Final Project: Classifying Chinese Poems with Markov Models

Roujia Wen

Minerva Schools

CS156

Prof. Shekhar

April 20th, 2018

Problem Definition

The problem I'm trying to solve is, given a Chinese ancient poem, how can we tell which category of themes it belongs to? This is a supervised learning problem and the main goal is to train a classifier that categorize the poems accurately.

A possible concern could be that themes are very subjective and whether I can label the poems correctly in the first place. The good thing is that the themes in ancient Chinese poems are very well-defined, to the extent that they almost become cliches. For example, in a poem with a war theme, you would see common phrases and words associated with horses, the Northwest desserts, flames, smokes, and war drums. The most distinct poetic themes include war, rural life, nature, love, and so on. The online poem collection website, shicimingju.com, sorts poems by theme and time period. I selected only the poems from Tang dynasty, because poems during the same time period share the same structure and imageries, and therefore have more distinct characteristics which makes it easier to classify. Tang dynasty was one of the most active period in Chinese history in artistic creations and the poems have very regular structures (either five or seven characters per line, and four lines per poem). In addition, I chose five major theme categories, presented in the table below.

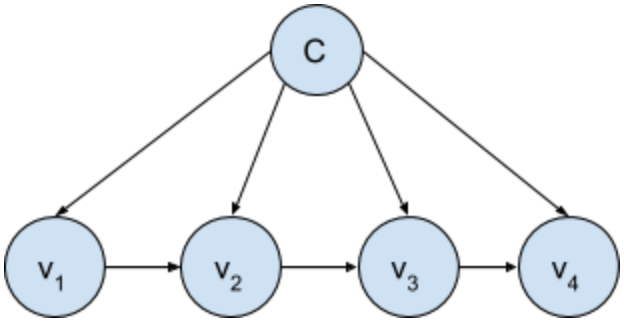
Category	**Translation**	**Number of Poems**
songbie	farewell	37
tianyuan	rural life	23
shanshui	natural scenery	53
zhanzheng	war	54
aiqing	love	41

Table 1 - Poem themes, translations and data size

Solution Specification

Poems are discrete and sequential data. If we consider each character as an observation, the entries in the joint probability distribution will grow exponentially as the number of observations increases. Therefore, Markov models can be a good way to model the poem generation process, conveniently parametrizing it into lower dimension. In this project, I'm making a few assumptions to simplify the problem. First, we assume that the probability of observing a character depends only on its previous character. We also assume that this transition probability is time-independent, or, in this context, that it does not depend on the location in the poem in which the transition takes place. Under these assumptions, we can use a discrete state first order stationary Markov chain to model poem generation.

An additional assumption made here is that poems within one category are generated using the same underlying probability distribution. Therefore, in this model we have one transition matrix for each poem category. The model can be summarized in the figure below:



Where c is a variable that represents the category in which a poem comes from, and $v_1, v_2, ...v_n$ are the observations (characters) of a poem.

Training

I trained one Markov chain for each category separately, using maximum likelihood. This is done by counting the number of transitions in the sequence and normalize so that probabilities sum to one.

Predicting

Assuming a uniform prior probability of a poem being in a category, we have

$$p(c|v_{1:T}) \propto p(c)p(v_{1:T}|c) \propto p(v_{1:T}|c) = \prod_{t=1}^{T} p(v_t|v_{t-1}, c)$$

Where c is a variable that represents the category and $v_{1:T}$ represents a given sequence of observations.

Therefore, when predicting which category a poem belongs to, we calculate $p(v_{1:T}|c)$ [1] (taking the log in actual implementation to avoid numerical issues) for each category and compare the magnitude. The category with largest value is the predicted category.

Testing and Analysis

Since the total number of data is quite small, I decided to use LOOCV as the validation method. In the first trial in which I trained the classifier with uneven numbers of data points in different categories, I obtained the following test accuracies:

songbie	tianyuan	shanshui	zhanzhen	aiqing	**overall**
30.6%	40%	57.2%	89.8%	41.2%	**51.76%**

Table 2 - Prediction accuracies by category, and overall accuracy (trained with unbalanced data)

[1] For unknown transitions, log probability is taken to be -8 (which is slightly smaller than the lowest nonzero entry in the transition matrix)

		True Class					
		songbie	tianyuan	shanshui	zhanzhen	aiqing	
Out-put	songbie	6.12%	0.74%	1.74%	1.3%	1.12%	55.54%
	tianyuan	0.04%	8%	0.64%	0.04%	0.04%	91.32%
	shanshui	6.44%	4.72%	11.44%	0.42%	5.52%	40.08%
	zhanzhen	5.58%	6.48%	2.4%	17.96%	5.08%	47.89%
	aiqing	1.82%	0.06%	3.78%	0.28%	8.24%	58.11%
		30.6%	40%	57.2%	89.8%	41.2%	51.76%

Table 3 - Confusion matrix (trained with unbalanced data)

Since the results seem to correlates with the number of data points by category, I was concerned that the large variation in prediction accuracies across categories was affected by the difference in data size. Therefore, I also tried to train the classifier with the same number of data points across category (n=22 per category), and obtained the following accuracies:

songbie	tianyuan	shanshui	zhanzhen	aiqing	**overall**
30.3%	57.2%	37.4%	85.1%	45.3%	**51.06%**

Table 4 - Prediction accuracies by category, and overall accuracy (trained with balanced data)

		True Class					
		songbie	tianyuan	shanshui	zhanzhen	aiqing	
Out-put	songbie	6.06%	1.28%	3.84%	0.62%	0.84%	47.94%
	tianyuan	1.5%	11.44%	2.54%	0.26%	2.72%	61.97%
	shanshui	5.66%	3.02%	7.48%	0.84%	4.12%	35.42%
	zhanzhen	4.52%	3.06%	2.24%	17.02%	3.26%	56.54%
	aiqing	2.26%	1.2%	3.9%	1.26%	9.06%	51.24%
		30.3%	57.2%	37.4%	85.1%	45.3%	51.06%

Table 5 - Confusion matrix (trained with balanced data)

We can see that the accuracies for shanshui and zhanzhen decreased while those for the others improved. The category zhanzhen still has a dominating accuracy comparing to the others. This is an interesting result and my hypothesis is that zhanzhen (war)-themed poems have more distinct words that make it easier to classify. This, however, does not explain everything since the classifier in general favors the zhanzhen category (30.1% of all test cases were predicted to be zhanzhen).

Limitations

There are a few limitations I would like to discuss about this study. First, the small data size comparing to the large size of alphabet[2] can be disadvantageous for fitting Markov models. The transition matrix is very sparse and there might not be a lot of information we can use when predicting the class of an unseen poem.

In addition, many assumptions in this project can be an oversimplification. For example, the assumption of first-order is naive and does not capture the complexity of poem writing. However, we see that the model did capture some important information, since we were able to achieve a prediction success rate of around 51% (a random guess would only be 20%).

[2] The data contains 2170 unique characters.

Appendix

I also used the trained Markov models to generate fake poems. I'll present the best poem[3] from each category.

songbie (farewell)	随风尘飞鸟， 嗟君迟楼船， 忽闻游子唱， 月向人归雁。	With wind and dust birds fly away, You Jun is late for the building boat, Suddenly hearing wanderer's singing, The moon faces people and returning wild-goose.
tianyuan (rural life)	童未已暮锄，宫帘挂玉弓， 蒸藜炊黍归，苍苍横翠微。	The children has not hoed in the evening, The palace curtain hangs the jade bow, Steam quinoa, cook millet and come back, Pale green horizontal trees are small.
shanshui (natural scenary)	孤舟子知不，荣光秋日如， 天门山鸟飞，巫山喷雪山， 正西秦地即，薄暮禽相邀， 汝意春看棹，竹溪三吴山， 秋霜岁欲渡，了语声晚来。	Lonely boat, and you do not know, Autumn day is like glorious light, In Tianmen Mountain birds fly, In Wu Mountain snow mountains are bursting, In the West the properties of Qin, In the thin dusk fowls are inviting, You think about spring and look at the table, Bamboo river and three Wu Mountains, In autumn frost the year wants to pass, Known voices come late.
zhanzheng (war)	沧海动江城，公家乡为新， 农死者问行，泽国自从贵， 雷鼓动摇峦，垢腻脚不息， 绝笔于获再，送徒振原秋， 匈奴不得牛，一引龙跃鳞， 黠虏坐金鞍，男儿日程期， 血洗兵五年，从军雄剑四， 蓬茸菊垂今，翻思不见明。	The deep blue sea is shaking the river city, The public's hometown is new, Farmers die and ask for directions, Blessed countries themselves are expensive, Thunder and drums shake the mountain range, Dirts are greasy and feet do not rest, Stop writing and gain again, Send on foot and shake the original autumn, Huns cannot acquire cows, Once triggered, the dragon jumps its scales, Clever enemies sit on golden saddles,

[3] These poems were selected by my mom.

		True man expects his schedule, Blood-washed soldiers for five years, Joining the army with four heroic swords, Radiant chrysanthemum bequeath until today, Think back and forth, cannot see the light.
aiqing (love)	一弦写恨，没人兮为， 更脱红蜡，能窥帘下， 潭清池苑，终堪恨王， 借问江清，芙蓉塘外。	A string of writing hate, No one is guilty, Pulling off the red waxm Can look under the curtain, Tan Qing ponds and gardens, Finally hate the king, Please ask the clear river, Outside the hibiscus pond.