Homework 5: Mixtures, EM, and Graphical Models

This homework assignment will have you work with mixtures, EM, and graphical models.

Please type your solutions after the corresponding problems using this LaTeX template, and start each problem on a new page.

Please submit the **writeup PDF to the Gradescope assignment 'HW5'**. Remember to assign pages for each question.

Please submit your **LaTeX file and code files to the Gradescope assignment 'HW5 - Supplemental'**.

You can use a **maximum of 2 late days** on this assignment. Late days will be counted based on the latest of your submissions.

Problem 1 (Expectation-Maximization for Categorical-Geometric Mixture Models, 25pts)

In this problem we will explore expectation-maximization for a Categorical-Geometric Mixture model. Each observation $\mathbf{x}_n$ is a positive integer scalar drawn from a geometric distribution (associated with the number of trials needed to get to the first success, if success occurs with probability p). We posit that each observation comes from *one* mixture component. For this problem, we will assume there are K components. Each component $k \in \{1, \dots, K\}$ will be associated with a probability $p_k \in [0, 1]$. Finally let the (unknown) overall mixing proportion of the components be $\boldsymbol{\theta} \in [0, 1]^K$, where $\sum_{k=1}^{K} \boldsymbol{\theta}_k = 1$.

Our generative model is that each of the N observations comes from a single component. We encode observation n's component-assignment as a one-hot vector $\mathbf{z}_n \in \{0, 1\}^K$ over components. This one-hot vector is drawn from $\boldsymbol{\theta}$; then, $\mathbf{x}_n$ is drawn from Geometric(p_k) where $\mathbf{x}_n$ belongs to class k.

Formally, data are generated in two steps (assuming $\mathbf{z}_n$ encodes the class k), where we define the PMF of the geometric distribution to be $p(x_n|p_k) = (1 - p_k)^{x_n - 1} p_k$:

$$\begin{aligned} \mathbf{z}_n &\sim \text{Categorical}(\boldsymbol{\theta}) \\ \mathbf{x}_n &\sim \text{Geometric}(p_k) \end{aligned}$$

1. **Intractability of the Data Likelihood** We are generally interested in finding a set of parameters p_k that maximize the data likelihood $\log p(\{\mathbf{x}_n\}_{n=1}^N | \{p_k\}_{k=1}^K)$. Expand the data likelihood to include the necessary sums over observations $\mathbf{x}_n$ and latents $\mathbf{z}_n$. Why is optimizing this loss directly intractable?

2. **Complete-Data Log Likelihood** Define the complete data for this problem to be $D = \{(\mathbf{x}_n, \mathbf{z}_n)\}_{n=1}^N$. Write out the complete-data negative log likelihood. Note that optimizing this loss is now computationally tractable if we know $\mathbf{z}_n$.

$$\mathcal{L}(\boldsymbol{\theta}, \{p_k\}_{k=1}^K) = -\ln p(D \,|\, \boldsymbol{\theta}, \{p_k\}_{k=1}^K).$$

3. **Expectation Step** Our next step is to introduce a mathematical expression for $\mathbf{q}_n$, the posterior over the hidden topic variables $\mathbf{z}_n$ conditioned on the observed data $\mathbf{x}_n$ with fixed parameters, i.e $p(\mathbf{z}_n | \mathbf{x}_n; \boldsymbol{\theta}, \{p_k\}_{k=1}^K)$.
   - **Part 3.A** Write down and simplify the expression for $\mathbf{q}_n$.
   - **Part 3.B** Give an algorithm for calculating the expression for $\mathbf{q}_n$ found in Part 3.A for all n, given the observed data $\{\mathbf{x}_n\}_{n=1}^N$ and settings of the parameters $\boldsymbol{\theta}$ and $\{p_k\}_{k=1}^K$.

4. **Maximization Step** Using the $\mathbf{q}_n$ estimates from the Expectation Step, derive an update for maximizing the expected complete data log likelihood in terms of $\boldsymbol{\theta}$ and $\{p_k\}_{k=1}^K$.
   - **Part 4.A** Derive an expression for the expected complete-data log likelihood using $\mathbf{q}_n$.
   - **Part 4.B** Find an expression for $\boldsymbol{\theta}$ that maximizes this expected complete-data log likelihood. You may find it helpful to use Lagrange multipliers in order to enforce the constraint $\sum \theta_k = 1$. Why does this optimized $\boldsymbol{\theta}$ make intuitive sense?
   - **Part 4.C** Apply a similar argument to find the values of $\{p_k\}_{k=1}^K$ that maximizes the expected complete-data log likelihood.

5. Suppose that this had been a classification problem. That is, you were provided the "true" categories $\mathbf{z}_n$ for each observation $\mathbf{x}_n$, and you were going to perform the classification by inverting the provided generative model (i.e. now you're predicting z given x). Could you reuse any of your inference derivations above?

6. Finally, implement your solution (see `T5_P1.py` for starter code). You are responsible for implementing the `loglikelihood`, `e_step` and `m_step` functions. Test it out with data given 10 samples from 3 components with $p_1 = .1$, $p_2 = .5$, and $p_3 = .9$. How does it perform? What if you increase the number of samples to 1000 from each of the components? What if you change $p_2 = .2$? Hypothesize reasons for the differences in performance when you make these changes. You may need to record five to ten trials (random restarts) in order to observe meaningful insights.

Solution

Problem 2 (PCA, 15 pts)

For this problem you will implement PCA from scratch. Using `numpy` to call SVDs is fine, but don't use a third-party machine learning implementation like `scikit-learn`.

We return to the MNIST data set from T4. You have been given representations of 6000 MNIST images, each of which are 28×28 greyscale handwritten digits. Your job is to apply PCA on MNIST, and discuss what kinds of structure is found.

As before, the given code in `T5_P3.py` loads the images into your environment as a 6000x28x28 array.

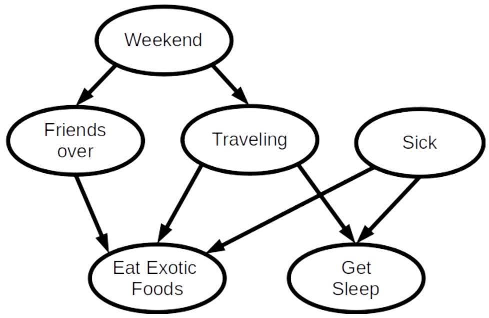
1. Compute the PCA. Plot the eigenvalues corresponding to the most significant 500 components in order from most significant to least. Make another plot that describes the cumulative proportion of variance explained by the first k most significant components for values of k from 1 through 500. How much variance is explained by the first 500 components? Describe how the cumulative proportion of variance explained changes with k.
2. Plot the mean image as well as the images corresponding to the first 10 principle components. How do the images compare to the cluster centers from K-means? Discuss any similarities and differences.
3. Compute the reconstruction error on the data set using the mean image. Then compute the reconstruction error using the first 10 principal components. How do these errors compare to the final objective loss achieved by using K-means on the dataset? Discuss any similarities and differences.

Include your plots in your PDF. There may be several plots for this problem, so feel free to take up multiple pages.

Solution

Problem 3 (Bayesian Networks, 10 pts)

In this problem we explore the conditional independence properties of a Bayesian Network. Consider the following Bayesian network representing a fictious person's activities. Each random variable is binary (true/false).



The random variables are:

- Weekend: Is it the weekend?
- Friends over: Does the person have friends over?
- Traveling: Is the person traveling?
- Sick: Is the person sick?
- Eat exotic foods: Is the person eating exotic foods?
- Get Sleep: Is the person getting sleep?

For the following questions, $A \perp B$ means that events A and B are independent and $A \perp B|C$ means that events A and B are independent conditioned on C.

Use the concept of d-separation to answer the questions and show your work (i.e., state what the blocking path(s) is/are and what nodes block the path; or explain why each path is not blocked).

Example: Is Friends over $\perp$ Traveling? If NO, give intuition for why.

Answer: NO. The path from Friends over – Weekend – Traveling is not blocked following the d-separation rules. Thus, the two are not independent. Intuitively, this makes sense as if say we knew that the person was traveling, it would make it more likely to be the weekend. This would then make it more likely for the person to have friends over.

1. Is Sick $\perp$ Weekend? If NO, give intuition for why.
2. Is Sick $\perp$ Friends over | Eat exotic foods? If NO, give intuition for why.
3. Is Friends over $\perp$ Get Sleep? If NO, give intuition for why.
4. Is Friends over $\perp$ Get Sleep | Traveling? If NO, give intuition for why.
5. Suppose the person stops traveling in ways that affect their sleep patterns (as various famous people have done). Travel still affects whether they eat exotic foods. Draw the modified network.
6. For this modified network, is Friends over $\perp$ Get Sleep? If NO, give an intuition why. If YES, describe what observations (if any) would cause them to no longer be independent.

Solution

Name

Collaborators and Resources

Whom did you work with, and did you use any resources beyond cs181-textbook and your notes?

Calibration

Approximately how long did this homework take you to complete (in hours)?