COURSE FINAL PROJECT

AUTOMATIC EPOCH SEGMENTATION OF DARWIN'S READING NOTEBOOKS

March 30, 2016

Liang Chen

Course Project for STAT-S 675

INTRODUCTION

Murdock *et al.* [1] shed lights on Charles Darwin's reading trajectory via mining both local and global patterns in the evolvement of his reading flavor. They analyzed 669 books Darwin read from 1838 to 1860. A probabilistic topic model was used to represent each book with a "semantic vector" lying on a $(k-1)$-dimensional simplex, given the topic number k. The reading trajectory was traced by connecting these data points in temporal order. The experiment also performs *epoch* segmentation to show different reading patterns of text-to-text (local) and past-to-text (global) *surprise* in this trajectory.

In this report, we want to explore a new way to segment and interpret the reading *epochs* in an unsupervised manner. We offer insights to the epoch segmentation problem via k-means clustering and Hidden Markov Model (HMM).

THE MODEL

According to [1], the reading surprise is defined as the KL divergence between the probabilistic distributions for two different books over a certain number of topics. Suppose the two distributions are denoted as p and q, then

$$D(p,q) = \sum_{i=1}^{k} p_i \log_2 \frac{p_i}{q_i} \tag{1}$$

where k is the number of topics, p_i is the probability of the first book assigned to the i-th topic, and analogous is q_i for the second book. The KL divergence describes how one distribution (p) deviates from the other (q). This measurement is asymmetric; in this report, we only use the one-way comparison from current book to previous ones. In particular, we just use $D(p_{current}, q_{previous})$ and $D(p_{current}, AVG(q_{previous}))$ as text-to-text surprise and past-to-text surprise. This is consistent with the definitions in the original paper draft.

Different from the paper draft, which analyzes local and global surprises independently, we look at the *joint* distribution of these two variables. For each book, we assign a 2-dimensional feature vector containing its text-to-text and past-to-text surprises. The feature space is defines as:

$$\mathtt{F} = \{f_i | f_i = (x_i, y_i), x_i \in \mathtt{L}, y_i \in \mathtt{G}\} \tag{2}$$

where L and G denote the sets of text-to-text and past-to-text surprises respectively. We use $P(f_i)$ to represent the probability of feature f_i.

We define several different states corresponding to different patterns of f_i. For instance, if the state implies high local exploitation but low global exploration, then feature f_i would be characterized as a pair of $(high, low)$ value.

The difficult part is that we do not know the patterns a prior so that we need to discover the surprise patterns and also perform segmentation based on these patterns. Here we apply unsupervised clustering to generate pattern candidates from data to avoid bias from artificial declarations. We generate multiple clusters in different scales such that a fine-grained set of pattern candidates will be exploited for the later segmentation task.

We use k-means to produce this candidate set, varing k from 2 to 4. We obtain $9(2+3+4)$ clusters in total, each representing its own surprise pattern. We train multivariate normal distributions over these 9 clusters independently, to form the likelihood model of pattern candidates:

$$\mathtt{C} = \{c_i | c_i : \text{i-th cluster}, 1 \le i \le 9\} \tag{3}$$

$$\mathtt{M} = \{m_i | m_i \sim MVN(\mu_i, \Sigma_i | c_i), c_i \in C\} \tag{4}$$

m_i is the candidate model we will use in epoch segmentation. The state space for epoch inference is defined as

$$\mathtt{E} = \{e | e = (m, i), m \in M, 1 \le i \le k\} \tag{5}$$

i is the index of epoch state e, which ranges from 1 to k, and k is the number of epochs to be estimated.

We formulate the epoch segmentation using *maximum likelihood estimation*, subject to the hard constraints enforcing the sequence to be continuously segmented:

$$L^* = argmax_{\{e_i\}} \sum_i \log P(f_i | e_i), 1 \le i \le n \tag{6}$$

$$\text{subject to } \phi(e_i, e_{i+1}) \tag{7}$$

where n is the total number of observations (surprise features of books). The constraints ϕ is intuitively illustrated by Fig. 1. In this diagram, we assume the number of epochs is 3, and display the possible transitions between states, without losing the generalizability for the models with an arbitrary number of epochs.

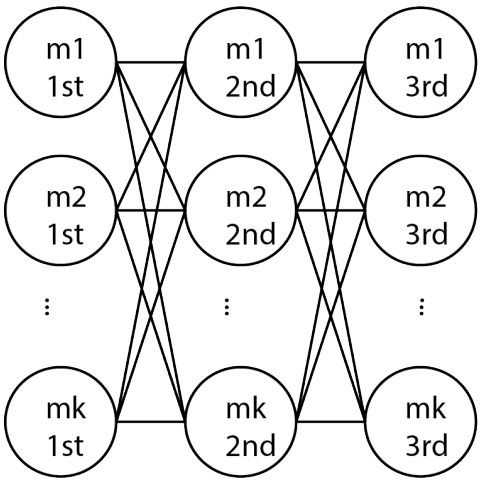


Figure 1: Legal Epoch State Transitions

We perform the inference of $\{e_i\}$ via Dynamic Programming and estimate the optimal sequence of epochs according to Enq. 7. This inference simultaneously estimates the locations of epochs and the best candidate model for each epoch.

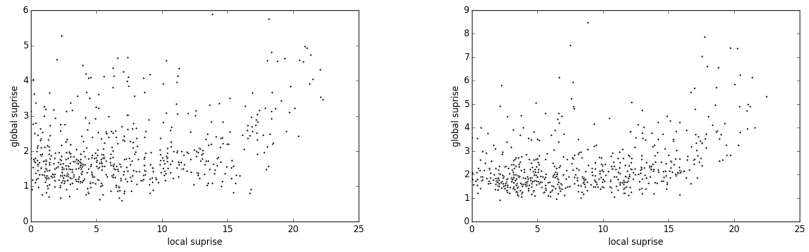
The experiments will be elaborated in the following section. We hypothesize on different epoch numbers and infer the optimal segmentation underneath each hypothesis. This approach computes the log likelihood of the observation given a specific model. It can be used to measure the fitness of different models to the same data.

EXPERIMENT

We first compute the KL-divergence of current book and the most recent book or the average of all the past books after converting them into k-dimensional vectors where k is the number of topics. This process was done over the 20, 40, 60, and 80 topics dataset respectively. Then we represent each book's surprise feature with a pair of text-to-text and past-to-text surprises. Fig. 2 shows the surprises for all the books using 20-topics model, each data point corresponding to one book.

We firstly use k-means to generate a constant number of model candidates, as described in the previous section. We can see the distributions of the clusters generated by k-means in Fig. 3- 6.

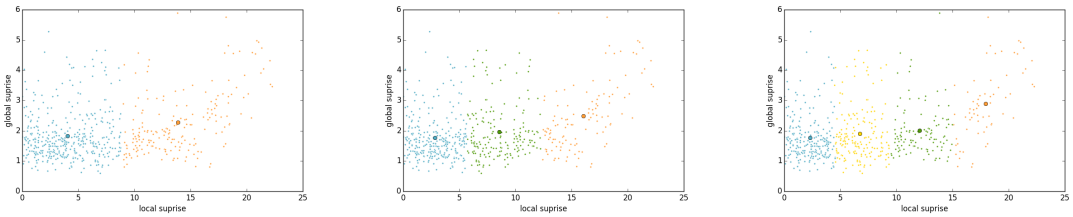
A multivariate normal distribution is learnt over each cluster. This is done simply by computing the mean and covariance matrix for these clusters. In Fig. 7- 10, the distribution



(a) Topic-20 model **(b)** Topic-40 model

(c) Topic-60 model **(d)** Topic-80 model

Figure 2: Surprise feature plots under different topic models

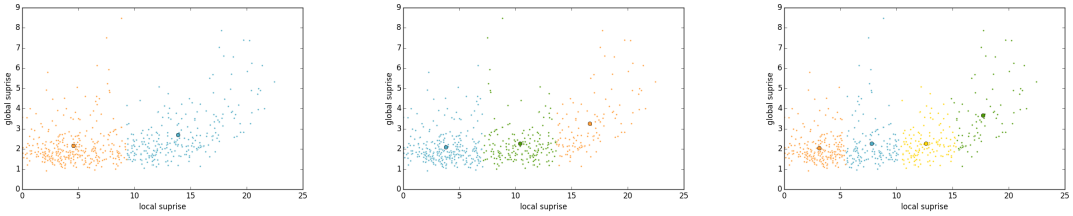


(a) cluster num = 2 **(b)** cluster num = 3 **(c)** cluster num = 4

Figure 3: K-means clustering of surprise features using topic-20 model

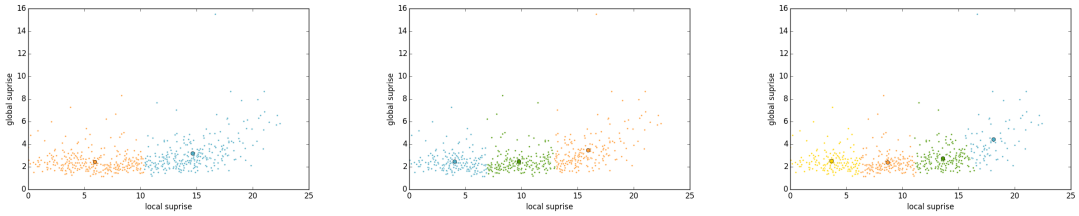
contours are overplayed on the data points and these models are used as candidates for epoch decoding.

The numerical results of epoch segmentation is collectively presented in Table 1. The first column shows which dataset we are working on and what epoch hypothesis we've made. The second column is the optimal segments to separate the sequence into epochs. The third column corresponds to the mean value of (text-to-text surprise, past-to-text surprise) model in each epoch. The last column shows the log likelihood of the data by applying different hypothesized models. Almost all the tests have shown the same pattern in surprise change: from low surprise pairs to high surprise pairs and back to low. This pattern seems to oscillate back and forth as the number of epochs increases. The topic-60 result is not consistent with others since the segmentation is not well-aligned with the results generated from other datasets. There is one epoch heavily overlapped across different datasets, which spans from



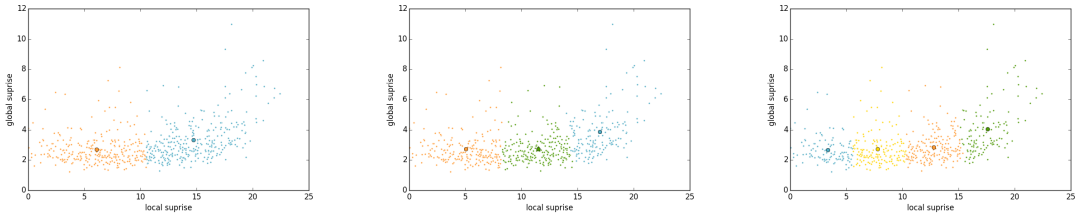
(a) cluster num = 2 **(b)** cluster num = 3 **(c)** cluster num = 4

Figure 4: K-means clustering of surprise features using topic-40 model



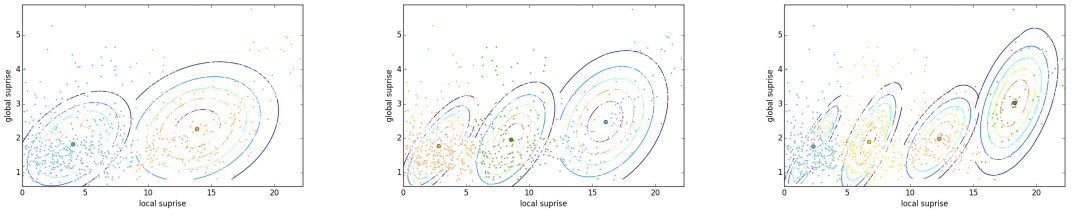
(a) cluster num = 2 **(b)** cluster num = 3 **(c)** cluster num = 4

Figure 5: K-means clustering of surprise features using topic-60 model



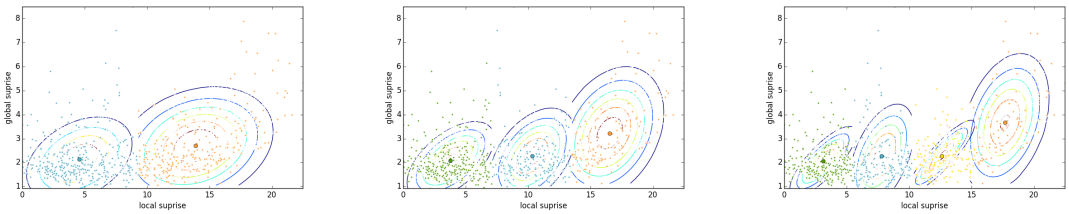
(a) cluster num = 2 **(b)** cluster num = 3 **(c)** cluster num = 4

Figure 6: K-means clustering of surprise features using topic-80 model



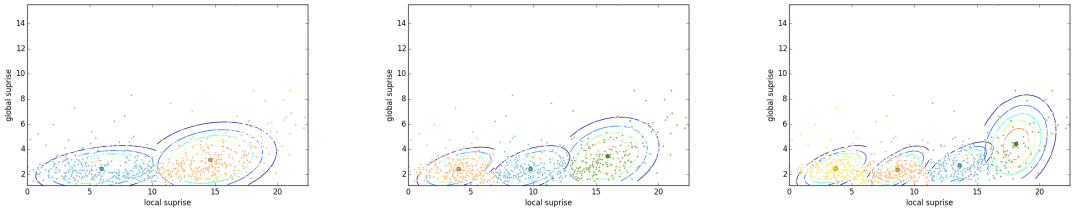
(a) cluster num = 2 **(b)** cluster num = 3 **(c)** cluster num = 4

Figure 7: Bivariate normal distribution generated on clusters over topic-20 model



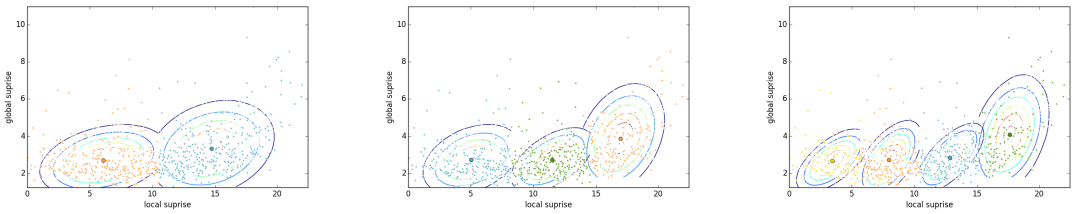
(a) cluster num = 2 **(b)** cluster num = 3 **(c)** cluster num = 4

Figure 8: Bivariate normal distribution generated on clusters over topic-40 model



(a) cluster num = 2 **(b)** cluster num = 3 **(c)** cluster num = 4

Figure 9: Bivariate normal distribution generated on clusters over topic-60 model



(a) cluster num = 2 **(b)** cluster num = 3 **(c)** cluster num = 4

Figure 10: Bivariate normal distribution generated on clusters over topic-80 model

Topics Epochs	Segments	Models	Log-Likelihood
20 & 2	1859-07-15	(4.02 , 1.83), (13.87 , 2.28)	-3813.24
20 & 3	1845-03-00, 1847-12-28	(4.02 , 1.83), (13.87 , 2.28), (4.02, 1.83)	-3601.02
20 & 4	1845-03-00, 1847-12-28, 1859-07-15	(4.02, 1.83), (13.87, 2.28), (4.02, 1.83), (13.87, 2.28)	-3576.87
40 & 2	1840-02-18	(4.53, 2.13), (13.8, 2.71)	-3970.22
40 & 3	1840-02-18, 1851-01-27	(4.57, 2.15), (13.8, 2.71), (4.57, 2.15)	-3970.12
40 & 4	1840-02-18, 1851-01-27, 1856-11-15	(4.53, 2.13), (13.8, 2.71), (4.53, 2.13), (13.8, 2.71)	-3897.37
60 & 2	1838-00-00	(14.62, 3.21),(5.92, 2.46)	-4007.32
60 & 3	1845-01-30, 1848-06-07	(5.92, 2.46), (14.62, 3.21), (5.92, 2.46)	-3896.46
60 & 4	1838-00-00, 1845-01-30, 1848-06-07	(14.62, 3.21), (5.92, 2.46), (14.62, 3.21),(5.92, 2.46)	-3827.37
80 & 2	1851-02-03	(14.73, 3.34), (6.09, 2.70)	-3906.13
80 & 3	1851-02-03, 1858-00-00	(14.73, 3.34), (6.09, 2.70), (14.73, 3.34)	-3811.82
80 & 4	1840-02-18, 1851-02-03, 1858-00-00	(6.09, 2.70), (14.73, 3.34), (6.09, 2.70), (14.73, 3.34)	-3719.57

Table 1: Epoch segmentation over different dataset, the segment dates and the model used for each epoch.

1840's to 1850's. In this period, the data shows significant increase of local surprise and moderate rise of global surprise in Darwin's reading.

DISCUSSION

In this report, we proposed an entirely unsupervised approach to segment epochs on Darwin's reading data. We applied clustering to extract local patterns and estimate the optmal sequence of patterns that interpret the data using HMM.

One question arises in the candidate model generation. Using k-means multiple times still doesn't guarantee the possibility to find the plausible candidates for epoch segmentation. The clustering is totally data-driven so it's almost blind to the temporal order of data. One possibility is to incorporate the temporal coherence into the clustering to generate more useful models that can be used. Another possibility is to initialize and direct the clustering by prior knowledge so that it's biased to produce more interpretable and informative models.

In the segmentation model, we only looked for best match in terms of the log likelihood under some hard constraints, but didn't penalize the complexity of model. This might lead to overfitting to the data if an over-complicated model is used. We need to introduce a penalty term on the number of parameters, which can be separated into the inference as transition cost or an overall regularization term. As such, we would be able to compare the effectiveness of different models more fairly by looking at the complexity-penalized likelihood.

Bibliography

[1] Jaimie Murdock, Simon DeDeo, and Colin Allen. Exploration and Exploitation of Victorian Science in Darwin's Reading Notebooks, 1838-1860.