CS 181 LECTURE 3/26/24

Scribe Notes

Contents

1 Recap from last time

We talked about PCA, which finds a transformation of our data $\tilde{x} = Uz = U(U^\top x)$. This is a linear encoding method. We can understand this as an encoding and decoding process, where $U^\top x$ is the encoding and $U(U^\top x)$ is the decoding. PCA is non-probablistic and today we will look at supervised probablistic methods. The shared feature between today's content and last class' content is the generative model (there are latent z's that generate the x's).

2 Factor Analysis

We have the following data generation process.

We draw $z_n \sim N(0, I_K)$, so z_n is a K-dimensional vector. Then we draw $A \sim N(0, I_{D \text{ x } K})$, so A is a D x K matrix that is shared across all datapoints.

Finally, we draw $x_n \sim N(Az_n, \sigma^2 I_D)$. So, x_n is a D-dimensional vector.

With this, we can write the likelihood. So our objective is:

$$p(x|z, A) = N(x; Az, \sigma^2 I)$$
$$p(x|A) = \int N(x; Az, \sigma^2 I) N(z; 0, I_K) dz$$

The E-step: is computing $p(z|A)$. The M-step: is computing the maximum of the likelihood with respect to A.

3 Topic Models

Consider the scenario where you have text data on diseases and patients. And each disease produces a signature. For instance, for a cold, the doctor may write "sore throat, blood test, vitamins, pain medications", or for heard disease, the doctor writes "blood test, ECG, hypertension," or for broken leg, the doctor writes "pain medication, cast, physical therapy." There is a mixed membership of what treatments are given for each disease.

Let us say that we observe a patient with a record that says "cast, vitamins, and pain medication." Given a patient with that profile, we can ask which disease is responsible for the profile. There could be more than one condition or none out of the conditions we have seen and a different condition all together. In topic modeling, cold, heart, and broken leg would be our topics, and the patient profile would be the data that we are observing.

So, in topic modeling, we model data as admixtures of "topics". Admixtures means that we can get multiple underlying categories - like multiple diseases, in this case.

To formalize the setting, we will have

- vocabulary - size D, meaning there are D unique words.
- N documents, x_n, and assume for simplicity they all have the same length L, so there are L tokens in each document
- K topics

- Parameters: Θ_{dk} - the proportion of word d in topic k, π_{nk} - the proportion of topic k in document n

3.1 Data Generation Process

For each document x_n,

$$\pi_n \sim \text{Dirchlet}(\alpha)$$

$$x_n \sim \text{Multinomial}(\theta\pi_n, L)$$

So, we are first drawing π_n which is the mixture of topics in the document x_n, then we draw the L tokens in the document x_n.

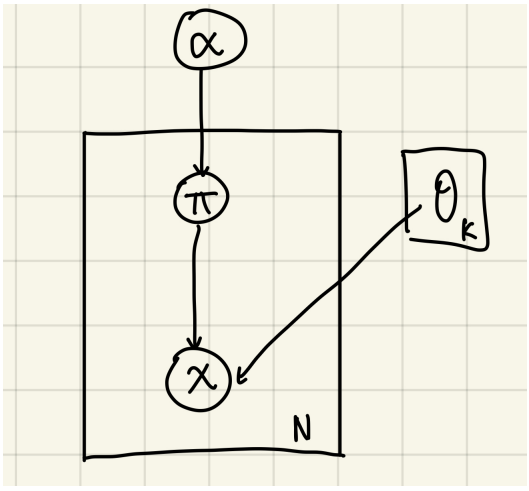

Figure 1: Topic model data generation.

$\Theta\pi_n$ gives us the probability of each word occuring in the document.

3.1.1 What is a Dirichlet?

If you haven't seen the Dirichlet before, it is a distribution over an n-dimensional vector whose components sum to 1. For example, a sample from a dirichlet distribution in 3-dimensions could produce a sample that is the vector

$$\begin{bmatrix} 0.2 \\ 0.5 \\ 0.3 \end{bmatrix}$$

.

A Dirichlet is a multivariate version of beta. If we want a uniform distribution, we use $\alpha > 1$. If we want a distribution that has concentrations at the corners, then we use $\alpha < 1$. This is a useful prior because it allows us to model distributions with different properties.

3.2 Applications of Topic Models

"Reconstructing Pompeian Households" by Mimmo '12. The authors took archeaological data and found a map of Pompei. There is a big table of different Roman objects in the

city. Then the authors build a topic model, where the topics are the functional groups of objects. The documents are the rooms and the words are the objects.

Another big use of topic models is social media analysis. In addition to the topic, there is another important variable that we must model: the creator of the post. The person creating the posts has a strong impact on what types of topics are posted about.

3.3 Back to the Math

$$p(x|\pi,\theta) \propto \prod_{d\in[D]} (\sum_k \theta_{dk}\pi_k)^{x_d}$$
$$p(\pi) \propto \prod_{k\in[K]} \pi_k^{\alpha_k - 1}$$
$$p(x,\pi|\theta) = (\prod_{k\in[K]} \pi_k^{\alpha_k-1})(\prod_{d\in[D]} (\sum_k \theta_{dk}\pi_k)^{x_d})$$

Since taking the log of this expression is challenging, we will introduce a new variable, z, which is the topic of each word.

For each token l, we do two things, we sample the topic $z_{nl} \sim Cat(\pi_n)$. So, z_{nl} is some number between 0 and k. Then $w_{nl} \sim Cat(\theta_{z_{nl}})$.

For example, we could sample "cold" for z_{nl} and then we could sample "cough" for the word.

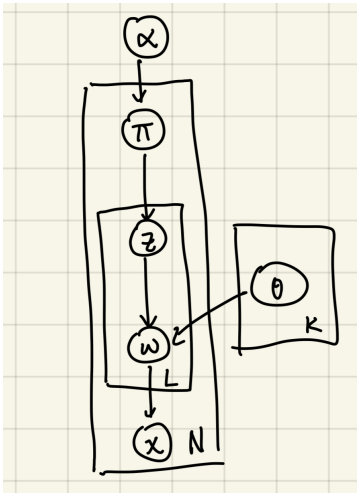

Figure 2: Modified data generation model.

3.4 Why?

Now we can simplify our likelihood more easily.

$$p(x_n|z,\pi,\theta) \propto \prod_L \prod_D \prod_K \theta_{dk}^{\mathbf{I}(w_{nl}=d)\mathbf{I}(z_{nl}=k)}$$
$$p(z|\pi) \propto \prod_K \pi^{\mathbf{I}(z=k)}$$
$$p(\pi) \propto \prod_K^{\alpha_k-1} \pi_k^{\alpha_k-1}$$

So,

$$p(x,z,\pi|\theta) = \prod_n [\prod_k \pi_{kn}^{\alpha_k-1}][\prod_{l,k} \pi^{\mathbf{I}(z_{nl}=k)}][\prod_{l,k,d} \theta_{dk}^{\mathbf{I}(w_{nl}=d)\mathbf{I}(z_{nl}=k)}]$$

$$\log p(x,z,\pi|\theta) = \sum_n \left[\sum_k (\alpha_k - 1)\log \pi_{kn}\right] + \sum_{n,l,k} \mathbf{I}(z_{nl}=k)\log\pi + \sum_{n,l,k,d} \mathbf{I}(w_{nl}=d)\mathbf{I}(z_{nl}=k)\log\theta_{dk}$$

E-Step: We take the expectation over z, and we introduce new variable q_{nlk} which models $p(z_{nl}=k|\pi,w,\theta)$.

$$\mathbf{E}_z[\log p(x,z,\pi|\theta)] = \sum_n \sum_k (\alpha_k - 1 + \sum_l q_{nlk})\log\pi_{nk} + \sum_d \sum_k (\sum_l q_{nlk}\mathbf{I}(w_{nl}=d)\log\theta_{dk})$$

M-step: We maximize with respect to π and θ.

$$\max_{\pi_{nk}} \propto (\alpha_k - 1) + \sum_l q_{nlk}$$

$$\max_{\theta_{kd}} \propto \sum_n \sum_l q_{nlk}\mathbf{I}(w_{nl}=d)$$

4 Concept Check

Suppose we have 2 dirichlets for $K=3$.

$\alpha_1 = [10,10,10]$

$\alpha_2 = [.01,.01,.01]$

1) Which came from α_1, α_2:

A) [1/3, 1/3, 1/3], [1/3, 1/3, 1/3] from α_1

B) [1/2, 1/2, 0], [2/5, 3/5, 0]

C) [0, 1, 0], [1, 0, 0] from α_2

D) [1/4, 1/2, 1/4], [3/5, 1/5, 2/5]

2) You see data $\{0,0,1,0\}$ What are the posteriors for α_1 and α_2?

3) Are the draws sparse after this update?

5 Concept Check Answer

1)

A) [1/3, 1/3, 1/3], [1/3, 1/3, 1/3] is from α_1

C) [0, 1, 0], [1, 0, 0] is from α_2

2) To get the posterior, you take the prior and you add the counts.

$\alpha_1 = [13, 11, 10]$ $\alpha_2 = [3.01, 1.01, .01]$

There are 3 occurrences of 0 in the data, so we add 3 to the first values in α_1 and α_2. There is 1 occurrence of 1 and 0 occurrences of 2.

3) No.