

Principled Generative-Discriminative Hybrid Hidden Markov Model

Oksana Yakhenko, Lucian Vlad Lita, Romer Rosales, Stefan Niculescu

Extended abstract In this work, we propose a new probabilistic model which generalizes discriminative properties of a Conditional Random Field, and generative properties of a Hidden Markov Model for labeling and segmenting sequence data. We also present promising preliminary results of application of our model to several natural language processing tasks.

Modeling and learning the probability distribution has a variety of applications in machine learning and classification in particular. The parameters for probabilistic models can be learned in order to maximize the objective function of interest. When the goal is classification, for instance, one is interested in maximizing the probability of the correct label of the input given the input (discriminative models). On the other hand, one may also be interested in learning the entire probability distribution of the input and the output (generative models). For the past several years there has been a great deal of research about which objective function needs to be maximized, and what trade-offs exist between the two training regimes. Several algorithms that combine the strengths of generative and discriminative models were also proposed. Recently, Minka [3] suggested that the trade-off between generative and discriminative models are the choice of the priors for the model. This has been taken further, and Lasserre et al [2] suggested that there is only one way to train a probabilistic model in order to combine generative and discriminative properties of the model.

In the context of structured prediction, Hidden Markov Model (HMM) has been widely used in many applications, such as natural language processing, protein structure prediction, phoneme classification to name a few. HMM is a generative model. Its discriminative equivalent is Conditional Random Field (CRF) [1]. One of the limitations of a CRF is that it is not easy to incorporate the unlabeled data, and a lot of training data is needed for the CRF to achieve good performance accuracy.

We adapt the hybrid generative-discriminative framework to combine generative and discriminative trade-offs for Hidden Markov models and the CRF's¹ and propose a new model which is a generalization of CRF and HMM.

Let \mathcal{X} be the input alphabet and \mathcal{Y} be the alphabet; $X_i = \{x_1 \dots x_l\}$, $x_l \in \mathcal{X}$ be the input sequence and $Y_i = \{y_1 \dots y_l\}$, $y_l \in \mathcal{Y}$ be the output sequence. Given the labeled dataset $D = \{X_i, Y_i\}_{i=1}^N$ and unlabeled dataset $UD = \{X_i\}_{i=1}^U$ the goal is to find a function $f: 2^{\mathcal{X}} \rightarrow 2^{\mathcal{Y}}$. One can find such function by assuming a model with parameters θ , estimating the probability $P(X, Y|\theta)$, and use maximum a posteriori principle to classify a new instance. As in [2] $\theta = \{\theta_G, \theta_D\}$ are assumed of two types: generative θ_G , and discriminative θ_D . The joint distribution is given by:

$$\begin{aligned} p(X, Y|\theta_G, \theta_D) &= p(\theta_G \theta_D) p(Y|X, \theta_D) p(X|\theta_G) \\ &= p(\theta_G \theta_D) \prod_{i=1}^N \frac{p(X_i Y_i|\theta_D)}{\sum_{Y_k \in \mathcal{Y}} p(Y_k X_i|\theta_G)} \prod_{i=1}^{N+U} \sum_{Y_k \in \mathcal{Y}} p(Y_k X_i|\theta_G) \end{aligned}$$

Probabilities $p(X_i Y_i|\theta_G)$ and $p(X_i Y_i|\theta_D)$ are modeled by HMM, and so the summation $\sum_{Y_k \in \mathcal{Y}} p(Y_k X_i|\theta_G)$ can be efficiently computed using backward-forward procedure. Thus training reduces to maximizing the probability above, subject to constraints on local emission and transition probabilities of HMM ($\sum_{x \in \mathcal{X}} p(x|y, \theta) = 1, \forall y \in \mathcal{Y}$ and $\sum_{y \in \mathcal{Y}} p(y|y_j) = 1, \forall y_j \in \mathcal{Y}$).² The maximization is performed using Quadratic Penalty method

¹For brevity we will refer to a discriminative hidden markov model with local probability constraints as a CRF (non-deficient probability distribution for transition and emission weights), however the general formulation of CRF is unconstrained.

²We also note that due to the product of $\sum_{Y_k \in \mathcal{Y}} p(Y_k X_i|\theta_G)$, this framework is difficult to adapt to undirected models, since the joint probability would require normalization over all possible inputs and outputs, or the objective function is unconstrained otherwise.

[5]. The constrained problem is turned into unconstrained problem as

$$\max_{\theta_G, \theta_D} p(X, Y | \theta_G \theta_D) + C_k \left(\sum_{y \in \mathcal{Y}, \theta \in \{\theta_D, \theta_G\}} \left(1 - \sum_{x \in \mathcal{X}} p(x|y, \theta) \right)^2 + \left(1 - \sum_{y_j \in \mathcal{Y}} p(y_j|y, \theta) \right)^2 \right)$$

and is solved sequentially: once the problem converged at step k for a fixed C_k using LBFGS [4], in the next step C_{k+1} is incremented, and the problem is initialized with the solutions from the previous step.

Our preliminary results on natural language processing tasks such as base noun-phrase prediction and chunking (the goal is to predict noun and verb phrases), Japanese named-entity recognition (the goal is to predict 'organization', 'location', and other entities from Japanese text), and segmentation of Chinese phrases (the goal is to predict where the phrase begins and ends). The experimental results and the accuracy of the hybrid model compared to generative HMM and a CRF are promising and are summarized in the Table below. The results are obtained using the data split into train/test set. Our current research includes a thorough experiments on large datasets, applications to real-world named-entity recognition tasks, and finding better/faster solutions to the optimization problem.

Task	CRF	HMM	Hybrid
basenp	78.7	74.9	81.5
chunking	70.8	72.8	76
japaneseNE	91.8	91.5	91.6
segmentation	75.8	70.8	72.8
average	79.3	77.7	80.5

Table 1: Comparison of accuracies of CRF, HMM, and Hybrid HMM on natural language tagging tasks

References

- [1] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning (ICML-01)*, 2001.
- [2] Julia A. Lasserre, Christopher M. Bishop, and Thomas P. Minka. Principled hybrids of generative and discriminative models. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006.
- [3] Tom Minka. Discriminative models, not discriminative training. Technical report, Microsoft Research, 2005.
- [4] J. Nocedal and D.C. Liu. On the limited memory method for large scale optimization. *Mathematical Programming*, 3(45):503–528, 1989.
- [5] Jorge. Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, 2nd edition, 2006.