CSCI 567, Fall 2021
Haipeng Luo

Written Assignment #5

Due: Nov 23, 2021, 11:59 pm, PT

Instructions

Submission: Assignment submission will be via `courses.uscden.net`. By the submission date, there will be a folder 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, 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. Please also follow the rules below:

- The file should be named as `firstname_lastname_USCID.pdf` (e.g., Joe_Doe_1234567890.pdf).
- Do not have any spaces in your file name when uploading it.
- Please include your name and USCID in the header of the report as well.

Total points: 40 points

Notes 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.
- $\|.\|$ means L2-norm unless specified otherwise i.e. $\|.\| = \|.\|_2$

Problem 1 Principal Component Analysis (22 points)

In the class we showed that PCA is finding the directions with the most variance. In this problem, you will show that PCA is in fact also minimizing reconstruction error in some sense.

1.1 Specifically, suppose we have a dataset $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{R}^D$ with zero mean, and we would like to compress it into a one-dimensional dataset $c_1, \ldots, c_N \in \mathbb{R}$. To reconstruct the dataset (approximately), we also keep a direction vector $\mathbf{v} \in \mathbb{R}^D$ with unit norm (i.e. $\|\mathbf{v}\|_2 = 1$) so that the reconstructed dataset is $c_1\mathbf{v}, \ldots, c_N\mathbf{v} \in \mathbb{R}^D$.

The way we find $c_1, \ldots, c_N$ and $\mathbf{v}$ is to minimize the reconstruction error in terms of the squared L2 distance, that is, we solve

$$\arg\min_{c_1,\ldots,c_N,\mathbf{v}:\|\mathbf{v}\|_2=1} \sum_{n=1}^{N} \|\mathbf{x}_n - c_n\mathbf{v}\|_2^2. \tag{1}$$

Prove that the solution of (1) is exactly the following

1) $c_n = \mathbf{x}_n^T\mathbf{v}$ for each $n = 1, \ldots, N$; **(3 points)**

2) $\mathbf{v}$ is the first principal component of the dataset (if needed, you can use what have been derived from the lecture directly). **(3 points)**

Hint: first prove 1) by fixing $\mathbf{v}$, then prove 2) using the conclusion of 1).

For any fixed $\mathbf{v}$ (with unit norm), clearly we can optimize over each c_n independently: **(1 point)**

$$\begin{aligned}
\arg\min_{c_n} \|\mathbf{x}_n - c_n\mathbf{v}\|_2^2 &= \arg\min_{c_n} \left(c_n^2\|\mathbf{v}\|_2^2 - (2\mathbf{x}_n^T\mathbf{v})c_n + \|\mathbf{x}_n\|_2^2\right) \\
&= \arg\min_{c_n} \left(c_n^2 - (2\mathbf{x}_n^T\mathbf{v})c_n\right) && (\|\mathbf{v}\|_2^2 = 1) \\
&= \arg\min_{c_n} \left(c_n - \mathbf{x}_n^T\mathbf{v}\right)^2 && \text{(1 point)} \\
&= \mathbf{x}_n^T\mathbf{v}. && \text{(1 point)}
\end{aligned}$$

Next plugging $c_n = \mathbf{x}_n^T\mathbf{v}$ back into the objective, we see that $\mathbf{v}$ is the solution of

$$\begin{aligned}
\arg\min_{\mathbf{v}:\|\mathbf{v}\|_2=1} \sum_{n=1}^{N} \|\mathbf{x}_n - \mathbf{v}\mathbf{x}_n^T\mathbf{v}\|_2^2 &= \arg\min_{\mathbf{v}:\|\mathbf{v}\|_2=1} \sum_{n=1}^{N} \left(\|\mathbf{x}_n\|_2^2 - 2(\mathbf{x}_n^T\mathbf{v})^2 + (\mathbf{x}_n^T\mathbf{v})^2\|\mathbf{v}\|_2^2\right) \\
&= \arg\min_{\mathbf{v}:\|\mathbf{v}\|_2=1} \sum_{n=1}^{N} \left(-(\mathbf{x}_n^T\mathbf{v})^2\right) && (\|\mathbf{x}_n\|_2^2 \text{ is irrelevant and } \|\mathbf{v}\|_2^2 = 1) \\
&= \arg\min_{\mathbf{v}:\|\mathbf{v}\|_2=1} -\sum_{n=1}^{N} \mathbf{v}^T\mathbf{x}_n\mathbf{x}_n^T\mathbf{v} \\
&= \arg\max_{\mathbf{v}:\|\mathbf{v}\|_2=1} \mathbf{v}^T\left(\sum_{n=1}^{N} \mathbf{x}_n\mathbf{x}_n^T\right)\mathbf{v}, && \text{(2 points)}
\end{aligned}$$

which is exactly the top eigenvector of the covariance matrix of the dataset (as discussed in the lecture), that is, the first principal component. **(1 point)**

Rubrics: The points shown above are for reference (there are of course many ways to write a correct proof). Give partial credits as appropriate.

1.2 Next, you are asked to generalize the same idea to an arbitrary compression dimension $p < D$. Specifically, we would like to compress the same zero-mean dataset into a p-dimensional dataset $\mathbf{c}_1, \ldots, \mathbf{c}_N \in \mathbb{R}^p$. To reconstruct the dataset (approximately), we also keep p orthogonal direction vectors $\mathbf{v}_1, \ldots, \mathbf{v}_p \in \mathbb{R}^D$ with unit norm. For notational convenience, we stack these vectors together as a matrix $\mathbf{V} \in \mathbb{R}^{D \times p}$ whose j-th column is $\mathbf{v}_j$.

1) Write down the reconstructed dataset using $\mathbf{c}_1, \ldots, \mathbf{c}_N$ and $\mathbf{V}$ (note: this is a set of points in $\mathbb{R}^D$). Then write down the analogue of (1), that is, the optimization problem (with variables $\mathbf{c}_1, \ldots, \mathbf{c}_N$ and $\mathbf{V}$) that minimizes the reconstruction error in terms of the squared L2 distance. Make sure to include the correct constraints in this optimization problem. **(6 points)**

   The reconstructed dataset is $\mathbf{V}\mathbf{c}_1, \ldots, \mathbf{V}\mathbf{c}_N \in \mathbb{R}^D$. Thus the optimization problem is

$$\arg\min_{\substack{\mathbf{c}_1,\ldots,\mathbf{c}_N \in \mathbb{R}^p \\ \mathbf{V} \in \mathbb{R}^{D\times p}:\ \mathbf{V}^\top \mathbf{V} = \mathbf{I}}} \sum_{n=1}^{N} \|\mathbf{x}_n - \mathbf{V}\mathbf{c}_n\|_2^2,$$

   where $\mathbf{I}$ is the $p \times p$ identity matrix.

   Rubrics:

   - Two points for the correct form of the reconstructed dataset.
   - Two points for the correct constraint of the optimization problem. It is okay to omit the dimension constraints, but the "orthonormal" constraint $\mathbf{V}^\top \mathbf{V} = \mathbf{I}$ is important to have (and there are of course different acceptable ways to express this constraint).
   - Two points for the correct objective. Do not deduct points if the mistake is only from the incorrect form of the reconstructed dataset.

2) Find the optimal solution of $\mathbf{c}_1, \ldots, \mathbf{c}_N$ while fixing $\mathbf{V}$. **(4 points)**

   Similarly, the optimization can be done independently for each $\mathbf{c}_n$: **(1 point)**

$$\begin{aligned} \arg\min_{\mathbf{c}_n} \|\mathbf{x}_n - \mathbf{V}\mathbf{c}_n\|_2^2 &= \arg\min_{\mathbf{c}_n} \left(\|\mathbf{x}_n\|_2^2 - 2\mathbf{x}_n^\top \mathbf{V}\mathbf{c}_n + \mathbf{c}_n^\top \mathbf{V}^\top \mathbf{V}\mathbf{c}_n \right) \\ &= \arg\min_{\mathbf{c}_n} \left(\mathbf{c}_n^\top \mathbf{c}_n - 2\mathbf{x}_n^\top \mathbf{V}\mathbf{c}_n \right). \quad \text{(1 point)} \end{aligned}$$

   For the last step, simply setting the gradient $2\mathbf{c}_n - 2\mathbf{V}^\top \mathbf{x}_n$ to $\mathbf{0}$ gives the solution $\mathbf{c}_n = \mathbf{V}^\top \mathbf{x}_n$. **(2 points)**

   Rubrics: Similarly to Problem 1.1, give partial credits as appropriate.

3) Plug the solution of the previous question into the optimization problem and find the optimal solution of $\mathbf{V}$. (Again, feel free to use conclusions from the lecture.) **(6 points)**

After plugging $\mathbf{c}_n = \mathbf{V}^\top \mathbf{x}_n$, the problem becomes

$$\begin{aligned}
\operatorname*{arg\,min}_{\mathbf{V}:\, \mathbf{V}^\top\mathbf{V}=\mathbf{I}} \sum_{n=1}^N \|\mathbf{x}_n - \mathbf{V}\mathbf{V}^\top \mathbf{x}_n\|_2^2 &= \operatorname*{arg\,min}_{\mathbf{V}:\, \mathbf{V}^\top\mathbf{V}=\mathbf{I}} \sum_{n=1}^N \left(\|\mathbf{x}_n\|_2^2 - 2\mathbf{x}_n^\top \mathbf{V}\mathbf{V}^\top \mathbf{x}_n + \mathbf{x}_n^\top \mathbf{V}\mathbf{V}^\top \mathbf{V}\mathbf{V}^\top \mathbf{x}_n\right) \\
&= \operatorname*{arg\,min}_{\mathbf{V}:\, \mathbf{V}^\top\mathbf{V}=\mathbf{I}} \sum_{n=1}^N \left(-2\mathbf{x}_n^\top \mathbf{V}\mathbf{V}^\top \mathbf{x}_n + \mathbf{x}_n^\top \mathbf{V}\mathbf{V}^\top \mathbf{x}_n\right) \\
&\qquad\qquad (\|\mathbf{x}_n\|_2^2 \text{ is irrelevant and } \mathbf{V}^\top\mathbf{V} = \mathbf{I}) \\
&= \operatorname*{arg\,max}_{\mathbf{V}:\, \mathbf{V}^\top\mathbf{V}=\mathbf{I}} \sum_{n=1}^N \left(\mathbf{x}_n^\top \mathbf{V}\mathbf{V}^\top \mathbf{x}_n\right) \qquad \text{(2 points)} \\
&= \operatorname*{arg\,max}_{\mathbf{V}:\, \mathbf{V}^\top\mathbf{V}=\mathbf{I}} \sum_{n=1}^N \left(\mathbf{x}_n^\top \left(\sum_{j=1}^p \mathbf{v}_j \mathbf{v}_j^\top\right) \mathbf{x}_n\right) \\
&= \operatorname*{arg\,max}_{\mathbf{V}:\, \mathbf{V}^\top\mathbf{V}=\mathbf{I}} \sum_{j=1}^p \left(\mathbf{v}_j^\top \left(\sum_{n=1}^N \mathbf{x}_n \mathbf{x}_n^\top\right) \mathbf{v}_j\right). \qquad \text{(2 points)}
\end{aligned}$$

The solution of the last problem is exactly the following: $\mathbf{v}_j$ is the j-th eigenvector of the covariance matrix $\sum_{n=1}^N \mathbf{x}_n \mathbf{x}_n^\top$ (that is, the j-th principal component of the dataset). There are many ways to argue this claim. For example, let $\sum_{i=1}^d \lambda_i \mathbf{u}_i \mathbf{u}_i^\top$ be the eigendecomposition of the covariance matrix such that $\lambda_1 \geq \cdots \geq \lambda_d$. Then the objective becomes

$$\sum_{j=1}^p \left(\mathbf{v}_j^\top \left(\sum_{i=1}^d \lambda_i \mathbf{u}_i \mathbf{u}_i^\top\right) \mathbf{v}_j\right) = \sum_{i=1}^d \lambda_i \sum_{j=1}^p \left(\mathbf{u}_i^\top \mathbf{v}_j\right)^2 = \sum_{i=1}^d \lambda_i \alpha_i. \qquad \text{(Define } \alpha_i = \textstyle\sum_{j=1}^p \left(\mathbf{u}_i^\top \mathbf{v}_j\right)^2\text{)}$$

Further note that $0 \leq \alpha_i = \mathbf{u}_i^\top \mathbf{V}\mathbf{V}^\top \mathbf{u}_i \leq \mathbf{u}_i^\top \mathbf{u}_i = 1$, and

$$\sum_{i=1}^d \alpha_i = \sum_{i=1}^d \sum_{j=1}^p \left(\mathbf{u}_i^\top \mathbf{v}_j\right)^2 = \sum_{j=1}^p \left(\mathbf{v}_j^\top \left(\sum_{i=1}^d \mathbf{u}_i \mathbf{u}_i^\top\right) \mathbf{v}_j\right) = \sum_{j=1}^p \mathbf{v}_j^\top \mathbf{v}_j = p,$$

so the maximum possible value of this objective is $\sum_{i=1}^p \lambda_i$.[1] This maximum value can be exactly attained by taking $\mathbf{v}_j = \mathbf{u}_j$. **(2 points)**

Rubrics: Similarly to Problem 1.1, give partial credits as appropriate. It is okay that the reasoning for the last claim is not as rigorous as the given solution.

[1] Think the following: there are d cups of water with volumes $\lambda_1 \geq \cdots \geq \lambda_d$, and you can take a fraction $\alpha_i \in [0, 1]$ of each of them subject to the total fraction $\sum_{i=1}^d \alpha_i$ being p. What is the maximum total amount of water you can get?

Problem 2 Hidden Markov Models (18 points)

Recall a hidden Markov model is parameterized by:

- initial state distribution $P(Z_1 = s) = \pi_s$,
- transition distribution $P(Z_{t+1} = s' \mid Z_t = s) = a_{s,s'}$,
- emission distribution $P(X_t = o \mid Z_t = s) = b_{s,o}$.

2.1 In the lecture, we discussed how to find the most likely hidden state path given only observations for the first $T_0 < T$ steps. In this problem, you need to generalize the algorithm to the case when you only observe data from an arbitrary subset of time steps. More concretely, for a given subset $\mathcal{M} \subset \{1, \ldots, T\}$, find

$$\arg\max_{z_{1:T}} P(Z_{1:T} = z_{1:T} \mid X_t = x_t, \ \forall t \in \mathcal{M}).$$

No derivation/reasoning is needed — simply fill in Lines 3, 6, and 7 of the pseudocode below. **(7 points)**

Algorithm 1: Viterbi Algorithm with Missing Data

1 **Input:** Observations $\{x_t\}_{t \in \mathcal{M}}$.
2 **Output:** The most likely path $z_1^*, \ldots, z_T^*$.
3 **Initialize:** For each $s \in [S]$, compute $\delta_s(1) = \begin{cases} \pi_s b_{s,x_1} & \text{if } 1 \in \mathcal{M}, \\ \pi_s & \text{else.} \end{cases}$
4 **for** $t = 2, \ldots, T$ **do**
5 **for** *each* $s \in [S]$ **do**
6 Compute

$$\delta_s(t) = \begin{cases} b_{s,x_t} \max_{s'} a_{s',s} \delta_{s'}(t-1) & \text{if } t \in \mathcal{M} \\ \max_{s'} a_{s',s} \delta_{s'}(t-1) & \text{else} \end{cases}$$

$$\Delta_s(t) = \arg\max_{s'} a_{s',s} \delta_{s'}(t-1).$$

7 Backtracking: let $z_T^* = \arg\max_s \delta_s(T)$. For $t = T, \ldots, 2$, set $z_{t-1}^* = \Delta_{z_t^*}(t)$.

Rubrics:

- 2 points for Line 3.
- 3 points for Line 6.
- 2 points for Line 7.

2.2 (The next two questions are unrelated to the first one.) Suppose we observe a sequence of outcomes $x_1, \ldots, x_{t-1}, x_{t+1}, \ldots, x_T$ with the outcome at time t missing ($2 \le t \le T-1$). Derive the conditional probability of the state at time t being s, that is,

$$P(Z_t = s \mid X_{1:t-1} = x_{1:t-1}, X_{t+1:T} = x_{t+1:T}).$$

Express you answer in terms of the forward message at time $t-1$:

$$\alpha_{s'}(t-1) = P(Z_{t-1} = s', X_{1:t-1} = x_{1:t-1}), \ \forall s' \in \{1, \ldots, S\}$$

and the backward message at time t:

$$\beta_{s'}(t) = P(X_{t+1:T} = x_{t+1:T} \mid Z_t = s'),\ \forall s' \in \{1, \ldots, S\}.$$

You can use the proportional sign in your derivation. However, to test if you fully understand its meaning, you need to express your final answer WITHOUT using the proportional sign. **(6 points)**

$$\begin{aligned}
&P(Z_t = s \mid X_{1:t-1} = x_{1:t-1}, X_{t+1:T} = x_{t+1:T}) \\
&\propto P(Z_t = s, X_{1:t-1} = x_{1:t-1}, X_{t+1:T} = x_{t+1:T}) && \text{(1 point)} \\
&= P(X_{t+1:T} = x_{t+1:T} \mid Z_t = s, X_{1:t-1} = x_{1:t-1}) P(Z_t = s, X_{1:t-1} = x_{1:t-1}) \\
&= P(X_{t+1:T} = x_{t+1:T} \mid Z_t = s) P(Z_t = s, X_{1:t-1} = x_{1:t-1}) && \text{(by conditional independence)} \\
&= \beta_s(t) P(Z_t = s, X_{1:t-1} = x_{1:t-1}) && \text{(2 points)} \\
&= \beta_s(t) \sum_{s'} P(Z_t = s, Z_{t-1} = s', X_{1:t-1} = x_{1:t-1}) && \text{(marginalizing)} \\
&= \beta_s(t) \sum_{s'} P(Z_t = s \mid Z_{t-1} = s') P(Z_{t-1} = s', X_{1:t-1} = x_{1:t-1}) && \text{(by conditional independence)} \\
&= \beta_s(t) \sum_{s'} a_{s',s} \alpha_{s'}(t-1). && \text{(2 points)}
\end{aligned}$$

Therefore, we have

$$P(Z_t = s \mid X_{1:t-1} = x_{1:t-1}, X_{t+1:T} = x_{t+1:T}) = \frac{\beta_s(t) \sum_{s'} a_{s',s} \alpha_{s'}(t-1)}{\sum_{s''} \beta_{s''}(t) \sum_{s'} a_{s',s''} \alpha_{s'}(t-1)} \quad \text{(1 point)}$$

Rubrics: The points shown above are for reference. Give partial credits as appropriate.

2.3 Continuing from the last question, derive the conditional probability of the outcome at time t being o, that is,

$$P(X_t = o \mid X_{1:t-1} = x_{1:t-1}, X_{t+1:T} = x_{t+1:T}).$$

You can express you answer using the quantity $P(Z_t = s \mid X_{1:t-1} = x_{1:t-1}, X_{t+1:T} = x_{t+1:T})$ from the last question. Similarly, express your final answer without using the proportional sign. **(5 points)**

$$\begin{aligned}
&P(X_t = o \mid X_{1:t-1} = x_{1:t-1}, X_{t+1:T} = x_{t+1:T}) \\
&= \sum_s P(X_t = o, Z_t = s \mid X_{1:t-1} = x_{1:t-1}, X_{t+1:T} = x_{t+1:T}) && \text{(marginalizing, 2 points)} \\
&= \sum_s P(X_t = o \mid Z_t = s) P(Z_t = s \mid X_{1:t-1} = x_{1:t-1}, X_{t+1:T} = x_{t+1:T}) && \text{(conditional independence, 2 points)} \\
&= \sum_s b_{s,o} P(Z_t = s \mid X_{1:t-1} = x_{1:t-1}, X_{t+1:T} = x_{t+1:T}) && \text{(1 point)}
\end{aligned}$$