December 2, 2017 Lecture 23, Nov. 29, 2017

Dimensionality Reduction

Lecturer: Justin Domke *Scribe: Lakshmi Vikraman*

1 Summary

Last class, we talked about the KMeans algorithm. Today, we are going to discuss Principal Components Analysis(PCA) which one of the most common method for dimensionality reduction. Reading for the lecture is ESL 14.5.1.

2 Dimensionality Reduction

Given a dataset X_1,X_2,....,X_N where $X_i \in R^p$, the goal is to find an encoder $t(X)$ and decoder $S(\lambda)$ such that

- $t(X) \in R^q$ where $q < p$
- $S(t(X_i)) \approx X_i$

This is highly data dependent and is in general not possible. Suppose we have $p = 2$ and $q = 1$. In the case where we have uniformly distributed data, as shown in figure 2.1 there is no obvious way to compress it.

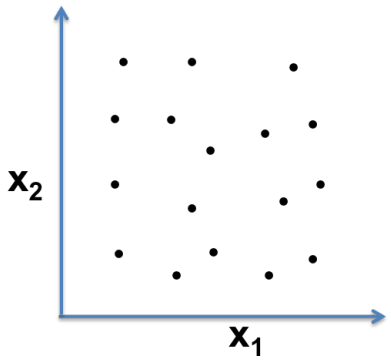
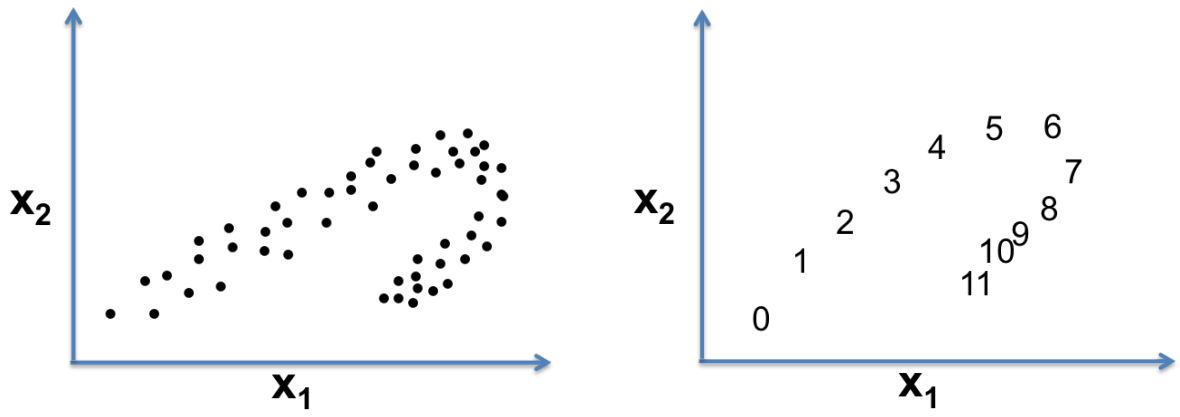


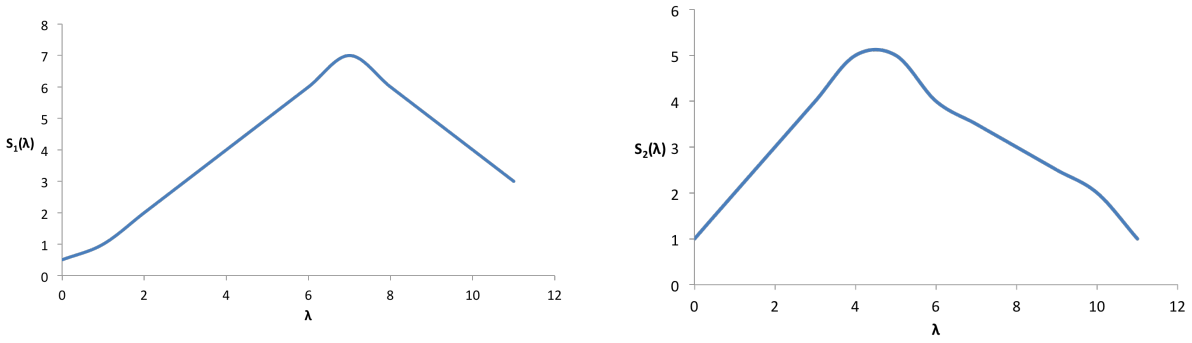
Figure 2.1: Uniformly distributed data

Consider a case where we have a dataset which has points only over a subset of the space as shown in figure 2.2a. The fig 2.2b shows the value of $t(X)$ for this data. Figures 2.3a and 2.3b shows how $S_1(\lambda)$ varies with λ and $S_2(\lambda)$ varies with λ respectively.



(a) Data (b) t(X) for Data

Figure 2.2



(a) Plot of $S_1(\lambda)$ with λ (b) Plot of $S_2(\lambda)$ with λ

Figure 2.3

2.1 Principal Components Analysis

The goal is to fit an encoder $t(X)$ and a decoder $S(\lambda)$ to minimize $\sum_{i=1}^{N} \|S(t(X_i)) - X_i\|$. This is called reconstruction error. In PCA we will focus on affine dimensionality reduction. We are interested in

$$S(\lambda) = \mu + V_q \lambda$$

where $\mu \in R^p, V_q \in R^{p \times q}$.

Assume that the columns of V_q are orthogonal where
$V_q = [v_1, v_2, ...v_q]$ where $v_i^T v_j = I[i = j]$
$V_q^T V_q = I$

The goal is to minimize $\sum_{i=1}^{N} \|X_i - S(t(X_i))\|^2$.

A few examples were shown in class. The first example was a dataset of faces and a codebook where each square corresponds to the column V_q. By combining the V_q with λ values, we get different reconstructions of the faces. The second example was a timeseries dataset where it was shown how the high dimensional input can be represented in terms of V_q and λ. The third example represents natural image patches and how it is represented by a combination of the codes.

What is the ideal encoder with decoder fixed?
Input : $S(\lambda), X$
Output : $\underset{\lambda}{\operatorname{argmin}} \|X - S(\lambda)\|$

What is λ with $S(\lambda) = \mu + V_q \lambda$?
Theorem : Best encoder is $\lambda = t(X) = V_q^T (X - \mu)$
Proof :

$$\min_{\lambda} \|X - S(\lambda)\|^2$$
$$\min_{\lambda} \|X - \mu - V_q \lambda\|^2$$
$$\frac{d}{dt} \|X - \mu - V_q \lambda\|^2 = 0$$
$$-2V_q^T (X - \mu - V_q \lambda) = 0$$
$$V_q^T (X - \mu) = V_q^T V_q \lambda$$
$$V_q^T (X - \mu) = \lambda$$

Story so far

- Want to fit an encoder $t(X)$ and a decoder $S(X)$ to minimize reconstruction error $\sum_{i=1}^{N} \|X_i - S(t(X_i))\|^2$.

- if we choose $S(\lambda) = \mu + V_q\lambda$, best possible encoder is $t(X) = V_q^T(X - \mu)$.
- Thus we need to solve

$$\min_{\mu,V_q} \sum_{i=1}^{N} \|X_i - S(t(X_i))\|^2$$

$$\min_{\mu,V_q} \sum_{i=1}^{N} \left\|X_i - S(V_q^T(X_i - \mu))\right\|^2$$

$$\min_{\mu,V_q} \sum_{i=1}^{N} \left\|X_i - \mu - V_qV_q^T(X_i - \mu)\right\|^2$$

Find best μ
Theorem : For any V_q, reconstruction error is minimized by
$\mu = \bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$
Proof :

$$\frac{d}{dt} \sum_{i=1}^{N} \left\|X_i - \mu - V_qV_q^T(X_i - \mu)\right\|^2$$

$$= -2\sum_{i=1}^{N} X_i - \mu - V_qV_q^T(X_i - \mu)(I + V_qV_q^T)$$

Best μ is not unique, but by setting $\mu = \bar{X}$, the derivative will be 0, since $\sum_{i=1}^{N}(X_i - \bar{X}) = 0$.

Updated Story:
We need to solve $\min_{V_q} \sum_{i=1}^{N} \left\|X_i - \bar{X} - V_qV_q^T(X_i - \bar{X})\right\|^2$
Without loss of generality, assume that $\bar{X} = 0$.
Then we have to solve $\min_{V_q} \sum_{i=1}^{N} \left\|X_i - V_qV_q^T(X_i)\right\|^2$
We will use SVD to solve this.

2.1.1 Review of SVD

If $X \in R^{N\times p}$, SVD is a factorization $X = UDV^T$ where ,

- U is an orthogonal matrix , $U \in R^{N\times p}$, $U^TU = I \in R^{p\times p}$.
- D is a diagonal matrix , $D \in R^{p\times p}$ where the diagonal elements $d_1, d_2..., d_p$ is non-increasing ie $d_1 \geq d_2..... \geq d_p$.
- V is an orthogonal matrix , $V \in R^{p\times p}$, $V^TV = I \in R^{p\times p}$.

SVD gives the best low rank approximation of a matrix.
Given $X = UDV^T$, we have to find the best approximation $\tilde{X}$ to minimize $\left\|X - \tilde{X}\right\|_F^2$ such that $\text{rank}(\tilde{X}) = q$.

Best $\tilde{X} = UD_qV_q^T$ where

$$D = \begin{bmatrix} d_1 & & \\ & \ddots & \\ & & d_q \end{bmatrix}$$

$$V_q = \begin{pmatrix} v_1 & v_2 & ..... & v_q \end{pmatrix}$$

2.1.2 PCA using SVD

Theorem:
Given dataset , $X \in R^{N\times p} = [x_1, x_2, ...x_N]^T$ and SVD of $X = UDV^T$, the optimal PCA projection is V_q.
Proof:

$$\begin{aligned}
&\sum_{i=1}^{N} \left\| X_i - V_qV_q^T(X_i) \right\|^2 \\
&= \left\| X - XV_qV_q^T \right\|_F^2 \\
&= \left\| X - UDV^TV_qV_q^T \right\|_F^2 \\
&= \left\| X - UD_qV_q^T \right\|_F^2 \text{ since } DV^TV_q = D_q \\
&= \left\| X - \tilde{X} \right\|_F^2
\end{aligned}$$

PCA was applied to a dataset of faces and the resultant images were shown in class for $k = 1$, $k = 2$ etc where $k = 1$ represents the first column of V_q , k_2, the second and so on.