CSCI 567, Summer'20
Victor Adamchik

Problem Set #4

Due: August 5, 2020 11:59 pm

Instructions

Submission: Assignment submission will be via `courses.uscden.net`. By the submission date, there will be a folder named `Written Assignment 4` set up in which you can submit your files. Please be sure to follow all directions outlined here.

You can submit multiple times, but only the last submission counts. That means if you finish some problems and want to submit something first and update later when you finish, that's fine. In fact you are encouraged to do this: that way, if you forget to finish the homework on time or something happens (remember Murphy's Law), you still get credit for whatever you have turned in.

Problem sets must be typewritten or neatly handwritten when submitted. In both cases, your submission must be a single PDF. It is strongly recommended that you typeset with LaTeX. There are many free online LaTeX editors that are convenient to use (e.g Overleaf). You can also use offline editor such as TeXShop.

Please follow the rules below while submitting:

- The file should be named as Firstname_Lastname_USCID.pdf e.g., Jeff_Dean_8675309045.pdf.
- Do not have any spaces in your file name when uploading it.
- Please include your name and USCID in the header of your report as well.

Collaboration: You may discuss with your classmates. However, you need to write your own solutions and submit separately. Also in your written report, you need to list with whom you have discussed for each problem. Please consult the syllabus for what is and is not acceptable collaboration.

Note on notation: Unless stated otherwise, scalars are denoted by small letter in normal font, vectors are denoted by small letters in bold font and matrices are denoted by capital letters in bold font.

Problem 1 EM (30 points)

In this problem, we will see how we can change our inductive bias on the distribution model for the expectation maximization (EM) clustering. In previous homework, we introduced the Laplace distribution. Let's use the Laplace Mixture Model (LMM) and, this time, we shall derive parameter estimates for the EM algorithm. To remind you the Laplace distribution has the form (for simplicity we will consider a one-dimensional case, i.e. $x \in R$):

$$\mathcal{L}(x; \mu, b) = \frac{1}{2b} \exp\left(- \frac{|x - \mu|}{b} \right)$$

HINT: You may find this example of informal derivation of the MLE GMM useful for the derivation of the parameter estimates.

1.1 Derive the estimate of the location parameter update μ_k for the EM algorithm. (15 points)

1.2 Derive the estimate of the scale parameter b_k for the EM algorithm. (10 points)

1.3 Provide steps of the LMM EM algorithm. What is the difference with the GMM EM algorithm? (5 points)

Problem 2 Alternative formulation of PCA (30 points)

In your lecture you saw Principal Component Analysis derived through maximizing the variance for the projected data. In this exercise we will look at an alternative formulation through minimizing the projection error.

For that, let us introduce a complete orthonormal set of D-dimensional vectors: $\{\mathbf{u}_i\}$, where vectors $\mathbf{u}_i, \mathbf{u}_j$ satisfy the following condition:

$$\mathbf{u}_i^T \mathbf{u}_j = \begin{cases} 1 & \text{if i=j} \\ 0 & \text{otherwise} \end{cases}$$

This basis is complete, so each data point $\mathbf{x}_n$ can be represented by a linear combination of $\{\mathbf{u}_i\}$:

$$\mathbf{x}_n = \sum_{i=1}^{D} \alpha_{ni} \mathbf{u}_i$$

The coefficients α_{ni} will be different for each point $\mathbf{x}_n$.

2.1 Show that $\alpha_{n,i} = \mathbf{x}_n^T \mathbf{u}_i$. (4 points)

Given what we just proved in Question 2.1, we can represent our datapoints as:

$$\mathbf{x}_n = \sum_{i=1}^{D} (\mathbf{x}_n{}^T \mathbf{u}_i) \mathbf{u}_i$$

We want to represent the data points with a smaller number of variables, let's denote that number $M < D$, that will correspond to the projection to a linear subspace. We use the first M vectors of $\mathbf{u}_i$, i.e. $(\mathbf{u}_1, \ldots, \mathbf{u}_M)$ as the subspace and let us denote the representation of $\mathbf{x}_n$ in a subspace spanned by the first M vectors as $\tilde{\mathbf{x}}_n$:

$$\tilde{\mathbf{x}}_n = \sum_{i=1}^{M} \mathbf{z}_{ni} \mathbf{u}_i + \sum_{i=M+1}^{D} b_i \mathbf{u}_i \tag{1}$$

Note, that in the Eq. 1, $\mathbf{z}_{ni}$ depends on $\mathbf{x}_n$ and will be different for different $\mathbf{x}_n$, while b_i are some constants and corresponds to a shifting of the subspace.

Given the above representation, we want to minimize the error between the projection and original vector. The objective therefore is:

$$J = \frac{1}{N} \sum_{n=1}^{N} \|\mathbf{x} - \tilde{\mathbf{x}}_n\|^2$$

and the minimization is w.r.t to $\mathbf{u}_i, \mathbf{z}_{ni}, b_i$.

2.2 Show that $\mathbf{z}_{ni}$ that minimizes J is $\mathbf{x}_n^T \mathbf{u}_i$ (4 points)

2.3 Show that b_i that minimizes J is $\mathbf{u}_i^T \left(\frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n \right)$ (4 points)

2.4 Using results from previous parts, show that

$$J = \sum_{i=M+1}^{D} \mathbf{u}_i^T \mathbf{S} \mathbf{u}_i$$

where $\mathbf{S}$ is the covariance matrix. If $\bar{\mathbf{x}} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n$ then

$$S = \frac{1}{N} \sum_{n=1}^{N} (\mathbf{x}_n - \bar{\mathbf{x}})^T (\mathbf{x}_n - \bar{\mathbf{x}})$$

. (10 points)

2.5 Show that $\mathbf{u}_i$, that minimize

$$J = \sum_{i=M+1}^{D} \mathbf{u}_i^T \mathbf{S} \mathbf{u}_i$$

are the eigenvectors of S corresponding to $D - M$ smallest eigenvalues. (8 points)

Hint:

- *Do not forget the orthonormality constraints on* $\mathbf{u}_i$.
- *Orthonormality conditions are useful to eliminate other variables* $\mathbf{u}_j$ *from the equations.*

Problem 3 HMM (20 points)

Let's consider a special case of a Hidden Markov Model where the state transition depends on previous two states (except for the state at second time step, i.e. it depends only on $t = 1$). In other words we are given the transition distribution $P(X_{t+1}|X_t, X_{t-1})$ of this model. On the other hand, our observation (sensor) model will depend on the current state only, i.e. we are given $P(O_t|X_t)$. These two distributions, along with the initial distribution, $P(X_0)$ completely define this modified HMM.

We will derive recursive equations for state belief updates of this model, i.e. equations for computing $P(X_{t+1}|O_{1:t+1})$ that incorporates the entire history of observations.

First, let's express the prediction step. For that:

3.1 Express prediction $P(X_{t+1}|O_{1:t})$ in terms of joint belief $P(X_{t+1}, X_t|O_{1:t})$ that will be tracked in our model.

(3 points)

3.2 Now, express the joint predicted belief $P(X_{t+1}, X_t|O_{1:t})$ in terms of known transition distribution $P(X_{t+1}|X_t, X_{t-1})$ and the prior $P(X_t, X_{t-1}|O_{1:t})$ (i.e. maintained belief over the joint distribution of states from previous time steps). (4 points)

recursion

Second, we will derive the update (correction) equation of the model. For that:

3.3 Express belief over the state X_{t+1} after incorporating new observation O_{t+1}, i.e. $P(X_{t+1}|O_{1:t+1})$ in terms of updated joint distribution $P(X_{t+1}, X_t|O_{1:t+1})$. (3 points)

3.4 Finally, express updated belief of the joint distribution $P(X_{t+1}, X_t|O_{1:t+1})$ using observation (sensor) model $P(O_{t+1}|X_{t+1})$ and the predicted joint belief $P(X_{t+1}, X_t|O_{1:t})$ computed earlier. (10 points)

Problem 4 Viterbi Algorithm (20 points)

In this problem, we want to fit DNA sequence data with a generative model. In particular, we assume that they are generated by a hidden Markov model (HMM). Let $O_{1:N}$ which denote the sequence $[O_1 O_2 \dots O_N]$ be random variables corresponding to a DNA sequence of length N, controlled by hidden states $X_{1:N} = [X_1, X_2 \dots X_N]$. Each O_n takes a value in $\{A, C, G, T\}$ and each X_n takes one of the two possible states $\{s_1, s_2\}$. This HMM has the following parameters:

$$\theta = \{\pi_i, a_i j, b_{ik}\} \text{ for } i \in \{1,2\}, j \in \{1,2\}, \text{and } k \in \{A, C, G, T\}$$

- Initial state distribution π_i for $i \in \{1, 2\}$:

$$\pi_1 = P(X_1 = s_1) = 0.6; \pi_2 = P(X_1 = s_2) = 0.4$$

- Transition probabilities $a_{ij} = P(X_{n+1} = s_j | X_n = s_i)$ for any $n \in N^+, i = \{1, 2\}$ and $j = \{1, 2\}$:

$$a_{11} = 0.7, a_{12} = 0.3, a_{21} = 0.2, a_{22} = 0.8$$

- Emission probabilities $b_{ik} = P(O_n = k | X_n = s_i)$ for any $n \in N^+, i \in \{1, 2\}$ and $k \in \{A, C, G, T\}$:

$$b_{1A} = 0.4, b_{1C} = 0.1, b_{1G} = 0.4, b_{1T} = 0.1$$

$$b_{2A} = 0.2, b_{2C} = 0.3, b_{2G} = 0.2, b_{2T} = 0.3$$

We observe a sequence $o_{1:4} = [o_1, o_2 \dots o_4]$ = [AGCT], please answer the following questions with step-by-step computations:

4.1 Compute probability of the observed sequence, i.e. compute $P(O_{1:4} = o_{1:4}; \theta)$. (5 points)

4.2 Find out the most likely explanation i.e. compute $s^*_{1:4} = [s^*_1, s^*_2 \dots s^*_4] = \arg\max_{s_{1:4}} P(X_{1:4} = s_{1:4} | O_{1:4} = o_{1:4}; \theta)$. (10 points)

4.3 Predict most likely observation for the next step i.e. compute $o^* = \arg\max_{o_5} P(O_5 = o_5 | O_{1:4} = o_{1:5}; \theta)$. (5 points)