Semisupervised Hidden Markov Models for Part-of-Speech tagging

Camila Da Souza
Matilda Grahn

Abstract

Bayesian semisupervised Hidden Markov Models are used for Part-of-speech-tagging in the following study. Three models with different priors are compared, where the results indicate that the third model which has the most informative prior provides more reliable results, regardless of predictive capacity.

1. Introduction

In the following project, Hidden Markov Models (HMM) are studied and used in a light version of Part-of-speech (POS) tagging.

In summary, the HMM is a probabilistic model that can be used to model hidden non-observable states by observing past sequences. Here, we have observed sequences of words, which are also known as tokens, where their respective POS category such as noun, verb, adjective, etc., are seen as hidden latent states.

Given a new sentence with an unobserved state sequence, the model predict the most probable hidden state POS sequence using the Viterbi algorithm.

This type of stochastic method for POS-tagging is frequently used as a preprocessing task in other natural language processing (NLP) analyses, such as sentiment analysis (Martinez, 2011).

In this study the analysis starts with a semi-supervised HMM with an arbitrarily chosen non-informative prior, which is then updated and modified in two subsequent models.

2. Data

The data used is a subset of the Name Entity Recognition (NER) data set provided by Kaggle user Debasis Samal (Samal), which contain words of sentences taken from a VOA news article about an anti-war protest in London from 2009 (News). The subset contains 98 words, and their respective Part-Of-Speech (POS) category. We have limited the categories to four: Noun (NN), Verb (VB), Adjective (JJ), and Other.

We further have constructed a sentence using words and a POS-tag sequence available in the data subset, which we use to evaluate how well our Hidden Markov Models predict the hidden sequences. The sentence is the following:

The, party, was, ruling, in, the, Houses, of, Parliament.

Other, NN, VB, VB, Other, Other, NN, Other, NN.

3. Models

The Hidden Markov Model (HMM) is a probabilistic model that is used to model non-observable (”hidden”) state variables z by their associated observed variables w. An assumption of the HMM is that the latent states form a Markov chain, such that hidden state z_t is independent of other variables conditioned on z_{t-1} (Guide).

In this study, a sequence of words is observed, in this context commonly referred to as tokens, w, which have an associated part-of-speech, which are the hidden states z. A simple Hidden Markov Model is visualized in Figure 1.

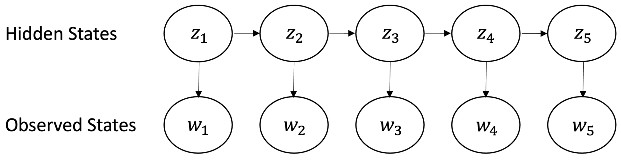

Figure 1. Visualization of simple Hidden Markov Model.

The Markov chain is parameterized by a $K \times K$ transition matrix θ, containing the probabilities of transition to a state z_t from a previous state $z_t - 1$, that is, the probability of a POS category, given some previous POS category, where K equals the total number of POS categories. Further, the HMM has a $K \times V$ emission matrix, where V equals the number of unique tokens, containing the probabilities of a token at time t, given a POS category at time t.

The priors for these parameters, i.e. emission and transition probabilities, are set to follow a Dirichlet distribution with parameter α and β respectively, as follows:

$$\theta_k \sim Dir(\alpha_k) \text{ for } k\text{=1,2,...,}K \quad (Transition) \tag{1}$$

$$\phi_k \sim Dir(\beta_v) \text{ for } k\text{=1,2,...,}K \quad (Emission) \tag{2}$$

where α is a vector with K positive real values and β is a vector with V positive real values.

The continuous multivariate Dirichlet distribution is often used to describe the probabilities of discrete probability distributions, such as the categorical distribution for which it is a conjugate prior, and is parameterized by a vector with K positive elements determining the distribution and concentration.

The observed variables w_t and the states z_t are set to follow a categorical distribution, which is a generalized Bernoulli distribution with (K >2), with parameters ϕ_{Z_t} and $\theta_{z_{t-1}}$.

$$w_t \sim Cat(\phi_{zt}) \text{for } t\text{=1,2,...,}T \quad (Observedvariables) \tag{3}$$

$$z_t \sim Cat(\theta_{z_{t-1}}) \text{ for } t\text{=1,2,...,}T \qquad (Hidden\, states) \tag{4}$$

The description above applies to a general supervised HMM, where both a sequence of tokens w and a hidden state sequence z are observed.

In a second step, the forward algorithm is used to build a semi-supervised HMM. The algorithm estimates the probability of an observed unsupervised sequence of T_{unsup} tokens w_{new}, with an unobserved state sequence z, and it is in this fashion we get a log likelihood $log\, P(w_{new}|z)$. Here, T_{unsup} stands for the total number of unsupervised tokens.

Then, it is the Viterbi algorithm that is used to find the most probable sequence of states z^*, given the observed sequence of T_{unsup} tokens w_{new} and the semi-supervised HMM with transition matrix θ and emission matrix ϕ:

$$z^* = \underset{z}{\operatorname{argmax}} \prod_{t=1}^{T_{unsup}} P(w_{new,t}|z_t)P(z_t|z_{t-1}). \tag{5}$$

The models are run in Stan via R. After generating the predictions z^* a more simple evaluation and comparison of different models is done by looking at the prediction accuracy level. Further, we carry out Pareto smoothed importance-sampling leave-one-out cross-validation (PSIS-LOO CV) for the supervised HMM, given that evaluations using LOO CV or Watanabe Information Critera (WAIC) can not be done for semisupervised HMM using the Viterbi algorithm to predict the most probable states (Vehtari).

4. Results

The analysis starts with a semi-supervised HMM with arbitrarily chosen non-informative priors α and β, which is then updated and modified in two subsequent models. The prior parameter values of α and β for each model is seen in Table 1, and are commented on in subsections below.

Table 1. Dirichlet prior values

MODEL	α	β
1	0.1	0.1
2	0.8	0.8
3	2	2

4.1. Convergence and efficiency

The convergence diagnostic $\hat{R}$ yields the ratio between- and within-chain estimates, thus a value close to 1 is desirable. The $\hat{R}$ used is the Rank normalized $\hat{R}$ on split chains, where the ranks of the drawn parameters are inverse normal transformed and hence close to normally distributed. Since the maximum values for the models does not exceed the recommended limit $\hat{R} \leq 1.01$, as seen in Table 2 showing the values for each model, there is no indication that either of the models fail to converge. It can be noted that the value for Model 2 is slightly lower than for Model 3.

Table 2. Percentage of correctly predicted POS-category, Divergent transitions, Maximum $\hat{R}$ and Minimum Effective Sample Size

	MODEL 1	MODEL 2	MODEL 3
CORRECT PRED.	88.89 %	66.67 %	44.44 %
DIV. TRANS.	622	-	-
RHAT	1.006	1.003	1.004
ESS	1310	2021	2051

$\hat{R}$=Maximum $\hat{R}$
ESS=Minimum Effective Sample Size

Further, the minimum effective sample size is presented in Table 3. The value is the ratio between the samples and the autocorrelations, which for each model is 372, 1267 and 3897 respectively. The result shows that Model 3 has the lowest autocorrelations and therefore the highest information content.

Table 3. Pareto k diagnostics

	MODEL 1	MODEL 2	MODEL 3
	K (ESS)	K (ESS)	K (ESS)
(-INF, 0.5]	31.6 (372)	60.2 (1267)	95.9 (3897)
(0.5, 0.7]	11.2 (164)	34.7 (809)	4.1 (3994)
(0.7, 1]	40.8 (19)	5.1 (530)	0 (NA)
(1, INF]	16.3 (3)	0 (NA)	0 (NA)

ESS=Minimum Effective sample size
K= percent of k-values

4.2. Prediction results

Percentage of correctly predicted POS-category, Divergent transitions, Maximum $\hat{R}$ and Minimum Effective Sample Size are reported in Table 2.

The first model reached 622 divergent transitions when compiling, which is one of the symptoms when the simulated Hamiltonian MC course departs from the true course and fail to approach the target distribution. This strongly suggests that Model 1 is misspecified. To ease the exploration, one solution is to assign a slightly more informative prior for the parameters (Table 1). The values for the parameters in Model 1 are both 0.1, and in Model 2 and 3, the parameters are set to 0.8 and 2 respectively, which eliminates the divergent transition issues.

Looking at the predictions, the number of correctly predicted states is by far the highest for Model 1 (88.89 %), closely followed by Model 2 (66.67 %), as presented in Table 2. Model 3 has the lowest number of accurate predictions (44.44 %).

Pareto smoothed importance-sampling leave-one-out cross-validations (PSIS-LOO CV) are done for the supervised Hidden Markov Models, given that evaluations using LOO CV or Watanabe Information Critera (WAIC) can not be done for semisupervised HMM using the Viterbi algorithm to predict the most probable states (Vehtari). The results are reported in Table 4. It is still valuable to carry out the analyses for the supervised HMM to evaluate how each model generalizes to unseen data, and understand the effect of different priors.

Table 4. Elpd, Effective number of parameters and loo information criteria

	MODEL 1	MODEL 2	MODEL 3
ELPD LOO	431.5 (15.1)	-406.4 (9.1)	-405 (6)
P LOO	128.6 (7.9)	58.5 (2.8)	32 (1)
LOOIC	863.1 (30.2)	812.8 (18.3)	809 (12)

The $elpd_{loo}$-values, which are the unbiased estimates of the expected log pointwise predictive density (LOO-estimate of out-of-sample prediction), (431.5, -406.4, -405), are all approximately zero when exponentiated. However, when evaluating the estimates it is seen (Table 3) that a large proportion of the Pareto k values are too high for Model 1 (57.1 $\%$ of $k > 0.7$) and Model 2 (5.1 $\%$ of $k > 0.7$). Because of the high k-values, the Monte Carlo standard error (MCSE) of $elpd_{loo}$ is NA. This indicates that the $elpd_{loo}$-estimates are not reliable. For Model 3, all $k \leq 0.7$ and the Monte Carlo SE of $elpd_{loo}$ is 0.1. The p_{loo}-values, which are the estimates of the effective number of parameters, i.e. the difference between the log pointwise predictive density for the posterior simulations, and the loo-posterior are 128.6, 58.5 and 31.5. All values are less than the total number of parameters in the model (p=312), which indicates that importance sampling can correct the difference between the full posterior and the loo-posterior. Hence, there is no evidence of misspecification of the models, with regards to the effective number of parameters. The $looic$-values (loo-cv information criteria) are 863.1, 812.8 and 809 for each model respectively, showing that Model 3 produces the lowest prediction error.

5. Conclusions

When examining the reliability of the estimates and whether the models are correctly specified, the results are mixed. When looking only at the accuracy level of predicted states, Model 1 is clearly the model of choice. However, one must not forget about the issues of divergent transitions which indicate that the model is misspecified. Meanwhile, the results of the PSIS-LOO CV analysis suggest that the model of choice is Model 3, which has the lowest prediction error. This could indicate that the previously mentioned proportion of correct predictions are overoptimistic for Model 1 and 2. Clearly, such mixed results leads to the conclusion that further studies are warranted. Potential improvements include using a bigger data set, and using a longer sequence of tokens to predict states for. Further, more variations of the semisupervised Hidden Markov Model could be studied, with other priors or prior parameter values.

References

Guide, S. U. Hidden markov models. URL https://mc-stan.org/docs/stan-users-guide/hmms.html.

Martinez, A. R. Part-of-speech tagging. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(1):107, 2011. doi: https://doi-org.ezproxy.its.uu.se/10.1002/wics.195.

News, V. Thousands protest iraq war in london. URL https://www.voanews.com/a/a-13-thousands-protest-iraq-war-in-london/303183.html.

Samal, D. Name entity recognition (ner) dataset. URL https://www.kaggle.com/datasets/debasisdotcom/name-entity-recognition-ner-dataset.

Vehtari, A. Stan user's mailing list: Waic or loo-cv for a hidden markov model of animal movement. URL https://groups.google.com/g/stan-users/c/Nb_yXztKVFY/m/y5S7qiZWAAAJ.