Rohan Mahendra Chaudhari

USCID : 6675-5653-85

1 Problem 1 - Principal Component Analysis

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 $x_1, ..., x_N \in R^D$ with zero mean, and we would like to compress it into a one-dimensional dataset $c_1, ..., c_N \in R$. To reconstruct the dataset (approximately), we also keep a direction vector $v \in R^D$ with unit norm (i.e. $||v||_2 = 1$) so that the reconstructed dataset is $c_1\boldsymbol{v}$,...,$c_N\boldsymbol{v} \in R^D$.

Solution

1. Consider the hint of fixing $\boldsymbol{v}$

For any fixed $\boldsymbol{v}$, we optimize over each c_n independently:

$\text{argmin}_{c_n} ||\boldsymbol{x}_n - c_n\boldsymbol{v}||_2^2 = \text{argmin}_{c_n}(c_n^2||\boldsymbol{v}||_2^2 - (2\boldsymbol{x}_n^T\boldsymbol{v})c_n + ||\boldsymbol{x}_n||_2^2)$

$= \text{argmin}_{c_n}(c_n^2 - (2\boldsymbol{x}_n^T\boldsymbol{v})c_n)$, since $||v||_2^2 = 1$

$= \text{argmin}_{c_n}(c_n - \boldsymbol{x}_n^T\boldsymbol{v})^2$

$= \boldsymbol{x}_n^T\boldsymbol{v}$

2. Substitute c = $\boldsymbol{x}_n^T$ into $\text{argmin}_{c_1,...,c_N,\boldsymbol{v}:||\boldsymbol{v}||_2=1} \sum_{n=1}^N ||\boldsymbol{x}_n - c_n\boldsymbol{v}||_2^2$, we see that $\boldsymbol{v}$ is the solution of,

$\text{argmin}_{\boldsymbol{v}:||\boldsymbol{v}||_2=1} \sum_{n=1}^N ||\boldsymbol{x}_n - \boldsymbol{v}\boldsymbol{x}_n^T\boldsymbol{v}_2^2|| = \text{argmin}_{\boldsymbol{v}:||\boldsymbol{v}||_2=1} \sum_{n=1}^N(||\boldsymbol{x}_n||_2^2 - 2(\boldsymbol{x}_n^T\boldsymbol{v})^2 + (\boldsymbol{x}_n^T\boldsymbol{v})^2||\boldsymbol{v}_n||_2^2)$

$= \text{argmin}_{\boldsymbol{v}:||\boldsymbol{v}||_2=1} \sum_{n=1}^N(-(\boldsymbol{x}_n^T\boldsymbol{v})^2)$

$= \text{argmin}_{\boldsymbol{v}:||\boldsymbol{v}||_2=1} - \sum_{n=1}^N(\boldsymbol{v}^T\boldsymbol{x}_n\boldsymbol{x}_n^T\boldsymbol{v})$

$= \text{argmin}_{\boldsymbol{v}:||\boldsymbol{v}||_2=1} \boldsymbol{v}^T(\sum_{n=1}^N \boldsymbol{x}_n\boldsymbol{x}_n^T)\boldsymbol{v}$

As discussed in the lecture, the above mentioned equation is the top eigenvector of the covariance matrix of the dataset and can thus be concluded as the first principal component.

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 $\boldsymbol{c}_1, ..., \boldsymbol{c}_N \in R^p$. To reconstruct the dataset (approximately), we also keep p orthogonal direction vectors $\boldsymbol{v}_1, ..., \boldsymbol{v}_p \in R^D$ with unit norm. For notational convenience, we stack these vectors together as a matrix $\boldsymbol{V} \in R^{D*p}$ whose j-th column is $\boldsymbol{v}_j$.

Solution

1. The reconstructed dataset is $\boldsymbol{V}_{\boldsymbol{c}_1}, ..., \boldsymbol{V}_{\boldsymbol{c}_N}$. Thus the optimization problem is

$$\text{argmin}_{\boldsymbol{c}_1,...,\boldsymbol{c}_N \in R^p, \boldsymbol{V} \in R^{D*p}: \boldsymbol{V}^T\boldsymbol{V}=\boldsymbol{I}} \sum_{n=1}^N ||\boldsymbol{x}_n - \boldsymbol{V}\boldsymbol{c}_n||_2^2$$

Here, $\boldsymbol{I}$ is a identity matrix of order p x p.

2. Here, we do the optimization independently for each c_n,

$\text{argmin}_{\boldsymbol{c}_n} \sum_{n=1}^N ||\boldsymbol{x}_n - \boldsymbol{V}\boldsymbol{c}_n||_2^2 = \text{argmin}_{\boldsymbol{c}_n}(||\boldsymbol{x}_n||_2^2 - 2\boldsymbol{x}_n^T\boldsymbol{V}\boldsymbol{c}_n + \boldsymbol{c}_n^T\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{c}_n)$
$= \text{argmin}_{\boldsymbol{c}_n}(\boldsymbol{c}_n^T\boldsymbol{c}_n - 2\boldsymbol{x}_n^T\boldsymbol{V}\boldsymbol{c}_n)$
Thus, set the gradient
$2\boldsymbol{c}_n - 2\boldsymbol{V}^T\boldsymbol{x}_n = 0$
Thus,
$\boldsymbol{c}_n = \boldsymbol{v}^T\boldsymbol{x}_n$

3. Substitute $\boldsymbol{c}_n = \boldsymbol{v}^T\boldsymbol{x}_n$,
$\text{argmin}_{\boldsymbol{V}:\boldsymbol{V}^T\boldsymbol{V}=\boldsymbol{I}} \sum_{n=1}^N ||\boldsymbol{x}_n - \boldsymbol{V}\boldsymbol{V}^T\boldsymbol{x}_n||_2^2 = \text{argmin}_{\boldsymbol{V}:\boldsymbol{V}^T\boldsymbol{V}=\boldsymbol{I}} \sum_{n=1}^N(||\boldsymbol{x}_n||_2^2 - 2\boldsymbol{x}_n^T\boldsymbol{V}\boldsymbol{V}^T\boldsymbol{x}_n + \boldsymbol{x}_n^T\boldsymbol{V}\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{V}^T\boldsymbol{x}_n)$
$= \text{argmin}_{\boldsymbol{V}:\boldsymbol{V}^T\boldsymbol{V}=\boldsymbol{I}} \sum_{n=1}^N(-2\boldsymbol{x}_n^T\boldsymbol{V}\boldsymbol{V}^T\boldsymbol{x}_n + \boldsymbol{x}_n^T\boldsymbol{V}\boldsymbol{V}^T\boldsymbol{x}_n)$
$= \text{argmin}_{\boldsymbol{V}:\boldsymbol{V}^T\boldsymbol{V}=\boldsymbol{I}} \sum_{n=1}^N(\boldsymbol{x}_n^T\boldsymbol{V}\boldsymbol{V}^T\boldsymbol{x}_n)$
$= \text{argmin}_{\boldsymbol{V}:\boldsymbol{V}^T\boldsymbol{V}=\boldsymbol{I}} \sum_{n=1}^N(\boldsymbol{x}_n^T(\sum_{j=1}^p \boldsymbol{v}_j\boldsymbol{v}_j^T)\boldsymbol{x}_n)$
$= \text{argmin}_{\boldsymbol{V}:\boldsymbol{V}^T\boldsymbol{V}=\boldsymbol{I}} \sum_{j=1}^p(\boldsymbol{v}_j^T(\sum_{n=1}^N \boldsymbol{x}_n\boldsymbol{x}_n^T)\boldsymbol{v}_j)$
To solve the last problem, we first find $\boldsymbol{v}_1$ to maximize $\boldsymbol{v}_1^T(\sum_{n=1}^N \boldsymbol{x}_n\boldsymbol{x}_n^T)\boldsymbol{v}_1$ with the constraint that $||\boldsymbol{v}_1||_2^2 = 1$ is the top eigenvector of the covariance matrix.
Next we find $\boldsymbol{v}_2$ to maximize $\boldsymbol{v}_2^T(\sum_{n=1}^N \boldsymbol{x}_n\boldsymbol{x}_n^T)\boldsymbol{v}_2$ with the constraint that $\boldsymbol{v}_2$ is orthogonal to $\boldsymbol{v}_1$ and $||\boldsymbol{v}_2||_2^2 = 1$. This will be the second eigenvector.
Thus, similarly $\boldsymbol{v}_j$ will be the j-th eigenvector or the j-th principle component.

2 Problem 2 - Hidden Markov Model

2.1 In the lecture, we discussed how to find the most likely hidden state path given only observations for the first T0 < 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 M 1,..., T, find

$arg\max z_{1:T} P(Z_{1:T} = z_{1:T} | X_t = x_t, \forall t \in M)$

Solution

Algorithm 1 Viterbi Algorithm with missing data

Input : Observations $\{x_t\}_{t \in M}$

Output : The most likely path $z_1^*, ..., z_T^*$

Initialize : For each s ∈ [S], compute $\delta_s(1) = \begin{Bmatrix} \pi_s b_{s,x_1} & 1 \in M \\ \pi_s & else \end{Bmatrix}$

for t = 2,...,T **do**

  for each s ∈ [S] **do**

    Compute

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

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

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

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

P$(Z_t =$ **s** $| \; X_{1:t-1} = x_{1:t-1}, X_{t+1:T} = x_{t+1:T})$

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.

Solution

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

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

$= P(X_{t+1:T} = x_{t+1:T}|Z_t = s) \sum_{s'} P(Z_t = s, Z_{t-1} = s', X_{1:t-1} = x_{1:t-1})$
$= \beta_s(t) \sum_{s'} P(Z_t = s|Z_{t-1} = s')P(Z_{t-1} = s', X_{1:t-1} = x_{1:t-1})$
$= \beta_s(t) \sum_{s'} a_{s',s}\alpha_s(t-1)$

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

Solution
$P(X_t = o|X_{1:t-1} = x_{1:t-1}, X_{t+1:T} = x_{t+1:T})$
$= \sum_s P(X_t = o, Z_t = s|X_{1:t-1} = x_{1:t-1}, X_{t+1:T} = x_{t+1:T})$
$= \sum_s P(X_t = o|Z_t = s)P(Z_t = s|X_{1:t-1} = x_{1:t-1}, X_{t+1:T} = x_{t+1:T})$
$= \sum_s b_{s,o} \sum_{s'} P(Z_t = s, Z_{t-1} = s'|X_{1:t-1} = x_{1:t-1}, X_{t+1:T} = x_{t+1:T})$
Thus, the above answer is written in terms of the equation in Q2.2
On substituting the answer of Q2.2,
$= \sum_s b_{s,o}(\beta_s(t) \sum_{s'} a_{s',s}\alpha_s(t-1))$