CS 181: Machine Learning April 21, 2023

Midterm II Review Questions

George Cai (georgecai@college.harvard.edu) *Matthew Qu (matthewqu@college.harvard.edu)*

1. (PCA on transformed data[1])

Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be a data matrix, with the j^{th} row corresponding to the j^{th} observation $\mathbf{x}_j^\top \in \mathbb{R}^d$. Assume $\mathbf{X}$ is centered, and suppose that the PCA of this data has principal components (eigenvectors) $\mathbf{v}_1, \ldots, \mathbf{v}_d$ with associated variances (eigenvalues) $\lambda_1 \geq \ldots, \geq \lambda_d \geq 0$. Now, let $\mathbf{Q}$ be a $d \times d$ orthonormal matrix and let $\mathbf{y}_j = \mathbf{Q}\mathbf{x}_j$ for all $j = 1, \ldots, n$.

(a) In words, briefly explain how the matrix $\mathbf{Q}$ transforms the data $\mathbf{x}_j$. Then, find an expression for $\mathbf{Y} \in \mathbb{R}^{n \times d}$, the new data matrix where the j^{th} row corresponds to $\mathbf{y}_j^\top$, in terms of $\mathbf{X}$ and $\mathbf{Q}$.

(b) Show that the PCA of $\mathbf{Y}$ yields principal components $\mathbf{Q}\mathbf{v}_1, \ldots, \mathbf{Q}\mathbf{v}_d$ with associated variances $\lambda_1, \ldots, \lambda_d$. Briefly explain why this result is intuitive.

(c) Now, suppose that $\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^\top$ is the SVD of $\mathbf{X}$. Noting that the columns of $\mathbf{V}$ contain the eigenvectors of the covariance matrix of $\mathbf{X}$, find an equation that relates the singular values of $\mathbf{S}$, s_j, with the PCA variances λ_j (assuming that the diagonal values of $\mathbf{S}$ are sorted in decreasing order).

(d) *With a graphical explanation*, show how performing PCA on non-centered data can yield incorrect principal components and variances.

[1] Adapted from STAT 185, Fall 2022

2. (**Expectation maximization on multinomial data**[2])

Recall that the multinomial distribution is a generalization of the binomial distribution where there are $k \geq 3$ possible categories. If $\mathbf{x} = (x_1, \ldots, x_k) \sim \text{Mult}(n, \boldsymbol{\pi})$, its PMF is given by

$$p(\mathbf{x}) = \frac{n!}{x_1! \cdots x_k!} \pi_1^{x_1} \ldots \pi_k^{x_k},$$

where $x_j \geq 0$ for all $j = 1, \ldots, k$, $\sum_{j=1}^k x_j = n$, and $\sum_{j=1}^k \pi_j = 1$. Suppose we have a single observation $\mathbf{x} = (x_1, x_2, x_3, x_4)$ from a $\text{Mult}(n, \boldsymbol{\pi}_\theta)$ distribution, where

$$\boldsymbol{\pi}_\theta = \left(\frac{1}{2} + \frac{1}{4}\theta, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{1}{4}\theta \right).$$

However, assume that the complete data is given by $\mathbf{z} = (z_0, z_1, x_2, x_3, x_4) \sim \text{Mult}(n, \boldsymbol{\pi}_\theta^*)$, where

$$\boldsymbol{\pi}_\theta^* = \left(\frac{1}{2}, \frac{1}{4}\theta, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{1}{4}\theta \right).$$

That is, we have the latent variables z_0 and z_1, but we only observe $x_1 = z_0 + z_1$.

(a) Write out both the observed data log-likelihood, $\ell(\theta; \mathbf{x})$, and the complete data log-likelihood, $\ell(\theta; \mathbf{x}, z_1)$, up to an additive constant with respect to θ. Why does the complete data log-likelihood not depend on z_0?

(b) What is the conditional distribution $p(z_1 \mid \mathbf{x}, \theta)$? Briefly justify your answer.

(c) Let θ^t denote the current value of θ at the t^{th} iteration of the EM algorithm. Write out the ELBO function $\text{ELBO}(\theta \mid q, \theta^t)$, where $q(z_1)$ is the posterior distribution $p(z_1 \mid \mathbf{x}, \theta^t)$. You may leave your answer in terms of named distributions (i.e., you do not need to plug in the PMF/PDFs of any distributions in your answer).

(d) Now, write out the expression which must be maximized in the M-step of the EM algorithm. Simplify as much as possible: your answer should be in the form $\underset{\theta}{\operatorname{argmax}}\, g(\theta)$, where $g(\theta)$ contains no expectations or other terms not dependent on θ.

(e) Finally, compute θ^{t+1} in terms of θ^t and $\mathbf{x}$ by finding the maximum value of the expression you derived in part (d).

[2] Adapted from `http://www.columbia.edu/~mh2078/MachineLearningORFE/EM_Algorithm.pdf`

3. (Markov decision process of a caterpillar[3])

George the very hungry caterpillar loves to eat. Because George wants to grow up and become a butterfly, he is trying to eat as many calories as possible. At every meal, he decides between eating watermelon and strawberry ice cream. Eating watermelon gives him 4 calories, while eating ice cream gives him 10 calories. However, eating too much ice cream may cause George to become sick, and eating ice cream while sick may cause him to die! (This would be bad because then George can no longer eat.) On the other hand, eating watermelon will keep George healthy and make him healthy if he is sick. George will always be in one of these three states: healthy, sick, or dead—the transitions are given in the table below.

Health condition	Watermelon or Ice Cream?	Next condition	Probability
healthy	watermelon	healthy	1
healthy	ice cream	healthy	1/4
healthy	ice cream	sick	3/4
sick	watermelon	healthy	1/4
sick	watermelon	sick	3/4
sick	ice cream	sick	7/8
sick	ice cream	dead	1/8

(a) Model this problem as an MDP by specifying the states $\mathcal{S}$, actions $\mathcal{A}$, transition functions $T^a(s, s') = P(s' \mid s, a)$, and reward function $R(a)$. Note that in this context, the reward function does not depend on the current state s or subsequent state s'.

(b) Run **value iteration** for 2 iterations on this MDP with $\gamma = 0.8$, specifying how the functions $Q_t(s, a)$ and $V_t(s)$ change over each iteration. That is, start with $Q_0 = 0, V_0 = 0$ and compute $Q_1(s, a), V_1(s), Q_2(s, a)$, and $V_2(s)$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$.

(c) Let π_0 be the policy that George always eats ice cream regardless of his health condition. Find the value of this policy, $V^{\pi_0}(s)$, for each state $s \in \mathcal{S}$ in terms of the discount factor γ. Then, for each state-action pair (s, a), find $Q^{\pi_0}(s, a)$ in terms of V^{π_0}.

(d) Now, run **policy iteration**, starting with the initial policy π_0 where George always eats ice cream. Setting $\gamma = 0.8$, use your results in part (c) to find the optimal policy. How many iterations does your algorithm take?

(e) Run policy iteration again, but now with $\gamma = 0.9$. How many iterations does it take for your policy to converge? If your optimal policy differs that of part (d), briefly describe your intuition for why this might occur.

(f) (**Bonus**) Does there exist a reward function $R'(a)$ with $R'(\text{watermelon}) < R'(\text{ice cream})$ and some $\gamma \in [0, 1)$ such that the optimal policy π^* satisfies $\pi^*(\text{healthy}) = \text{watermelon}$? If so, find R' and γ, or prove that such an R' cannot exist.

[3] Adapted from CS 182, Fall 2020

4. (**Everything is a graphical model**)

For the following models[4], write the following: (i) graphical model, (ii) "generative story" of the model, (iii) observed data log likelihood, (iv) ELBO function, (v) E-step (what are you maximizing, with respect to what?), (vi) M-step (what are you maximizing, with respect to what?).

(a) pPCA: latent variables $\mathbf{z}_n \sim \mathcal{N}(0, I)$, $\mathbf{z} \in \mathbb{R}^k$, observed variables $\mathbf{x}_n | \mathbf{z}_n \sim \mathcal{N}(\mathbf{W}\mathbf{z}_n, \sigma^2\mathbf{I})$, $\mathbf{x} \in \mathbb{R}^d$. You may find the multivariate Gaussian $\mathbf{r} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\mathbf{r} \in \mathbb{R}^d$ PDF helpful:

$$p(\mathbf{r}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-d/2}(\det \boldsymbol{\Sigma})^{-1/2} \exp - \left(\frac{(\mathbf{r} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{r} - \boldsymbol{\mu})}{2} \right)$$

(b) Hidden Markov Model: latent variables $(\mathbf{s}_1, \ldots, \mathbf{s}_n), \mathbf{s}_1 \sim \text{Cat}(\boldsymbol{\theta})$, $\mathbf{s}_t \in [0, 1]^K$ (i.e. the s_t are one-hot encoded vectors), $\boldsymbol{\theta} \in \mathbb{R}^K$, observed variables $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$, $\mathbf{x}_t | \mathbf{s}_t \sim p(\mathbf{x}_t | \mathbf{s}_t)$ (this is arbitrary, one example is linear Gaussian noise $\mathcal{N}(\mathbf{D}s_t + \mathbf{E}, \sigma^2\mathbf{I})$, where $\mathbf{D} \in \mathbb{R}^{K \times d}, \mathbf{E} \in \mathbb{R}^d$, and $\mathbf{I}$ is the $d \times d$ identity matrix), $\mathbf{x}_t \in \mathbb{R}^d$, latent state transition probabilities $\mathbf{T}_{ij} \in \mathbb{R}^{K \times K}$, $\mathbf{T}_{ij} = P(s_{t+1} = j | s_t = i)$.

[4] For more practice, go through the midterm skills checklist and do this process for each model listed.