION AND MAIN RESULT

Let (X,Y,Z) be random variables with joint probability mass function P_{XYZ}. A random encoder $P_{U|X}$ wishes to extract from X information about a collection of labels $Y|Z=z$, with $z\in\mathcal{Z}$, when the randomly chosen task-request index Z is available only at the decoder, as shown in Fig. 1. The best decoder $P_{Y|UZ}$ will depend on the encoder selection $P_{U|X}$ and P_{XYZ} and is given by

$$P_{Y|UZ}=\frac{\sum_x P_{U|X}P_{XYZ}}{\sum_x P_{U|X}P_{XZ}}. \tag{1}$$

It is easy to see that finding the encoder that minimizes the *log-loss risk* is equivalent to search for the encoder maximizing the mutual information $I(Y;U|Z)$ (relevance). We will focus on maximizing this mutual information subject to a bound on the complexity (Shannon information rate) $I(X;U|Z)$. The reason why these concepts are defined in this way is evident in the multi-letter characterization [16].

Definition 1 (Rate-Relevance Region)**.** *A pair rates (R,μ) is achievable iff it belongs to the rate-relevance region:*

$$\begin{aligned}\mathcal{R}=\big\{(\mu,R)\in\mathbb{R}^2_{\geq 0}:\ \exists\, P_{U|X}/\quad & R\geq I(X;U|Z),\\ & \mu\leq I(Y;U|Z),\quad P_{XYZU}=P_{U|X}P_{XYZ}\big\}\end{aligned} \tag{2}$$

The corresponding relevance-rate function is defined by

$$L(R,P_{XYZ})=\max_{P_{U|X}:\, I(U;X|Z)\leq R} I(U;Y|Z). \tag{3}$$

We easily see that the relevance-rate function, as the upper-boundary of $\mathcal{R}$ provides an alternative and complete characterization of this region. Using duality theory, (3) can be shown to be equivalent:

$$V_\lambda=\max_{P_{U|X}}\ \lambda I(Y;U|Z)-(1-\lambda)I(X;U|Z),\quad \lambda\in[0,1]. \tag{4}$$

We call $f(\lambda,P_{U|X})=\lambda I(Y;U|Z)-(1-\lambda)I(X;U|Z)$ in order to simplify the notation. Our approach is to obtain an algorithm which is able to find for every $\lambda\in[0,1]$ the optimal pmf $P^{*,\lambda}_{U|X}$ that achieves the maximum in (4) and evaluating the corresponding mutual informations. The function $f(\lambda,P_{U|X})$ can be written as:

$$f(\lambda,P_{U|X})=(2\lambda-1)I(X;U|Z)-\lambda I(X;U|Y,Z). \tag{5}$$

It is appropriate to define the algorithm in two different ways depending on the value of λ. This is similar to the approach in [11]. If $\lambda\in[0,0.5]$, both terms of (5) are non-positive, and the solution is trivial ($V_\lambda=0$). Clearly, this is achieved for all pmf that satisfies $P_{U|X}=P_U$ (U independent of X). The more interesting case is when $\lambda\in(0.5,1]$. In that case the proposed algorithm is an iterative one. Let an initial condition $P^{(0)}_{U|X}$, the algorithm iterate until convergence between:

$$Q^{(n+1)}_{U|YZ}=\sum_x P^{(n)}_{U|X}P_{X|YZ},\ \ Q^{(n+1)}_{X|ZU}=\frac{P^{(n)}_{U|X}P_{X|Z}}{\sum_{x'}P^{(n)}_{U|X'}P_{X'|Z}} \tag{6}$$

$$\begin{aligned}P^{(n+1)}_{U|X}=k_x\cdot\exp\Bigg\{\frac{2\lambda-1}{\lambda}\sum_z P_{Z|X}\log(Q^{(n+1)}_{X|ZU})\\ +\sum_{y,z}P_{Y|XZ}P_{Z|X}\log(Q^{(n+1)}_{U|YZ})\Bigg\}\end{aligned} \tag{7}$$

where k_x are constant such that $\sum_u P^{(n+1)}_{U|X}=1\quad\forall\, x\in\mathcal{X}$.

3. ALGORITHM ANALYSIS

In this section we will provide an analysis of (6) and (7). We define the function $F(\lambda,P_{U|X},Q_{U|YZ},Q_{X|ZU})$ as:

$$\begin{aligned}&F(\lambda,P_{U|X},Q_{U|YZ},Q_{X|ZU})\\ &\quad=(2\lambda-1)\sum_{x,z,u}P_{U|X}P_{XZ}\log\left(\frac{Q_{X|ZU}}{P_{X|Z}}\right)\\ &\quad\quad-\lambda\sum_{x,y,z,u}P_{U|X}P_{XYZ}\log\left(\frac{P_{U|X}}{Q_{U|YZ}}\right),\end{aligned} \tag{8}$$

where $Q_{U|YZ},Q_{X|ZU}$ are arbitrary pmfs. This new function has some properties.

Theorem 1. *Consider any $P_{U|X}$ and let $\lambda\in(0.5,1]$. The following properties are true:*

1. *$f(\lambda, P_{U|X}) \geq F(\lambda, P_{U|X}, Q_{U|YZ}, Q_{X|ZU})$, and equality is achieved iff $Q_{U|YZ} = P_{U|YZ}$ $\forall\ (y,z) \in \mathcal{Y} \times \mathcal{Z}$ and $Q_{X|ZU} = P_{X|ZU}$ $\forall (z,u) \in \mathcal{Z} \times \mathcal{U}$.*

2. *The value V_λ satisfies*

$$V_\lambda = \max_{P_{U|X}} \max_{Q_{U|YZ}, Q_{Y|ZU}} F(\lambda, P_{U|X}, Q_{U|YZ}, Q_{X|ZU}).$$

3. *For any $Q_{U|YZ}, Q_{X|ZU}$ and $\lambda \in (0.5, 1]$, F is concave in $P_{U|X}$ and it achieves its maximum iff*

$$P_{U|X} = k_x \cdot \exp\left\{ \frac{2\lambda - 1}{\lambda} \sum_z P_{Z|X} \log(Q_{X|ZU}) + \sum_{y,z} P_{Y|XZ} P_{Z|X} \log(Q_{U|YZ}) \right\}, \quad (9)$$

where k_x are constants such that $\sum_u P_{U|X} = 1$ $\forall x$.

The proof and other algorithm properties can be reader in [15]. We see that function $F(\lambda, P_{U|X}, Q_{U|YZ}, Q_{X|ZU})$ provides an achievable and easy way to optimize a lower bound to the objective function $f(\lambda, P_{U|X})$ for each $P_{U|X}$. Interestingly enough, $P_{U|X} \mapsto F(\lambda, P_{U|X}, Q_{U|YZ}, Q_{X|ZU})$ is concave for each $(Q_{U|YZ}, Q_{X|ZU})$, guaranteeing that any local optimum is also a global one. These facts lead naturally to the iterative process in (6) (7) to perform the double maximization which should result in V_λ. For a given $\lambda \in (0.5, 1]$ and starting from an initial condition $P^{(0)}_{U|X}$, and according to 2) in Theo. 1 we find $Q^{(1)}_{U|YZ}, Q^{(1)}_{X|ZU}$ such that the maximum of $F(\lambda, P^{(0)}_{U|X}, Q_{U|YZ}, Q_{X|ZU})$ for fixed $P^{(0)}_{U|X}$ is achieved. Next, from 3) in the previous theorem, we find $P^{(1)}_{U|X}$ as the argument that maximize $F(\lambda, P_{U|X}, Q^{(1)}_{U|YZ}, Q^{(1)}_{X|ZU})$. It is easy to show that the sequence of values $F(\lambda, P^{(n)}_{U|X}, Q^{(n)}_{U|YZ}, Q^{(n)}_{X|ZU})$ provided by the iterative process is monotone non-decreasing. This clearly guarantees that the process is convergent. A global convergence analysis can be seen in [15].

4. NUMERICAL EXAMPLES

In this section, we will exemplify the use of the proposed algorithm for different problems and applications.

4.1. Compression-based regularization learning

In previous sections, we have shown that the problem of maximizing the relevance $I(U;Y|Z)$ subject to a mutual-information constraint $I(U;X|Z) \leq R$ is equivalent to the maximization problem of $f(\lambda, P_{U|X})$. We now show that this constraint can act as a regularization when applied to situations where the joint statistics controlling the observations P_{XYZ} is not known but it is estimated via training samples. Indeed, Shamir *et al.* [17] have already showed evidence that this term can help to prevent "overfitting". It should mentioned that these analysis were performed for the classical IB setting in which side information is not present. In this section, we provide numerical evidence that the desired regularization effects also hold in our MTL setup. Consider a multi-task supervised classification problem and define the average risk:

$$\text{Risk}\big(P_{U|X}, P_{\hat{Y}|UZ}\big) = \mathbb{E}_{P_{XYZ}P_{U|X}}[-\log P_{\hat{Y}|UZ}(Y|UZ)], \quad (10)$$

with respect to an encoder $P_{U|X}$ and a decoder $P_{\hat{Y}|UZ}$. Finding the optimal encoder in (10) requires knowledge of the underlying distribution P_{XYZ}. From a practical perspective, as the input to the proposed algorithm, we will use the data sampling distribution $\hat{P}_{XYZ}$ based on n training labeled examples. By introducing the rate constraint (or penalization), the optimization problem is reduced to optimizing $L(R, \hat{P}_{XYZ})$ in (3) from which the resulting encoder $\hat{P}^{*,\lambda}_{U|X}$ is derived while the decoder $\hat{P}^{*,\lambda}_{\hat{Y}|UZ}$ follows from expression (1)[1]. Our measure of merit will be the *Excess-risk*:

$$\text{Excess-risk} = \text{Risk}\big(\hat{P}^{*,\lambda}_{U|X}, \hat{P}^{*,\lambda}_{\hat{Y}|UZ}\big) - H(Y|XZ). \quad (11)$$

The source parameters for the experiment show in Fig. 2 can be read in [15]. In Fig. 2, we plot the excess risk curve as a function of the rate constraint for different size of training samples. With dash lines we denoted the rate R such that the excess risk achieves its minimum. When the number of training samples increases the optimal rate R approaches its maximum possible value: $H(X|Z)$ (dashed in black). Notice that for every curve there exists a different limiting rate $\hat{H}(X|Z)$, such that for each $R \geq \hat{H}(X|Z)$, the excess-risk remains constant with value $\hat{I}(X;Y|Z)$. In addition, for every size of training samples, there is an optimal rate value which provides the lowest value for the excess-risk in (11). In a sense, this is indicating that the rate R can be interpreted as an effective regularization term and hence, it can provide robustness for learning in practical situations in which the true

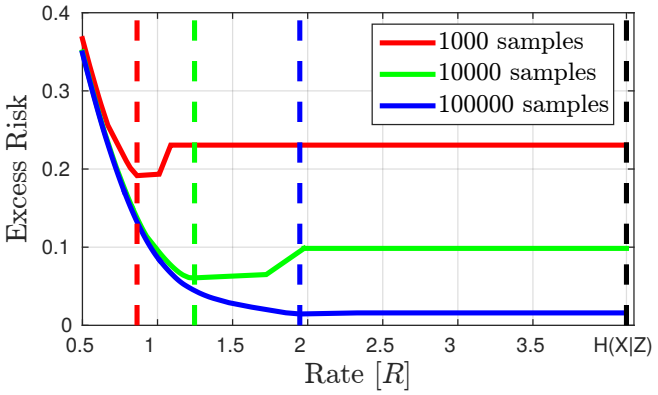

Fig. 2. Excess risk (11) as a function of information rate.

[1] Note that when the decoder is chosen as in (1), $\text{Risk}\big(P_{U|X}, P_{\hat{Y}|UZ}\big) = H(Y|UZ) \geq H(Y|XZ)$.

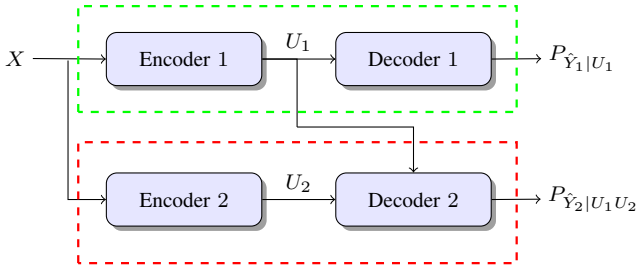

Fig. 3. Hierarchical text categorization scheme.

input distribution is not known and the empirical data distribution is used. It is worth to mention that when more data is available then the optimal value of the regularizing rate becomes less critical.

4.2. Hierarchical text categorization

High dimensionality of text can become a severe deterrent to the task of text classification. This issue can be alleviated by intelligently grouping different classes in disjoint subcategories. In this way a first classification problem can be set over the generated sub-categories and the information extracted can be used in a second classification problem to discriminate better between the original classes. This is the case in hierarchical text classification [8], [18].

In order to solve this problem, we propose the scheme of Fig. 3. Consider a document d consisting of different words X. We want to estimate the class Y_2 to which the document belongs by using information related to a sub-category Y_1 (typically related to the text topic) to which the same document also belongs. We consider the following setting. A pair of encoder 1-decoder 1 tries to estimate the document subcategory $\hat{Y}_1$ using our algorithm without side information (Z is a constant) and with input P_{XY_1}. This is clearly a standard classification problem where U_1 is the feature that encoder 1 extracted from X. The encoder 2-decoder 2 pair tries to generate the final classification in $\hat{Y}_2$ using our algorithm with input $P_{XY_1U_1}$. In this case, U_1 can be considered as side information available at the decoder 2. We see that this problem can be interpreted as MTL problem where the different classification problems (tasks) to be solved by the decoder 2 are induced by the features extracted by encoder 1.

Assume a training set consisting of documents belonging to $|\mathcal{Y}_2|$ classes. The distribution $P_{Y_1|Y_2}$ is known because the sub-category Y_1 is a function of the more refined class Y_2. The class priors P_{Y_2} are replaced by the empirical distribution and the words distribution conditional to the class $P_{X|Y_2}$, is estimated using Laplace rule of succession [19]. Once pmfs $P_{U_1|X}$ and $P_{U_2|X}$ are calculated using the proposed algorithm, we estimate the class of the document $\hat{y}_2(d)$ using the maximum a posteriori probability:

$$\hat{y}_2(d) = \arg\max_{y\in\mathcal{Y}_2} P_{Y_2|D}(y|d) \tag{12}$$

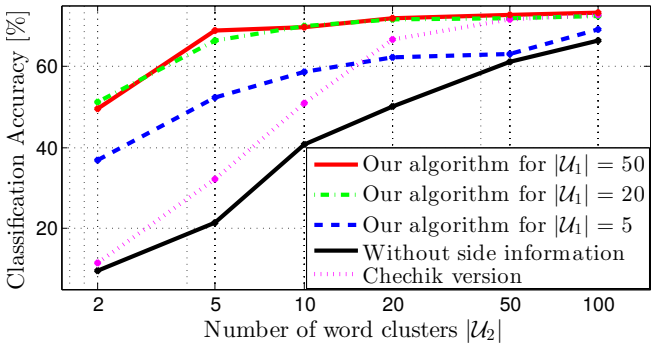

Fig. 4. Classification Accuracy in the hierarchical text categorization problem.

$$= \arg\max_{y\in\mathcal{Y}_2} \left[\log\left(P_{Y_2}\right) + \sum_{u_1,u_2} n(u_1,u_2,d)\log P_{U_1U_2|Y_2}\right],$$

where $n(u_1,u_2,d)$ is the number of jointly occurrences of clusters (u_1,u_2) in the document d computed with:

$$u_j(x) = \arg\max_{u} P_{U_j|X}(u|x), \quad j\in\{1,2\}. \tag{13}$$

We test the above proposed classification procedure on the 20 Newsgroups (20Ng) data set [20]. The 20 Newsgroups correspond to 6 topics. In this case, the sub-category Y_1 is the topic and the refined classification Y_2 is the Newsgroup. In Fig. 4, our algorithm performance ($\lambda = 0.99$) versus $|\mathcal{U}_2|$ is compared with the algorithm without side information (which is a single-task setup) and the one proposed in [8]. It is interesting to mention that the single-task setting and the one in [8] can be covered using our algorithm (details can be read in [15]). Our setting and the one in [8] show an improvement with respect to the single task setup (without side information). This suggests that exploiting the common features in MTL may be advantageous. Further our setting uses the additional information in a structured manner to show an improvement with respect to the other proposals.

5. SUMMARY AND DISCUSSION

From information-theoretic methods, we have investigated the MTL framework in which an encoder builds a common representation for several related tasks. We derived an iterative learning algorithm that uses compression as a natural safeguard against overfitting. Empirical evidence shows that there exists an optimal information rate minimizing the *excess risk* according to the amount of available training data. It is observed that these optimal rates increase with the size of the training set. An application to hierarchical text categorization was also investigated.

At present, several open questions remain regarding the statistical regularization properties of building compact data representations. Applications of the proposed algorithm to other MTL problems also deserve some efforts.

6. REFERENCES

[1] Claude Elwood Shannon, "Coding theorems for a discrete source with a fidelity criterion," in *Claude Elwood Shannon: collected papers*, N. J. A. Sloane and A. D. Wyner, Eds., pp. 325–350. IEEE Press, 1993.

[2] Kenneth Rose, "Deterministic annealing for clustering, compression, classification, regression, and related optimization problems," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2210–2239, Nov 1998.

[3] Naftali Tishby, Fernando C. Pereira, and William Bialek, "The information bottleneck method," in *Proceedings of the 37-th Annual Allerton Conference on Communication, Control and Computing*, 1999, pp. 368–377.

[4] Noam Slonim and Naftali Tishby, "The power of word clusters for text classification," in *23rd European Colloquium on Information Retrieval Research (ECIR)*, 2001, pp. 1–12.

[5] Ron M. Hecht, Elad Noor, and Naftali Tishby, "Speaker recognition by gaussian information bottleneck.," in *INTERSPEECH*. 2009, pp. 1567–1570, ISCA.

[6] Suguru Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," *IEEE Trans. Inform. Theory*, vol. 18, no. 1, pp. 14–20, 1972.

[7] Richard Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Trans. Inform. Theory*, vol. 18, no. 4, pp. 460–473, 1972.

[8] Gal Chechik and Naftali Tishby, "Extracting relevant structures with side information," in *Advances in Neural Information Processing Systems 15 [Neural Information Processing Systems, NIPS 2002, December 9-14, 2002, Vancouver, British Columbia, Canada]*, 2002, pp. 857–864.

[9] Gowtham Kumar and Andrew Thangaraj, "Computation of secrecy capacity for more-capable channel pairs," in *Information Theory (ISIT), 2008 IEEE International Symposium on*, Toronto, Canada, July 2008.

[10] Kensuke Yasui, Tota Suko, and Toshiyasu Matsushima, "An algorithm for computing the secrecy capacity of broadcast channels with confidential messages," in *Information Theory (ISIT), 2007 IEEE International Symposium on*, Nice, France, June 2007.

[11] Kensuke Yasui and Toshiyasu Matsushima, "Toward computing the capacity region of degraded broadcast channel," in *Information Theory (ISIT), 2010 IEEE International Symposium on*, Texas, U.S.A, June 2010.

[12] Rich Caruana, "Multitask learning," *Machine Learning*, vol. 28, no. 1, pp. 41–75, Jul 1997.

[13] Jonathan Baxter, "A model of inductive bias learning," *Journal of Artificial Intelligence Research*, vol. 12, pp. 149–198, 2000.

[14] Pei Yang, Qi Tan, Hao Xu, and Yehua Ding, "An information-theoretic approach for multi-task learning," in *in International Conference on Advanced Data Mining and Applications (ADMA 2009)*, 2009, pp. 386–396.

[15] Matias Vera, Leonardo Rey Vega, and Pablo Piantanida, "Compression-based regularization with an application to multi-task learning," *ArXiv e-prints, https://arxiv.org/abs/1711.07099*, Nov. 2017.

[16] Matias Vera, Leonardo Rey Vega, and Pablo Piantanida, "The two-way cooperative information bottleneck," in *IEEE International Symp. on Information Theory, ISIT 2015*, 2015, pp. 2131–2135.

[17] Ohad Shamir, Sivan Sabato, and Naftali Tishby, "Learning and generalization with the information bottleneck," *Theor. Comput. Sci.*, vol. 411, no. 29-30, pp. 2696–2711, June 2010.

[18] Alexei Vinokourov and Mark Girolami, "A Probabilistic Framework for the Hierarchic Organisation and Classification of Document Collections," *Journal of Intelligent Information Systems*, vol. 18, no. 2–3, pp. 153–172, Mar. 2002.

[19] Christopher M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[20] Ken Lang, "Newsweeder: Learning to filter netnews," in *in Proceedings of the 12th International Machine Learning Conference (ML95)*, 1995.