

---

# USING THE NAIVE BAYES AS A DISCRIMINATIVE CLASSIFIER

---

**Elie Azeraf\***  
 Watson Department  
 IBM GSB France  
 elie.azeraf@ibm.com

**Emmanuel Monfrini**  
 SAMOVAR, Telecom SudParis  
 Institut Polytechnique de Paris

**Wojciech Pieczynski**  
 SAMOVAR, Telecom SudParis  
 Institut Polytechnique de Paris

## ABSTRACT

For classification tasks, probabilistic models can be categorized into two disjoint classes: generative or discriminative. It depends on the posterior probability computation of the label  $x$  given the observation  $y$ ,  $p(x|y)$ . On the one hand, generative classifiers, like the Naive Bayes or the Hidden Markov Model (HMM), need the computation of the joint probability  $p(x,y)$ , before using the Bayes rule to compute  $p(x|y)$ . On the other hand, discriminative classifiers compute  $p(x|y)$  directly, regardless of the observations' law. They are intensively used nowadays, with models as Logistic Regression, Conditional Random Fields (CRF), and Artificial Neural Networks. However, the recent Entropic Forward-Backward algorithm shows that the HMM, considered as a generative model, can also match the discriminative one's definition. This example leads to question if it is the case for other generative models. In this paper, we show that the Naive Bayes classifier can also match the discriminative classifier definition, so it can be used in either a generative or a discriminative way. Moreover, this observation also discusses the notion of Generative-Discriminative pairs, linking, for example, Naive Bayes and Logistic Regression, or HMM and CRF. Related to this point, we show that the Logistic Regression can be viewed as a particular case of the Naive Bayes used in a discriminative way.

**Keywords** Probabilistic Graphical Models · Generative Classifier · Discriminative Classifier · Logistic Regression · Naive Bayes; Generative-Discriminative pair

## 1 Introduction

Given an input data  $y$ , also called observation, and a discrete finite output space  $\Lambda_X = \{\lambda_1, \dots, \lambda_N\}$ , a classification task consists in computing the different label probabilities  $p(x = \lambda_i|y)$ . They can be used in many areas as Natural Language Processing with Text Classification [1] or Sentiment Analysis [2], or Computer Vision with Image Classification [3] or Object Detection [4], among many others. These model classifiers are usually categorized into two disjoint classes: generative or discriminative. From [5], a generative classifier learns the joint probability  $p(x, y)$ , then uses the Bayes rule to compute the posterior probability  $p(x|y)$ :

$$\forall \lambda_i \in \Lambda_X, p(x = \lambda_i|y) = \frac{p(x = \lambda_i, y)}{\sum_{j \in \Lambda_X} p(x = \lambda_j, y)} \quad (1)$$

To train a generative model in a supervised context, one has to maximize the joint likelihood  $p(x, y)$  of the training data. For example, in the case where  $y$  is a discrete variable, it consists in counting the different patterns' occurrence. Among the most popular generative models, we can cite the Naive Bayes [6, 7, 8, 9], the Hidden Markov Model (HMM) [10, 11, 12], or the Gaussian Mixture Model [13, 14].

On the other hand, a discriminative classifier computes  $p(x|y)$  directly, which means that  $p(x,y)$  does not have to be known. We can cite the Logistic Regression [15, 16, 17, 18, 19], the Maximum Entropy Markov Model (MEMM) [20],

---

\*Elie Azeraf is also a member of SAMOVAR, Telecom SudParis, Institut Polytechnique de Paris

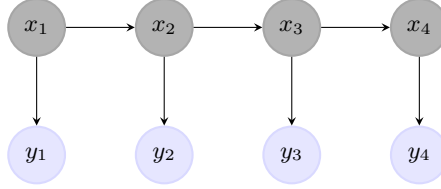


Figure 1: Probabilistic oriented graph of the HMM

the Conditional Random Fields (CRF) [21, 22], or even the Artificial Neural Networks (ANN) [23, 24], among the most popular ones. They are usually trained by fitting the posterior distribution  $p(x|y)$ , or minimizing a loss function, thanks to optimization algorithms like the gradient descent [25].

Many papers use these definitions [26, 27, 28, 29, 30, 22, 31, 32] and compare both approaches, with a general preference for the second category where the task does not have unlabeled data. Indeed, generative classifiers' main criticism concerns their learning strategy, imposing to learn the joint probability, and therefore the observation's distribution. This learning method constrains the observation's features during classification tasks. Indeed, it is considered impossible to handle arbitrary features with a generative model [33, 20, 22], except with a restrictive independence condition. The next section will illustrate this point.

We focus on the HMM, represented in figure 1, with an observed process  $y_{1:T} = (y_1, \dots, y_T)$ , and a hidden one  $x_{1:T} = (x_1, \dots, x_T)$ . This model is considered as a generative probabilistic one. Indeed, if one wants to compute  $p(x_t = \lambda_i | y_{1:T})$ , he can apply the Forward-Backward algorithm [11, 34] described in the appendices. In agreement with the generative model definition, this algorithm firstly computes the joint probability  $p(x_t = \lambda_i, y_{1:T})$  before computing the posterior one. Moreover, all the parameters are learned with maximum likelihood. However, a recent work about HMM presents the Entropic Forward-Backward algorithm [35], also described in the appendices. It allows computing the posterior distribution  $p(x_t = \lambda_i | y_{1:T})$  directly, with using neither the joint distribution, nor the observations' one. It also allows training the HMM with optimization algorithms as the gradient descent. Therefore, in this case, the HMM matches the discriminative classifier's definition.

This HMM example leads to question this categorization of probabilistic classifiers, which is the aim of the paper. We present two contributions:

1. We show that the Naive Bayes, a popular generative model, can also match the definition of a discriminative one. It is another example showing that the traditional definitions of both discriminative and generative models do not necessarily lead to disjoint categories.
2. We discuss the notion of Generative-Discriminative pairs [5, 22], linking a generative model with its discriminative counterpart, as for example, the Naive Bayes - Logistic Regression pair, or the HMM - CRF one. Considering the first example, we show that the Logistic Regression is a particular case of the Naive Bayes used in a discriminative way.

This paper is organized as follows. In the next section, we recall the Naive Bayes model's law with its oriented probabilistic graph, and we present the generative way, usually used, to compute the label probabilities. Then, we show how to compute these probabilities in a discriminative way, which does not use the observation's law. In the third section, after some recalls about the Logistic Regression model, we show that this latter can be viewed as a particular case of the Naive Bayes used in a discriminative way. Conclusion and perspectives lie at the end of the paper.

## 2 The Naive Bayes classifier as a discriminative model

### 2.1 Naive Bayes classifier

The Naive Bayes is a probabilistic graphical model [36] considering the observations  $y_{1:T} = (y_1, \dots, y_T)$ , with  $y_t$  taking its value in  $\Omega_{Y_t}$ , and a hidden variable  $x$  taking its value in  $\Lambda_X$ . It models the joint probability of  $(x, y_{1:T})$  with the following law:

$$p(x, y_{1:T}) = p(x)p(y_1|x)p(y_2|x)\dots p(y_{T-1}|x)p(y_T|x) = p(x) \prod_{t=1}^T p(y_t|x) \quad (2)$$

The oriented probabilistic graph is given in figure 2.

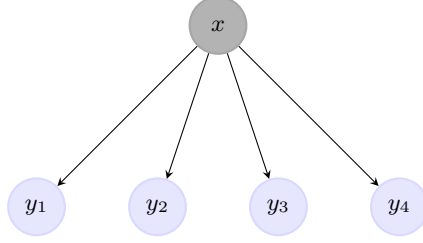


Figure 2: Probabilistic graphical representation of the Naive Bayes

In the generative way, as usually used, the Naive Bayes classifier computes, for each  $\lambda_i \in \Lambda_X$ ,  $p(x = \lambda_i | y_{1:T})$  as follows:

$$p(x = \lambda_i | y_{1:T}) = \frac{p(x = \lambda_i, y_{1:T})}{\sum_{\lambda_j \in \Lambda_X} p(x = \lambda_j, y_{1:T})} \quad (3)$$

$$= \frac{p(x = \lambda_i) \prod_{t=1}^T p(y_t | x = \lambda_i)}{\sum_{\lambda_j \in \Lambda_X} p(x = \lambda_j) \prod_{t=1}^T p(y_t | x = \lambda_j)} \quad (4)$$

$$= \frac{\pi(i) \prod_{t=1}^T b_i^{(t)}(y_t)}{\sum_{\lambda_j \in \Lambda_X} \pi(j) \prod_{t=1}^T b_j^{(t)}(y_t)} \quad (5)$$

using the notations:

- $\pi(i) = p(x = \lambda_i)$ ;
- $b_i^{(t)}(y) = p(y_t | x_t = \lambda_i)$ .

With (5), the Naive Bayes effectively matches the generative classifier definition, as it first computes the joint probability  $p(x = \lambda_i, y_{1:T})$ , and then the posterior  $p(x = \lambda_i | y_{1:T})$ .

Used in this way, it is facing difficulties to consider arbitrary features. Indeed, let us consider, for example,  $\Omega_{Y_t} = \Omega_Y^2$ , for each t, with  $\Omega_Y = \{\omega_1, \dots, \omega_M\}$ . Thus, each observation has two features:  $y_t = (\omega_i, \omega_j)$ . With this restoration method, the different parameters are learned by maximum likelihood estimation, which consists in counting the different patterns:

$$\pi(i) = \frac{f(i)}{L}, \quad b_i^{(t)}(y_t) = \frac{f_i^{(t)}(j, k)}{f(i)},$$

with L the number of training samples,  $f(i)$  the number of times  $x = \lambda_i$  in this training sample, and  $f_i^{(t)}(j, k)$  the number of times  $y_t = (\omega_j, \omega_k)$  when  $x = \lambda_i$ . Therefore, an observation  $y_t$  has its value  $b_i^{(t)}(y_t)$  different from 0 if and only if there is an observation in the training set having the same features at position t. If this estimation method is possible with a small number of features, it quickly becomes intractable when the number of features increases, making mandatory to suppose them independent. It is especially the case for Natural Language Processing tasks, where features can be suffixes of any length, prefixes, some word's characteristics, or large numerical vector issued from an embedding method [37].

## 2.2 Computing directly $p(x = \lambda_i | y_{1:T})$ with the Naive Bayes classifier

In this section, we show that one can use the Naive Bayes classifier by computing  $p(x = \lambda_i | y_{1:T})$  directly, without previous computation of  $p(x = \lambda_i, y_{1:T})$ . To simplify notations, we set, for each  $\lambda_i \in \Lambda_X$ :

$$L_{y_t}^{(t)}(i) = p(x = \lambda_i | y_t).$$

We can state the following result:

**Proposition 1.** Let  $p(x = \lambda_i | y_{1:T})$  be a Naive Bayes distribution. Then for each  $\lambda_i \in \Lambda_X$ :

$$p(x = \lambda_i | y_{1:T}) = \frac{\pi(i)^{1-T} \prod_{t=1}^T L_{y_t}^{(t)}(i)}{\sum_{j=1}^N \pi(j)^{1-T} \prod_{t=1}^T L_{y_t}^{(t)}(j)} \quad (6)$$

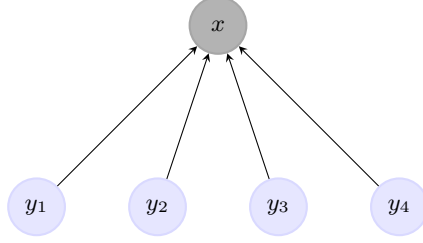


Figure 3: Probabilistic graphical representation of the Logistic Regression

**Proof.** For each  $\lambda_i \in \Lambda_X$ , we set:

$$\delta(i) = \pi(i)^{1-T} \prod_{t=1}^T L_{y_t}^{(t)}(i) \quad (7)$$

This function is linked with the joint probability by:

$$p(x = \lambda_i, y_{1:T}) = p(y_1) p(y_2) \dots p(y_{T-1}) p(y_T) \delta(i) \quad (8)$$

Indeed,

$$p(x = \lambda_i, y_{1:T}) = \pi(i) \prod_{t=1}^T b_i^{(t)}(y_t) = \pi(i) \prod_{t=1}^T \frac{p(y_t, x = \lambda_i)}{p(x = \lambda_i)} = \pi(i)^{1-T} \prod_{t=1}^T p(y_t) L_{y_t}^{(t)}(i) = \delta(i) \prod_{t=1}^T p(y_t)$$

Therefore

$$p(x = \lambda_i | y_{1:T}) = \frac{p(x = \lambda_i, y_{1:T})}{\sum_{j=1}^N p(x = \lambda_j, y_{1:T})} = \frac{\delta(i)}{\sum_{j=1}^N \delta(j)}, \quad (9)$$

which ends the proof.

Finally, given (4), the Naive Bayes model also matches the definition of a discriminative classifier: it allows computing  $p(x | y_{1:T})$  directly and does not use neither the observations' law nor the joint one. Therefore, through this example, we see that in some situations, “discriminative” and “generative” definitions do not describe the deep nature of a model, but rather the way it is used. When differentiable functions model  $L_{y_t}^{(t)}$ , one can train the Naive Bayes used in a discriminative way with (4) using gradient descent algorithms. In this case, the training process consists in computing the gradient of a given loss function by applying the backpropagation [38, 39] algorithm, and updating the model's parameters. In this way, the Naive Bayes can be applied for any classification tasks and consider arbitrary observation features. For example, a homogeneous Naive Bayes used in a discriminative way, modelling  $L_{y_t}^{(t)}$  with the same function independent of  $t$ , can be applied to classify text of various lengths. First of all, one has to use an embedding method like Flair [40], FastText [41], or BERT [42], to transform word into numerical vector. Then, one has to select a function to model  $L_{y_t}^{(t)}$ , for example, a feedforward neural network.

### 3 The Logistic Regression: a particular case of the Naive Bayes used in a discriminative way

#### 3.1 Logistic Regression classifier

Let us consider multinomial Logistic Regression, which is the more general one. For each  $t, \Omega_{Y_t} = \mathbb{R}$ , Logistic Regression computes  $p(x = \lambda_i | y_{1:T})$  with

$$p(x = \lambda_i | y_{1:T}) = \frac{\exp(W_i \cdot y_{1:T} + b_i)}{\sum_{j=1}^N \exp(W_j \cdot y_{1:T} + b_j)} \quad (10)$$

where for each  $i \in \{1, \dots, N\}$ ,  $W_i \in \mathbb{R}^T$ ,  $b_i \in \mathbb{R}$ , and  $\cdot$  denotes the matrix product.

Figure 3 represents the oriented graph of this classifier.

The Logistic Regression is a probabilistic graphical model used in numerous applications. One can train it with a gradient descent algorithm. It is considered as a discriminative model as it computes directly  $p(x | y_{1:T})$ . With this model, we notice that each probability computed is strictly greater than 0.

### 3.2 Logistic Regression and Naive Bayes

Logistic Regression is frequently linked with the Naive Bayes with the notion of Generative – Discriminative pairs [5, 22], assuming it is the discriminative counterpart of the Naive Bayes. However, we are going to show that the Logistic Regression is a particular case of the Naive Bayes.

We can state:

**Proposition 2.** Let us consider a Naive Bayes classifier used in a discriminative way with (4). If, for each  $i \in \{1, \dots, N\}$ ,  $t \in 1, \dots, T$ :

$$L_{y_t}^{(t)}(i) = \frac{\exp(a_i^{(t)} y_t + c_i^{(t)})}{\sum_{k=1}^N \exp(a_k^{(t)} y_t + c_k^{(t)})} \quad (11)$$

with  $a_i^{(t)} \in \mathbb{R}$  and  $c_i^{(t)} \in \mathbb{R}$ , then this Naive Bayes classifier is a Logistic Regression.

**Proof .** We start from (4), for each  $\lambda_i \in \Lambda_X$ :

$$\begin{aligned} p(x = \lambda_i | y_{1:T}) &= \frac{\pi(i)^{1-T} \prod_{t=1}^T L_{y_t}^{(t)}(i)}{\sum_{j=1}^N \pi(j)^{1-T} \prod_{t=1}^T L_{y_t}^{(t)}(j)} \\ &= \frac{\pi(i)^{1-T} \prod_{t=1}^T \frac{\exp(a_i^{(t)} y_t + c_i^{(t)})}{\sum_{k=1}^N \exp(a_k^{(t)} y_t + c_k^{(t)})}}{\sum_{j=1}^N \pi(j)^{1-T} \prod_{t=1}^T \frac{\exp(a_j^{(t)} y_t + c_j^{(t)})}{\sum_{k=1}^N \exp(a_k^{(t)} y_t + c_k^{(t)})}} \\ &= \frac{\pi(i)^{1-T} \prod_{t=1}^T \exp(a_i^{(t)} y_t + c_i^{(t)})}{\sum_{j=1}^N \pi(j)^{1-T} \prod_{t=1}^T \exp(a_j^{(t)} y_t + c_j^{(t)})} \\ &= \frac{\exp\left((1-T) \log(\pi(i)) + \sum_{t=1}^T a_i^{(t)} y_t + c_i^{(t)}\right)}{\sum_{j=1}^N \exp\left((1-T) \log(\pi(j)) + \sum_{t=1}^T a_j^{(t)} y_t + c_j^{(t)}\right)} \end{aligned}$$

We set:

- $b_i = (1-T) \log(\pi(i)) + \sum_{t=1}^T c_i^{(t)}$ ;
- $W_i = [a_i^{(1)}, a_i^{(2)}, \dots, a_i^{(T)}]$ .

Therefore, we can effectively verify that  $b_i \in \mathbb{R}$  and  $W_i \in \mathbb{R}^T$ .

Then:

$$p(x = \lambda_i | y_{1:T}) = \frac{\exp(W_i \cdot y_{1:T} + b_i)}{\sum_{j=1}^N \exp(W_j \cdot y_{1:T} + b_j)} \quad (12)$$

which is the equation (8) defining a Logistic Regression and ends the proof. Proposition 2 shows that one can view the Logistic Regression as a particular case of the Naive Bayes used in a discriminative way.

## 4 Conclusion

This paper presents how to use the Naive Bayes as a discriminative model. A practical consequence is that it can be used in both generative and discriminative ways; the latter allowing arbitrary feature consideration. Moreover, we show that the Logistic Regression, usually presented as the discriminative counterpart of the Naive Bayes, can be seen as a particular case of the latter used in a discriminative way. A general conclusion is that the usual definitions of generative and discriminative models do not lead to disjoint families; and there are models that simultaneously satisfy both of them. An interesting perspective would consist in extending ideas of the paper to other popular generative models and examining whether they can also be used in a discriminative way.

## References

- [1] Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald Brown. Text classification algorithms: A survey. *Information*, 10(4):150, 2019.
- [2] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.
- [3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [4] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.
- [5] Andrew Y Ng and Michael I Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in neural information processing systems*, pages 841–848, 2002.
- [6] David D Lewis. Naive (bayes) at forty: The independence assumption in information retrieval. In *European conference on machine learning*, pages 4–15. Springer, 1998.
- [7] Andrew McCallum, Kamal Nigam, et al. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48. Citeseer, 1998.
- [8] Irina Rish et al. An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46, 2001.
- [9] Geoffrey I. Webb. *Naive Bayes*, pages 713–714. Springer US, Boston, MA, 2010.
- [10] Olivier Cappé, Eric Moulines, and Tobias Rydén. Inference in hidden markov models. 2009.
- [11] Lawrence Rabiner and B Juang. An introduction to hidden markov models. *iee assp magazine*, 3(1):4–16, 1986.
- [12] Ruslan Leont’evich Stratonovich. Conditional markov processes. In *Non-linear transformations of stochastic processes*, pages 427–453. Elsevier, 1965.
- [13] Carl Edward Rasmussen. The infinite gaussian mixture model. In *Advances in neural information processing systems*, pages 554–560, 2000.
- [14] Douglas Reynolds. *Gaussian Mixture Models*, pages 659–663. Springer US, Boston, MA, 2009.
- [15] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.
- [16] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.
- [17] Scott Menard. *Applied logistic regression analysis*, volume 106. Sage, 2002.
- [18] Chao-Ying Joanne Peng, Kuk Lida Lee, and Gary M Ingersoll. An introduction to logistic regression analysis and reporting. *The journal of educational research*, 96(1):3–14, 2002.
- [19] Raymond E Wright. Logistic regression. 1995.
- [20] Andrew McCallum, Dayne Freitag, and Fernando CN Pereira. Maximum entropy markov models for information extraction and segmentation.
- [21] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.
- [22] Charles Sutton and Andrew McCallum. An introduction to conditional random fields for relational learning. *Introduction to statistical relational learning*, 2:93–128.
- [23] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.
- [24] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.
- [25] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.
- [26] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.
- [27] Guillaume Bouchard and Bill Triggs. The tradeoff between generative and discriminative classifiers. 2004.

- [28] Julia A Lasserre, Christopher M Bishop, and Thomas P Minka. Principled hybrids of generative and discriminative models. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 87–94. IEEE, 2006.
- [29] Tom Minka. Discriminative models, not discriminative training. Technical report, Technical Report MSR-TR-2005-144, Microsoft Research, 2005.
- [30] Wolfgang Roth, Robert Peharz, Sebastian Tschiatschek, and Franz Pernkopf. Hybrid generative-discriminative training of gaussian mixture models. *Pattern recognition letters*, 112:131–137, 2018.
- [31] Ilkay Ulusoy and Christopher M Bishop. Generative versus discriminative methods for object recognition. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 258–265. IEEE, 2005.
- [32] Oksana Yakhnenko, Adrian Silvescu, and Vasant Honavar. Discriminatively trained markov model for sequence classification. In *Fifth IEEE International Conference on Data Mining (ICDM'05)*, pages 8–pp. IEEE, 2005.
- [33] Dan Jurafsky. *Speech & language processing*. Pearson Education India, 2000.
- [34] Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.
- [35] Elie Azeraf, Emmanuel Monfrini, Emmanuel Vignon, and Wojciech Pieczynski. Hidden markov chains, entropic forward-backward, and part-of-speech tagging. *arXiv preprint arXiv:2005.10629*, 2020.
- [36] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.
- [37] Felipe Almeida and Geraldo Xexéo. Word embeddings: A survey. *arXiv preprint arXiv:1901.09069*, 2019.
- [38] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.
- [39] Yann LeCun, Bernhard E Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne E Hubbard, and Lawrence D Jackel. Handwritten digit recognition with a back-propagation network. In *Advances in neural information processing systems*, pages 396–404, 1990.
- [40] Alan Akbik, Duncan Blythe, and Roland Vollgraf. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, 2018.
- [41] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.
- [42] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

## Appendix

In this appendix, we present both HMM’s algorithms to restore, for each  $\lambda_i \in \Lambda_X$ ,  $t \in \{1, \dots, T\}$ ,  $y_{1:T} \in (\Omega_Y)^T$ ,  $p(x_t = \lambda_i | y_{1:T})$ : the Forward-Backward algorithm, matching the generative definition, and the Entropic Forward-Backward, which matches the discriminative one. We consider a homogeneous HMM with the following notations:

- $\pi(i) = p(x_t = \lambda_i)$ ;
- $a_i(j) = p(x_{t+1} = \lambda_j | x_t = \lambda_i)$ ;
- $b_i(y_t) = p(y_t | x_t = \lambda_i)$ ;
- $L_{y_t}(i) = p(x_t = \lambda_i | y_t)$ .

### A. HMM as a generative classifier: The Forward-Backward algorithm

The Forward-Backward algorithm consists in computing  $p(x_t = \lambda_i | y_{1:T})$  as follows:

$$p(x_t = \lambda_i | y_{1:T}) = \frac{\alpha_t(i) \beta_t(i)}{\sum_{j=1}^N \alpha_t(j) \beta_t(j)}$$

with forward probabilities  $\alpha$  computed with the following recursion:

$$\alpha_1(i) = \pi(i) b_i(y_1)$$

$$\alpha_{t+1}(i) = b_i(y_{t+1}) \sum_{j=1}^N \alpha_t(j) a_j(i)$$

And the backward probabilities  $\beta$ :

$$\beta_T(i) = 1$$

$$\beta_t(i) = \sum_{j=1}^N \beta_{t+1}(j) a_i(j) b_j(y_{t+1})$$

#### 4.1 B. HMM as a discriminative classifier: The Entropic Forward-Backward algorithm

The Entropic Forward-Backward algorithm consists in computing,  $p(x_t = \lambda_i | y_{1:T})$  as follows:

$$p(x_t = \lambda_i | y_{1:T}) = \frac{\alpha_t^E(i) \beta_t^E(i)}{\sum_{j=1}^N \alpha_t^E(j) \beta_t^E(j)}$$

with entropic forward probabilities  $\alpha^E$  computed as:

$$\alpha_1^E(i) = L_{y_1}(i)$$

$$\alpha_{t+1}^E(i) = \frac{L_{y_{t+1}}(i)}{\pi(i)} \sum_{j=1}^N \alpha_t^E(j) a_j(i)$$

And the entropic backward  $\beta^E$ :

$$\beta_T^E(i) = 1$$

$$\beta_t^E(i) = \sum_{j=1}^N \frac{L_{y_{t+1}}(j)}{\pi(j)} \beta_{t+1}^E(j) a_i(j)$$