Xue, J.H. and Titterington, D.M. (2008) *Short note on two output-dependent hidden Markov models.* Pattern Recognition Letters, 29 (9) pp. 1424-1426. ISSN 0167-8655

Short Note on Two Output-dependent Hidden Markov Models

Jing-Hao Xue [a,*], D. Michael Titterington [a]

[a] *Department of Statistics, University of Glasgow, Glasgow G12 8QQ, UK*

Abstract

The purpose of this note is to study the assumption of "mutual information independence", which is used by Zhou (2005) for deriving an output-dependent hidden Markov model, the so-called discriminative HMM (D-HMM), in the context of determining a stochastic optimal sequence of hidden states. The assumption is extended to derive its generative counterpart, the G-HMM. In addition, state-dependent representations for two output-dependent HMMs, namely HMMSDO (Li, 2005) and D-HMM, are presented.

Key words: Discriminative models; Generative models; Mutual information independence; Output-dependent hidden Markov model

* Corresponding author. Tel.: +44 141 330 2474; fax: +44 141 330 4814.
Email addresses: `jinghao@stats.gla.ac.uk` (Jing-Hao Xue), `mike@stats.gla.ac.uk` (D. Michael Titterington).

1 Introduction

Generative models like normal-based discriminant analysis and discriminative models like logistic regression are comprehensively investigated and compared in the machine learning literature (Rubinstein and Hastie, 1997; Ng and Jordan, 2001). Amongst the latent (hidden) variable models for structured data such as time series, hidden Markov models (HMMs) for discrete-valued hidden states and state-space models (SSMs) for continuous-valued hidden states are widely used.

Traditionally, an HMM is generative because it models a distribution $P(O_1^n|S_1^n)$, the data generation process (DGP) of the observed output sequence, $O_1^n = o_1, \ldots, o_n$, given the hidden state sequence, $S_1^n = s_1, \ldots, s_n$, and thus $P(O_1^n|S_1^n)$, a state-dependent term, is included in the criterion for determining a stochastic optimal sequence of hidden states. Recently, Zhou (2005) proposes a discriminative hidden Markov model (D-HMM), which includes output-dependent terms $P(s_t|O_1^n), t = 1, \cdots, n$, in the criterion, based on an assumption of "mutual information independence". Meanwhile, Li (2005) presents the so-called "hidden Markov models with states depending on observations" (HMMSDO), which assumes that the current state s_t depends not only on the last state s_{t-1} but also on the last output o_{t-1}, so that output-dependent terms $P(s_t|s_{t-1}, o_{t-1})$ are included in the criterion.

Both the D-HMM and HMMSDO show superior performance in determining the optimal state sequence for certain applications. Zhou (2005) shows that the D-HMM outperforms the corresponding generative hidden Markov model (G-HMM) for part-of-speech tagging and phrase chunking; Li (2005) shows that

HMMSDO outperforms the standard HMM for prediction of protein secondary structures when the training set is large enough.

In this note, we shall study the assumption of "mutual information independence" that is used for deriving the D-HMM (Zhou, 2005) in the context of determining an optimal state sequence, and then extend it to derive its generative counterpart, the G-HMM. In addition, state-dependent representations for these two output-dependent HMMs will be presented.

2 Generative HMM

Following the notation used by Zhou (2005), the definition of the optimal hidden state sequence S_1^n based on the observed output sequence O_1^n is that of the maximum a posteriori (MAP) estimator S^* of S_1^n:

$$S^* = \underset{S_1^n}{\operatorname{argmax}} \{\log P(S_1^n|O_1^n)\} \ . \tag{1}$$

The G-HMM rewrites the criterion (1) through applying Bayes' theorem and ignoring the item determined purely by O_1^n as

$$S^* = \underset{S_1^n}{\operatorname{argmax}} \{\log P(S_1^n) + \log P(O_1^n|S_1^n)\} \ ,$$

which is further factorised as

$$S^* = \underset{S_1^n}{\operatorname{argmax}} \left\{\log P(S_1^n) + \log \left(P(o_1|S_1^n) \prod_{k=2}^{n} P(o_k|O_1^{k-1}, S_1^n)\right)\right\} \ .$$

In order to make this formulation tractable, an assumption that O_1^n is conditionally independent given S_1^n is in general introduced as, for all $k \in \{2, \ldots, n\}$,

$$P(o_k|O_1^{k-1}, S_1^n) = P(o_k|S_1^n) \ , \tag{2}$$

and thus based on such a conditional independence assumption, the MAP estimator for the G-HMM is simplified to

$$S^* = \underset{S_1^n}{\operatorname{argmax}} \left\{ \log P(S_1^n) + \sum_{i=1}^{n} \log P(o_i|S_1^n) \right\} \ . \tag{3}$$

The G-HMM is regarded as being generative because it directly models the DGP $P(o_i|S_1^n)$ of the observed o_i from the hidden S_1^n.

In practice, as for the standard HMM, the assumption (2) is further simplified to

$$P(o_k|O_1^{k-1}, S_1^n) = P(o_k|S_1^n) = P(o_k|s_k) \ , \tag{4}$$

and thus the MAP estimator of the standard HMM is

$$S^* = \underset{S_1^n}{\operatorname{argmax}} \left\{ \log P(S_1^n) + \sum_{i=1}^{n} \log P(o_i|s_i) \right\} \ . \tag{5}$$

3 Discriminative HMM from Mutual Information Independence

The D-HMM rewrites the criterion (1) through applying Bayes' theorem, but not ignoring the item determined purely by O_1^n, as

$$S^* = \underset{S_1^n}{\operatorname{argmax}} \left\{ \log P(S_1^n) + \log \frac{P(S_1^n, O_1^n)}{P(S_1^n)P(O_1^n)} \right\} \ .$$

To make this formulation tractable, an assumption that the mutual information ($MI\,(S_1^n, O_1^n) = \log \frac{P(S_1^n, O_1^n)}{P(S_1^n)P(O_1^n)}$) between S_1^n and O_1^n is independent with respect to each hidden s_i was introduced by Zhou (2005) as

$$MI(S_1^n, O_1^n) = \sum_{i=1}^{n} MI(s_i, O_1^n) \ , \tag{6}$$

or, in more detail,

$$\log \frac{P(S_1^n, O_1^n)}{P(S_1^n)P(O_1^n)} = \sum_{i=1}^{n} \log \frac{P(s_i, O_1^n)}{P(s_i)P(O_1^n)} = \sum_{i=1}^{n} \log \frac{P(s_i|O_1^n)}{P(s_i)} \ . \tag{7}$$

Based on such a representation, the MAP estimator for the D-HMM is simplified as (Zhou, 2005)

$$S^* = \operatorname*{argmax}_{S_1^n} \left\{ \log P(S_1^n) + \sum_{i=1}^{n} \log P(s_i|O_1^n) - \sum_{i=1}^{n} \log P(s_i) \right\} . \tag{8}$$

The D-HMM is regarded as being discriminative because the criterion (8) includes directly the discriminative process $P(s_i|O_1^n)$, representing an output-dependence of a hidden state s_i on all the observed outputs O_1^n.

We shall make four observations about the D-HMM.

First, it is noted that the criterion (8) is simultaneously to maximise the maximum posterior marginal (MPM) estimator $\sum_{i=1}^{n} \log P(s_i|O_1^n)$ of $\log P(S_1^n|O_1^n)$ and to maximise the distance between the state transition model $\log P(S_1^n)$ and its independent-based counterpart $\sum_{i=1}^{n} \log P(s_i)$.

Second, in order to satisfy the assumption (7) underlying the D-HMM, it is required that

$$\prod_{k=2}^{n} \frac{P(s_k|S_1^{k-1}, O_1^n)}{P(s_k|S_1^{k-1})} = \prod_{k=2}^{n} \frac{P(s_k|O_1^n)}{P(s_k)} .$$

Since this is valid for any value of s_k, it follows that, for all $k \in \{2, \ldots, n\}$,

$$\frac{P(s_k|S_1^{k-1}, O_1^n)}{P(s_k|S_1^{k-1})} = \frac{P(s_k|O_1^n)}{P(s_k)} . \tag{9}$$

Third, the assumption (7) can be rewritten as

$$\log \frac{P(S_1^n, O_1^n)}{P(S_1^n)P(O_1^n)} = \sum_{i=1}^{n} \log \frac{P(s_i, O_1^n)}{P(s_i)P(O_1^n)} = \sum_{i=1}^{n} \log \frac{P(O_1^n|s_i)}{P(O_1^n)} . \tag{10}$$

Based on such a representation, the MAP estimator (8) for the D-HMM can be rewritten, with the term $\sum_{i=1}^{n} \log P(O_1^n)$ determined purely by O_1^n being ignored, as

$$S^* = \operatorname*{argmax}_{S_1^n} \left\{ \log P(S_1^n) + \sum_{i=1}^{n} \log P(O_1^n|s_i) \right\} . \tag{11}$$

Therefore, the D-HMM can also be represented as being generative because the criterion (11) includes a generative-like process $P(O_1^n|s_i)$, representing a state-dependence of all the observed outputs O_1^n on a hidden state s_i.

Fourth, it can be seen that, when the assumption (6) of mutual information independence develops from independence between pairs (s_i, O_1^n) into that between local pairs (s_i, o_i) such that $MI(S_1^n, O_1^n) = \sum_{i=1}^n MI(s_i, o_i)$, the criteria (11) and (8) degenerate into the criterion (5), indicating that the D-HMM degenerates into the standard HMM.

4 Generative HMM from Mutual Information Independence

Furthermore, similarly to the assumption (6) proposed by Zhou (2005), an assumption that mutual information between S_1^n and O_1^n is independent with respect to each observed o_i can be introduced here as

$$MI(S_1^n, O_1^n) = \sum_{i=1}^n MI(S_1^n, o_i) \ , \tag{12}$$

or, in more detail,

$$\log \frac{P(S_1^n, O_1^n)}{P(S_1^n)P(O_1^n)} = \sum_{i=1}^n \log \frac{P(S_1^n, o_i)}{P(S_1^n)P(o_i)} = \sum_{i=1}^n \log \frac{P(o_i|S_1^n)}{P(o_i)} \ . \tag{13}$$

Based on such a representation, we can obtain another generative model and its MAP estimator, with the term $\sum_{i=1}^n \log P(o_i)$ determined purely by O_1^n being ignored, as

$$S^* = \underset{S_1^n}{\operatorname{argmax}} \left\{ \log P(S_1^n) + \sum_{i=1}^n \log P(o_i|S_1^n) \right\} \ . \tag{14}$$

This estimator is in fact the estimator (3) of the G-HMM, *i.e.*, the G-HMM can be derived under the assumption (12), a type of mutual information independence.

Similarly, we shall make three observations about this G-HMM, which is derived from mutual information independence.

First, in order to satisfy the assumption (13) of the G-HMM, it is required that, for all $k \in \{2, \dots, n\}$,

$$\frac{P(o_k|O_1^{k-1}, S_1^n)}{P(o_k|O_1^{k-1})} = \frac{P(o_k|S_1^n)}{P(o_k)} \ . \tag{15}$$

Therefore, under the MAP criterion (1), the conditions (15) and (2) have the same effect on determining the optimal hidden S_1^n.

Second, the assumption (13) can be rewritten as

$$\log \frac{P(S_1^n, O_1^n)}{P(S_1^n)P(O_1^n)} = \sum_{i=1}^{n} \log \frac{P(S_1^n, o_i)}{P(S_1^n)P(o_i)} = \sum_{i=1}^{n} \log \frac{P(S_1^n|o_i)}{P(S_1^n)} \ . \tag{16}$$

Based on such a representation, the MAP estimator (14) for the G-HMM can be rewritten, with the terms related to $\log P(S_1^n)$ being combined, as

$$S^* = \underset{S_1^n}{\operatorname{argmax}} \left\{ (1-n) \log P(S_1^n) + \sum_{i=1}^{n} \log P(S_1^n|o_i) \right\} \ . \tag{17}$$

Therefore, in this sense, the G-HMM can also be represented as being discriminative because the criterion (17) includes a discriminative-like process $P(S_1^n|o_i)$, representing an output-dependence of all the hidden states S_1^n on an observed output o_i.

Third, it can be seen that, when the assumption (12) of mutual information independence develops from independence between pairs (S_1^n, o_i) into that between local pairs (s_i, o_i) such that $MI(S_1^n, O_1^n) = \sum_{i=1}^{n} MI(s_i, o_i)$, the criteria (17) and (14) degenerate into the criterion (5), indicating that the G-HMM degenerates into the standard HMM.

5 Equivalence between G-HMM and D-HMM

Once we assume a fully independent mutual information between any state-output combination (s_i, o_j) as

$$MI(S_1^n, O_1^n) = \sum_{i=1}^{n}\sum_{j=1}^{n} MI(s_i, o_j) \ , \tag{18}$$

or, in more detail,

$$\begin{aligned} \log \frac{P(S_1^n, O_1^n)}{P(S_1^n)P(O_1^n)} &= \sum_{i=1}^{n}\sum_{j=1}^{n} \log \frac{P(s_i, o_j)}{P(s_i)P(o_j)} \\ = \sum_{i=1}^{n}\sum_{j=1}^{n} \log \frac{P(o_j|s_i)}{P(o_j)} &= \sum_{i=1}^{n}\sum_{j=1}^{n} \log \frac{P(s_i|o_j)}{P(s_i)} \ , \end{aligned} \tag{19}$$

this assumption results in two criteria, one generative and the other discriminative, with the MAP estimators as

$$S^* = \underset{S_1^n}{\operatorname{argmax}}\{\log P(S_1^n) + \sum_{i=1}^{n}\sum_{j=1}^{n} \log P(o_j|s_i)\} \ , \tag{20}$$

$$S^* = \underset{S_1^n}{\operatorname{argmax}}\left\{\log P(S_1^n) + \sum_{i=1}^{n}\sum_{j=1}^{n} \log P(s_i|o_j) - \sum_{i=1}^{n} \{n \log P(s_i)\}\right\} \ , \tag{21}$$

respectively. These two criteria are equivalent.

In the context of determining an optimal sequence of hidden states, apart from the equivalence above, up to now, we find two occurrences of equivalence between a discriminative representation of the MAP criterion and its generative counterpart: one is for the D-HMM between the criteria (8) and (11), the other is for the G-HMM between the criteria (17) and (14).

We shall further illustrate such equivalence with two simple but related HMMs: one is a generative-like state-dependent model, which assumes that the current output o_t depends not only on the current state s_t but also on the last state s_{t-1}; the other is a discriminative-like output-dependent model, the so-called

HMMSDO (Li, 2005), which assumes that the current state s_t depends not only on the last state s_{t-1} but also on the last output o_{t-1}.

The joint distribution of the first generative-like state-dependent model is

$$P(S_1^n, O_1^n) = P(s_1)P(o_1|s_1)\prod_{i=2}^{n} P(s_i|s_{i-1})P(o_i|s_i, s_{i-1}) \ . \tag{22}$$

This distribution can be rewritten as

$$\begin{aligned} P(S_1^n, O_1^n) &= P(o_1, s_1)\prod_{i=2}^{n} P(s_i, o_i|s_{i-1}) \\ &= P(o_1)P(s_1|o_1)\prod_{i=2}^{n} P(o_i|s_{i-1})P(s_i|s_{i-1}, o_i) \ , \end{aligned} \tag{23}$$

which leads to a discriminative-like output-dependent part $P(s_i|s_{i-1}, o_i)$ in the distribution. In fact, the difference between the probabilistic directed acyclic graphs (DAGs) corresponding to the joint distributions (22) and (23) is only in that directions of edges from s_i to o_i are reversed.

Similarly, the joint distribution of the discriminative-like output-dependent HMMSDO, with $P(s_i|s_{i-1}, o_{i-1})$ included, is (Li, 2005)

$$P(S_1^n, O_1^n) = P(s_1)P(o_1|s_1)\prod_{i=2}^{n} P(s_i|s_{i-1}, o_{i-1})P(o_i|s_i) \ . \tag{24}$$

This distribution can be rewritten as

$$\begin{aligned} P(S_1^n, O_1^n) &= P(s_1)P(o_n|s_n)\prod_{i=2}^{n} P(s_i, o_{i-1}|s_{i-1}) \\ &= P(s_1)P(o_n|s_n)\prod_{i=2}^{n} P(s_i|s_{i-1})P(o_{i-1}|s_i, s_{i-1}) \ , \end{aligned} \tag{25}$$

which leads to a no longer discriminative-like output-dependence in the distribution. In fact, the difference between the DAGs corresponding to the joint distributions (24) and (25) is only in that directions of edges from s_i to o_{i-1} are reversed. In practice, whether or not $P(o_{i-1}|s_i, s_{i-1})$ is reasonable needs to be justified, because it means that the current output depends on the next state.

6 Summary

This note has suggested that the mutual information assumption (12) resulted in the G-HMM, while another mutual information assumption (6) resulted in the D-HMM. However, in practice, whether or not the assumptions are reasonable and how the corresponding HMMs perform can be data-dependent; research efforts to explore an adaptive switching between or combination of these two models may be worthwhile. Meanwhile, this note has suggested that the so-called output-dependent HMMs could be represented in a state-dependent manner, and vice versa, essentially by application of Bayes' theorem.

References

Li, Y., 2005. Hidden Markov models with states depending on observations. Pattern Recognition Letters 26 (7), 977–984.

Ng, A. Y., Jordan, M. I., 2001. On discriminative vs. generative classifiers: a comparison of logistic regression and naïve bayes. In: NIPS. pp. 841–848.

Rubinstein, Y. D., Hastie, T., 1997. Discriminative vs. informative learning. In: KDD. pp. 49–53.

Zhou, G. D., 2005. Direct modelling of output context dependence in discriminative hidden Markov model. Pattern Recognition Letters 26 (5), 545–553.