Intro

Statistical machine learning (Mark) lecture.

In this Case we going through the mathematical foundations that we need in order to do the dimensionality reduction with Principal Component Analysis

Data in real life is often higher dimensional

For eg if you want to estimate the price of housing in your time, we can use data to do this

Type, size, number bedrooms and bathrooms, value of houses in neighbourhood, when they were build, the distance to next ~~st~~ train station, park, number of crimes committed in neighbourhood, Economic climate, etc

∴ Many things that influence house price and we collect this information in dataset, that we can use to estimate price of a house

Another example is 640 x 480 pixel colour image, which is data point in one dim. space. Where every pixel correspond to 3 dim, one for each colour channel, red, green, blue.

Working with high dim data, comes with some difficulties

It's hard to analyse, interpretation is difficult, visualization is nearly impossible and from a practical point of view, storage can be quite expensive.

However, high dimensional data, also have some nice properties

Eg HDim data is often over complete, that means, many dimension are redundant.

and can be explained by combination of other dimension.

Eg if we took a color image of 4 channels red, green, blue, gray, the gray scale (~~chan~~) channel can be explained by combination of other 3 channels, and the image can ~~be~~ be equally represented by red, green and blue alone, as we can reconstruct the gray scale channel just by using that information.

Dimensionality reduction, exploit structure and correlation, and allows us to work with a more compact representation of the data

Ideally, without loosing information

We can think of Dim. reduction as a compression technique, similar to jpeg or mp3. which are compression algorithm for image and music

Let's take a small binary image of the hand written digit "8"

We look at 28x28 pixels, each of which can be either black or white

Therefore this image can be represented as a 784 Dim vector ((28x28) = 784)

However, in this eg. the pixels are not randomly black or white, there is structure

They often have the same values as the neighbouring pixels

In this dataset, there are many examples of an 8:

8 8 8 8 8 8

They differ a little bit, but they look sufficiently similar that we can identify them as 8.

We can use dim reduction, to find a lower dimensional representation of all 8's that is easier to work with, than a 784 dim vector.

The lower dim. representation of a higher dim. data point is often called a feature or a code. (5)

In this course we will look at a classical algorithm for linear dim. reduction, i.e. "principal component analysis" or PCA.

The purpose of this course is to go through the necessary mathematical details to derive PCA.

In Module 1, we will look at statistical representation of data, e.g. using means and variances, and will also describe how means and variance change, if we linearly transform our data set.

In Module 2, we will look at the geometry in vector spaces and define how to compute distances and angles between vectors.

In Module 3, we will use these results to project data onto lower dim. sub spaces.

In Module 4, we will derive PCA, as a way to reduce dim of data, using orthogonal projections