1:27:45

Module 4

Problem Setting.

In this session we will introduce the setting of PCA and the high level idea.

"Assume we have a dataset x in $\mathbb{R}^D$ consisting of N vectors"

$$X = \{x_1, \ldots x_N\} \quad x_i \in \mathbb{R}^D$$

Our objective is to find the low dimensional representation of the data that is as similar to X as possible

Before we start, let's briefly review 3 important concepts

First, every vector in $\mathbb{R}^D$ can be represented as a linear combination of the basis vectors

Let's write it down: "x_n can be written as the sum of i equal 1 to D, of $\beta(\text{beta})_i$ in times b_i, and we will assume that b_i

are an orthonormal basis of $\mathbb{R}^D$"

$$① \quad x_n = \sum_{i=1}^{D} \beta_{in} b_i$$

If we assume that we use the dotproduct as our inner product, and b_i to be D or orthonormal basis,

Second: we can also write "the β_{in} as x_n transpose times b_i", which means we can interpet β_{in} to be the orthogonal projection of x_n onto one dimensional subspace ~~(span b)~~ spanned by the i^{th} basis vector

$$② \quad \beta_{in} = x_n^T b_i$$

Third: If we have an orthonormal basis b_1 to b_m, of ~~$\mathbb{R}^D$~~ $\mathbb{R}^D$ and we define B to be the matrix that consists of these orthonormal basis vectors, $B(b_1, \ldots b_m)$", then the projection of x onto the subspace

"we can write $\tilde{x}$ (tilda) as B times Btranspose times X", that means $\tilde{x}$ (tilda) is the orthogonal projection of X onto the subspace, spanned by the M basis vectors

(3) $B = (b_1, \ldots b_M)$

$$\tilde{x} = BB^T x$$

And $BB^T x$ is the coordinate of $\tilde{x}$ w.r.t. the basis vectors collected in matrix B

This $(BB^T x)$ is also called the coordinates or <u>CODE</u>

Now lets have a look at PCA.

The key idea in PCA is to find a lower dim representation $\tilde{x}_n$ of x_n that can be expressed using fewer basis vectors

(see pic)

(1)

1. Centered data: $E[X] = 0$
2. ONB $b_1, \ldots, b_D$

(1) we assume the data is centered, that means the dataset has means 0,

(2) and we also assume that b_1 to b_D are an orthonormal basis (ONB) of $\mathbb{R}^D$

Formally we can write any $\tilde{x}_n$ in the following way

* $\tilde{x}_n$ can be written as a sum, $i = 1$ to M, of β_{ni} (β) in times b_i plus sum of $i = M+1$ to D of β in time b_i still living in $\mathbb{R}^D$ "

$$\tilde{x}_n = \underbrace{\sum_{i=1}^{M} \beta_{in} b_i}_{\text{(B)}} + \underbrace{\sum_{i=M+1}^{D} \beta_{in} b_i}_{\text{(A)}} \in \mathbb{R}^D$$

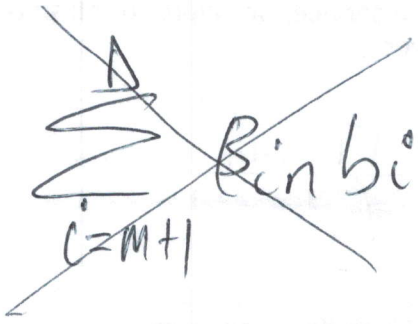
So we took our general way of writing any vector in $\mathbb{R}^D$ which came from property ①, and we split the sum in property ① into 2 sums.

One is living in an M dimensional subspace.

Other one is living in D-M dimensional subspace which is an orthogonal component to the particular subspace.

In PCA we ignore the second term.

So we get rid of $\sum_{i=M+1}^{D} \beta_{in} b_i$



this part

And then we call the subspace that is spanned by the basis vectors b_i to b_M, the principal subspace

Although $\tilde{X}_n$ is still a D dimensional vector, it lives in a M dim subspace of $\mathbb{R}^D$, and only M coordinates β_{ni} to β_{nm}, are necessary in order to represent it.

Do these ones of (A) and are the coordinates of the (B) (page 5)

The β (Betas)'s are also called the code or coordinates of $\tilde{X}_n$ wrt to basis vectors $b_1 \ldots b_M$

And the setting now is as follows:

Assuming we have data $X_1 \ldots X_N$, we want to find parameters β_{in} and orthonormal basis vectors b_i, such that the average squared reconstruction error is minimized

And we can write the average squared reconstruction error, as follows:

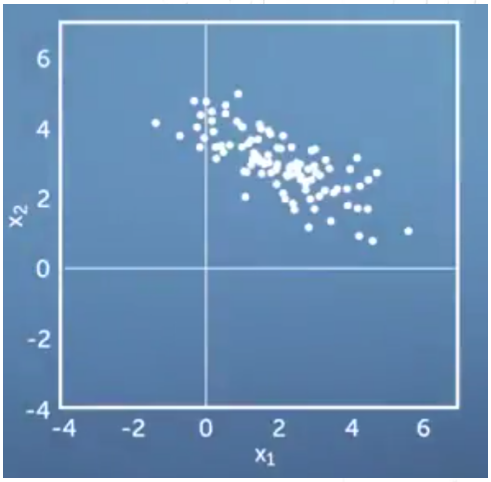
Can write "J" ← average squared reconstruction error

"J is going to be 1 over N times sum n equal 1 to N || then we write X_n minus $\tilde{X}_n$ (tilda) squared"

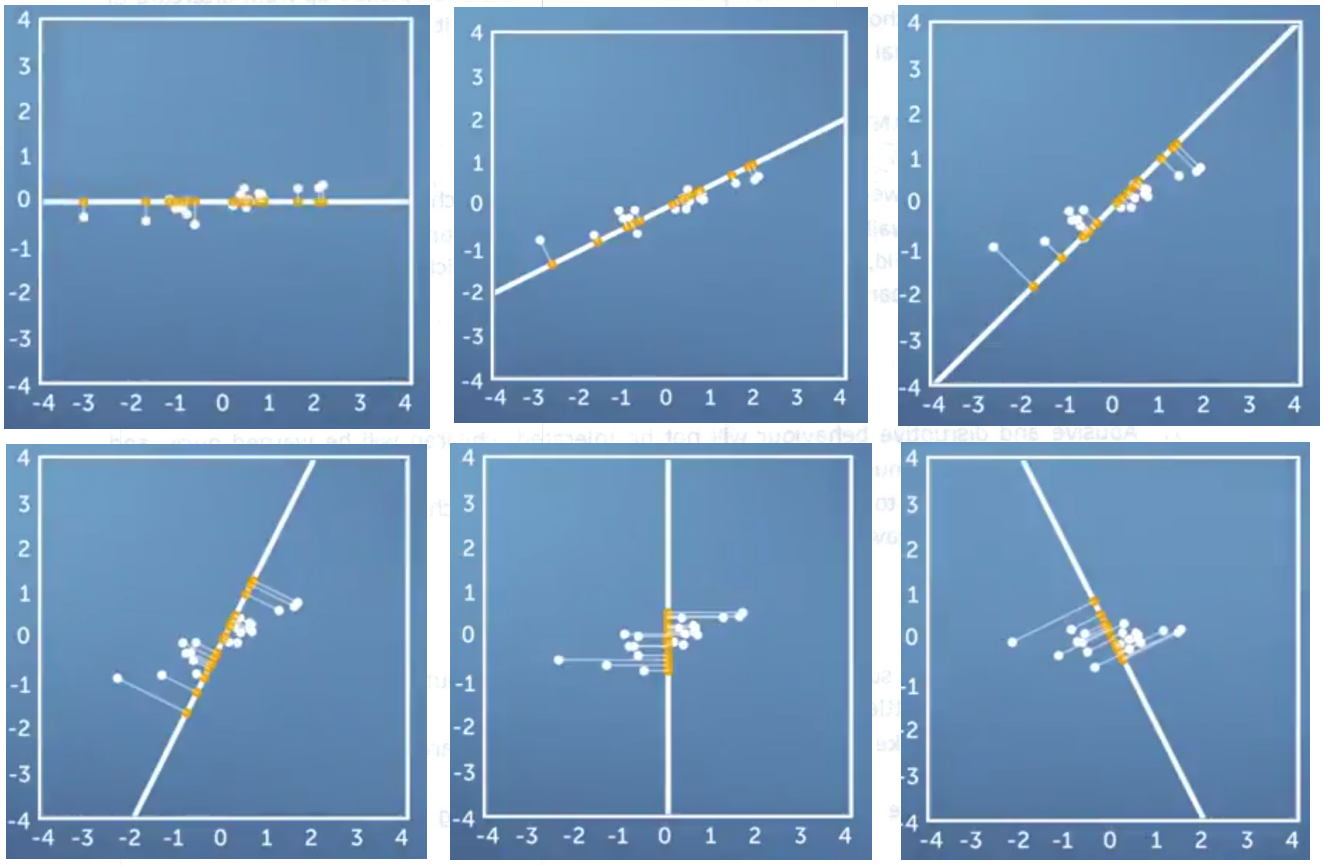
$$J = \frac{1}{N} \sum_{n=1}^{N} \| X_n - \hat{X}_n \|^2$$

Let's have a look at an example:

we have data living in 2 Dim, and now we want to find a good 1 Dim subspace such that the squared or average squared reconstruction error of the data, orginal data points in the corresponding projection is minimized (see pic)



Here I am plotting the original data set with the corresponding projections onto 1 Dim subspace, and cycling through a couple of options of subspaces and you can see that some of the projections are significantly more informative than other, and in PCA we going to find the best one: (see PCs)



Our approach is to compute partial derivatives of J (cost) w.r.t the parameters:

The parameters are:

- β_{in}
- b_i's

We set the partial derivatives of J w.r.t those parameters to zero(0), and solve for the optimal parameters.

But one observation we can already make. and that observation is that the parameters only enter this loss function (J) through

$\hat{X}_n$

This means in order to get our partial derivatives, we need to apply the Chain rule, Sum

Do we can write" ∂J over ∂ either β_{in} or b_i can be written as ∂J over $\partial \hat{x}_n$ times $\partial \hat{x}_n$ over ∂ either β_{in} or b_i

$$\frac{\partial J}{\partial \{\beta_{in}, b_i\}} = \frac{\partial J}{\partial \hat{x}_n} \frac{\partial \hat{x}_n}{\partial \{\beta_{in}, b_i\}}$$

And the first part we can already compute, and we get

" $\frac{\partial J}{\partial \hat{x}_n}$ = is minus 2 over N, times X_n minus $\hat{X}_n$ transpose "

$$\frac{\partial J}{\partial \hat{x}_n} = -\frac{2}{N}\left(x_n - \hat{x}_n\right)^T$$

and other derivatives we compute in next session.