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

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

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).

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)**

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

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)**

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∈M}.
2 Output: The most likely path z*_1, ..., z*_T.
3 Initialize:
4 for t = 2, ..., T do
5     for each s ∈ [S] do
6         Compute

7 Backtracking:
```

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 ($2 \leq t \leq T-1$). 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)**

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)**