Spectral Algorithm for learning HMM

1 Spectral method

The predominant learning algorithms have been local search heuristics such as Gibbs sampling and EM algorithm. An alternative approach to the HMM learning problem is the method of moments. The underlying idea is to sample empirical moments from experimental data, and then to find model parameters that yield expected values equal to the sampled quantities. Efficient learning algorithms for HMMs using this approach were enabled by a spectral decomposition technique. These algorithms can also be referred to as spectral methods. In contrast to the iterative approach used by the local search heuristics, spectral methods are not susceptible to local optima. Especially when learning parameter estimates from large observation spaces, this is an important advantage. Hsu's paper estimates HMM efficiently on trigrams of observations which have been projected onto a low dimensional space. Their method is to learn a matrix U that projects observations onto a space of the same dimension as the hidden state.

2 SVD

In linear algebra, the singular value decomposition (SVD) is a factorization of a matrix. It is the generalization of the eigen-decomposition of a positive semidefinite normal matrix (for example, a symmetric matrix with positive eigenvalues) to any $m \times n$ matrix via an extension of polar decomposition. It has many useful applications in signal processing and statistics.

The SVD theorem states:

$S = UDV^T$

Where $U^TU = I_{m\times m}$

$V^TV = I_{n\times n}$ (i.e. U and V are orthogonal)

Where the columns of U are the left singular vectors; D (the same dimensions as S) has singular values and is diagonal (mode amplitudes); and V^T has rows that are the right singular vectors (expression level vectors). SVD has the property that the best rank k approximation of a matrix can be obtained from the first k left singular vectors (corresponding to the first k singular values arranged in descending order), the first k right singular vectors, and the first k singular values. Stopping at k singular vectors, where $k \leq v$ is called "thin" SVD.

3 CCA

In statistics, canonical-correlation analysis (CCA) is a way of making sense of cross-covariance matrices. If we have two vectors $X = (X_1, ..., Xn)$ and $Y = (Y1, ..., Ym)$ of random variables, and there are correlations among the variables, then canonical-correlation analysis will find linear combinations of the X_i and Y_j which have maximum correlation with each other.

Canonical-correlation analysis seeks vectors a and b such that the random variables a^TX and b^TY maximize the correlation $\rho = corr(a^TX, b^TY)$.

Derivation Let $\Sigma_{XX} = cov(X, X)$ and $\Sigma_{YY} = cov(Y, Y)$. The parameter to maximize is

$\rho = \frac{a^T\Sigma_{XY}b}{\sqrt{a^T\Sigma_{XX}a}\sqrt{b^T\Sigma_{YY}b}}$

The first step is to define a change of basis and define

$c = \Sigma_{XX}^{1/2}a$, $d = \Sigma_{YY}^{1/2}b$.

And thus we have

$\rho = \frac{c^T\Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YY}^{-1/2}d}{\sqrt{c^Tc}\sqrt{d^Td}}$.

By the Cauchy-Schwarz inequality, we have

$c'\Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YY}^{-1/2}d \leq \left(c'\Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YY}^{-1/2}\Sigma_{YY}^{-1/2}\Sigma_{YX}\Sigma_{XX}^{-1/2}c\right)^{1/2}(d'd)^{1/2}$,

$\rho \leq \frac{\left(c'\Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}\Sigma_{XX}^{-1/2}c\right)^{1/2}}{(c'c)^{1/2}}$.

There is equality if the vectors d and $\Sigma_{YY}^{-1/2}\Sigma_{YX}\Sigma_{XX}^{-1/2}c$ are collinear. In addition, the maximum of correlation is attained if c is the eigenvector with the maximum eigenvalue for the matrix $\Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}\Sigma_{XX}^{-1/2}$. The subsequent pairs are found by using eigenvalues of decreasing magnitudes. Orthogonality is guaranteed by the symmetry of the correlation matrices.

4 Notation

Hidden state h_t

Observation x_t

Transition probability matrix with $T_{ij} = P(h_{t+1} = i|h_t = j) \in \mathbb{R}^{m*m}$

Observation probability matrix with $O_{ij} = P(x_t = i|h_t = j) \in \mathbb{R}^{n*m}$

Initial state distribution $\pi_i = P(h_1 = i) \in \mathbb{R}^m$

5 Observable representation for HMMs

Theorem 1. *The joint probability can be expressed as:*

$$P(x_1, ...x_t) = \sum_h {}_{t+1}P(h_{t+1}, x_1, ..., x_t) = 1_m^T A_{x_t}...A_{x_1}\pi \tag{1}$$

In which $A_x = Tdiag(O_{x_1}, ..., O_{x_m})$

Proof. We begin by making the assumption that $P(h_t, x_{t-1}, ...x_1) = A_{x_{t-1}}...A_{x_1}\pi$, since $P(h_1) = \pi$, and $P(h_2, x_1) = \sum_{h_1} P(h_2|h_1)P(x_1|h_1)P(h_1) = Tdiag(O_{x_1})\pi = A_{x_1}\pi$ Then for general t,

P(h$_{t+1}, x_t, ..., x_1)$
$= \sum_{h_t} P(h_{t+1}|h_t)P(h_t, x_{t-1}, ..., x_1)$
$= Tdiag(O_{x_t})A_{x_{t-1}}...A_{x_1}\pi$
$= A_{x_t}...A_{x_1}\pi$.

So $P(x_t, ..., x_1) = \sum_{h_{t+1}} P(h_{t+1}, x_t, ..., x_1) = 1_m^T A_{x_t}...A_{x_1}\pi$. □

However, A_x is not directly learnable.

Define a random variable $y{=}U^T\delta_{xt}$, where U has orthonormal columns and is a matrix mapping from observations to the reduced dimension space.

We show below that

Theorem 2.

$$P(x_{1:t}) = c_\infty^T C_{y_t}...C_{y_1}c_1 \tag{2}$$

where

$c_1 = \mu$

$c_\infty^T = \mu^T \Sigma^{-1}$

$C_y = C(y) = K(y)\Sigma^{-1}$

and $\mu = E(y_1)$, $\Sigma = E(y_2 y_1^T)$, *and* $K(y) = E(y_3 y_1^T y_2^T)$, *they are easy to estimate using the method of moments. K() is a tensor, it is linear in each of the three reduced dimension observations* y_1, y_2 *and* y_3.

And c_1, c_∞, C_y *follows from a telescoping product of the following items:*

$c_1 = U^T O \pi$

$c_\infty^T = 1^T (U^T O)^{-1}$

$C_Y = C(y) = U^T O A_x (U^T O)^{-1}$

Proof. The first three moments of the data from an HMM yield the follwing non-observable form:

$$\begin{aligned} &\text{E}(\text{X}_1) \\ &= E[E[\vec{x_1}|h]] \\ &= E[O\vec{e_h}] \\ &= O\pi \end{aligned}$$

$$\begin{aligned} &\text{E}(\text{X}_2 \otimes X_1) \\ &= E[E[\vec{x_2} \otimes \vec{x_1}|h]] \\ &= E[E[\vec{x_2}|h] \otimes E[\vec{x_1}|h]] \\ &= E[(OT\vec{e_h}) \otimes (O\vec{e_h})] \\ &= OTdiag(\vec{\pi})O^T \end{aligned}$$

$$\begin{aligned} &\text{E}(\text{X}_2 \otimes X_1) \\ &= E[E[(\vec{x_3} \otimes \vec{x_1}) < \vec{\eta}, \vec{x_2} > |h]] \\ &= E[E[\vec{x_3}|h] \otimes E[\vec{x_1}|h] < \eta, E[\vec{x_2}|h] >] \\ &= E[(OT\vec{e_h}) \otimes (O\vec{e_h} < \eta, OT\vec{e_h} >] \\ &= OTdiag((O^T\eta))Tdiag((\pi))O^T \end{aligned}$$

Then we can estimate the reduced dimensional form of the three moments with reduced data $y = U^T x$, based on the representation of moments of X above, we have:

$E(Y_1) = U^T O \pi$

$E(Y_2 \otimes Y_1) = U^T O T diag(\pi) O^T U$

$E(Y_3 \otimes Y_1 \otimes Y_2)(y) = U^T O T diag(O^T U y) T diag(\pi) O^T U$

Then we have:

$C(y) = E(Y_3 \otimes Y_1 \otimes OY_2)(y) E(Y_2 \otimes Y_1)^{-1}$

$c_1 = E(Y_1)$

$c_\infty^T = E(Y_1)^T E(Y_2 \otimes Y_1)^{-1}$

we will first write the characteristics μ, Σ and K represent moments of Y in terms of the theoretical matrices, T, O, U, and π:

$\mu = E(Y_1) = U^T O \pi$

$\Sigma = E(Y_2 \otimes Y_1) = U^T O T diag(\pi) O^T U$

$\Sigma^{-1} = E(Y_2 \otimes Y_1)^{-1} = (O^T U)^{-1} diag(\pi)^{-1} (U^T O)^{-1}$

$K(y) = E(Y_3 \otimes Y_1 \otimes Y_2)(y) = U^T O T diag(O^T U y) T diag(\pi) O^T U$

By definition, we have

$c_1 = \mu = U^T O\pi$

likewise,

$$\begin{aligned}
&c_\infty^T = \mu^T \Sigma^{-1} \\
&= (\pi^T O^T U)((O^T U)^{-1} diag(\pi)^{-1} T^{-1} (U^T O)^{-1}) \\
&= \pi^T diag(\pi)^{-1} T{-}1 (U^T O)^{-1} \\
&= 1^T T^{1} (U^T O)^{-1} \\
&= 1^T (U^T O)^{-1}
\end{aligned}$$

For C,

$$\begin{aligned}
&C(y) = K(y)\Sigma^{-1} \\
&= U^T O T diag(O^T U y) T diag(\pi) O^T U \Sigma^{1} \\
&= U^T O T diag(O^T U y)(U^T O)^{-1}
\end{aligned}$$

Note that UU^T is a projection operator and since its range is the same as that of O we have $O^T UU^T = O^T$. So, if $y = U^T \delta_x$, then:

$$\begin{aligned}
&C(y) = U^T O T diag(O^T UU^T \delta_x)(U^T O)^{-1} \\
&= U^T O T diag(O^T \delta_x)(U^T O)^{-1} \\
&= U^T O A_X (U^T O)^{-1}
\end{aligned}$$

□

Theorem 3. *If we define:*

$$b_1 = U^T P_1 \tag{3}$$

$$b_\infty = (P_{2,1} U)^+ P_1 \tag{4}$$

$$B_x = (U^T P_{3,x,1})(U^T P_{2,1})^+ \tag{5}$$

where

[$P_1]_i = P(x_1 = i)$.

[$P_{2,1}]_{ij} = P(x_2 = i, x_1 = j)$.

[$P_{3,x,1}]_{ij} = P(x_3 = i, x_2 = x, x_1 = j)$

Then we have

$$P[x_{1:t}] = 1_m^T A_{x_t} ... A_{x_1} \pi = b_\infty^T B_{x_t} ... B_{x_1} b_1 \tag{6}$$

In Hsu's paper, there are two conditions needed to be satisfied:

Condition1 (HMM Rank Condition) $\pi > 0$ element-wise, and O and T have rank m.

Condition2 (Invertibility Condition) $U^T O$ is invertible.

Theorem 4. *When condition 1 and 2 satisfied, then :*

$$b_1 = (U^T O)\pi \tag{7}$$

$$b_\infty^T = 1_m^T (U^T O)^{-1} \tag{8}$$

$$B_x = (U^T O) A_x (U^T O)^{-1} \tag{9}$$

Proof. Since $P_1 = O\pi$, then $b_1 = U^T P_1 = (U^T O)\pi$,

$P_1^T = \pi^T O^T = 1_m^T T diag(\pi) O^T = 1_m^T (U^T O)^{-1} U^T O T diag(\pi) O^T = 1_m^T (U^T O)^{-1} U^T P_{2,1}$

then b$_\infty^T = P_1^T (U^T P_2, 1)^+ = 1_m^T (U^T O)^{-1} U^T P_2, 1 (U^T P_2, 1)^+ = 1_m^T (U^T O)^{-1}$

P$_{3,x,1} = O A_x T diag(\pi) O^T = O A_x (U^T O)^{-1} U^T P_{2,1}$,

then B$_x = (U^T P_{3,x,1})(U^T P_2, 1)^+ = (U^T O) A_x (U^T O)^{-1}$

b$^T_\infty B_{x_t}...B_{x_1}b_1$
$= 1_m^T(U^TO)^{-1}(U^TO)A_{x_t}(U^TO)^{-1}...(U^TO)^{-1}(U^TO)A_{x_1}(U^TO)^{-1}(U^TO)\pi$
$= 1_m^T A_{x_t}...A_{x_1}\pi$
$= P[x_{1:t}]$ $\square$

6 Comparison of two papers

These two papers both implement a fast spectral method to approximate HMM in contrast to the usual slow methods like EM or Gibbs sampling.

The major modeling difference is to replace B_x in equation 3 with the lower dimensional tensor C(y) which depends on the reduced dimension projection $y = U^T\delta_x$ instead of the unreduced x.

Their models can be related as follows:

$$P(x_{1:t}) = 1_m^T A_{x_t}...A_{x_1}\pi \tag{10}$$

$$= b_\infty^T B_{x_t}...B_{x_1}b_1 \tag{11}$$

$$= c_\infty^T C_{y_t}...C_{y_1}c_1 \tag{12}$$

where

$$y = U^T\delta_x \tag{13}$$

where b_1, b_∞, B_x are as defined in equation (5), (6), (7). And

$c_1 = \mu$

$c_\infty^T = \mu^T\Sigma^{-1}$

$C_y = C(y) = K(y)\Sigma^{-1}$

and $\mu = E(y_1)$, $\Sigma = E(y_2y_1^T)$, and $K(y) = E(y_3y_1^Ty_2^T)$, and $K()$ is a tensor, it is linear in each of the three reduced dimension observations y_1, y_2 and y_3.

And c_1, c_∞, C_y follows from a telescoping product of the following items:

$c_1 = U^TO\pi$

$c_\infty^T = 1^T(U^TO)^{-1}$

$C_Y = C(y) = U^TOA_x(U^TO)^{-1}$

Both methods deal with HMMs with states h_1, h_1, h_3 which emit observations x_1, x_2, x_3. But they have different constraint. Equation (16) requires Condition 1 and 2 to be satisfied, and (17) requires $range(O) \subset range(U)$.

For Hsu's method, x_1 is hit by $(P_{2,1}^TU)^+$ to make a lower dimensional z_1, and x_3 is hit by U^T to reduce its dimension. However, x_2 is left unchanged in this method. B_x is estimated with $E(y_2z_1^T\delta_{x_2}^T)$, which is a tensor of size $m*m*v$.

For Rodu's method, x_1, x_2, x_3 are projected onto lower dimensional space with obervations y_1, y_2, y_3 by U, so the core statistic C_y is computed based on $K = E(y_3y_1^Ty_2^T)$ which is $m*m*m$ tensor.

Compared with Hus's method, for Rodu's method, by reducing the size of the matrix that is estimated, a lower sample complexity can be achieved. In particular, the sample complexity does not depend on the size of the vocabulary nor on the frequency distribution of the vocabulary.

For conditional probability, the two papers share the same idea of recursive update b_t, and $P(x_t|x_{1:t-1}) = b_\infty^TB_{x_t}b_t$ with recursive updates $b_{t+1} = \frac{B_{x_t}b_t}{b_\infty^TB_{x_t}b_t}$. And $b_t = (U^TO)h_t$, which is a linear function of the conditional expectation of the unobservable hidden state h_t.

References

[1] Foster, Dean P., Jordan Rodu, and Lyle H. Ungar. Spectral dimensionality reduction for hmms. preprint arXiv:1203.6130 (2012).

[2] Hsu, Daniel, Sham M. Kakade, and Tong Zhang. A spectral algorithm for learning hidden Markov models. Journal of Computer and System Sciences 78.5 (2012): 1460-1480.