Principal Components Analysis (PCA)

Sebastian J. Bentley — cph-sb287

April 22, 2022

1 Introduction

Principal Component analysis is a strong tool, in data science. It allows us to reduce the number of dimensions in our datasets, by capturing the maximum variation of the data. For instance, say we have a datasets with ID, and five features.

ID	f1	f2	f3	f4	f5
1	423	123	123	123	123
2	13	421	123	123	123
...	...	...	...	...	...

Table 1: Example: Dataset before PCA.

After using PCA, we can reduce the dimension, so our dataset has a lower dimension, and therefore could look more like this:

ID	f1	f2
1	423	123
2	13	421
...	...	...

Table 2: Example: Dataset after PCA.

These tables does not show what happens to the actual data in dataset, just that we have a lower dimension to work with.

2 Pros and Cons

2.1 Benefit

This is really beneficial, in the way that training our model is most likely much faster! Imagine having millions of data entries, with hundreds of dimensions. It would probably take a long time to train a model with such a large amount of data. With PCA, we can reduce the time it takes to train a model.

2.2 Downside

The downside of using PCA, is that we lose some accuracy of our model. Still, we can often keep a fairly high accuracy, and have it run a lot faster, than a model trained without PCA.

3 Example

You work as a data scientist at an advertisement company. You are given the task to make a model, that can predict sales, if most people stop getting physical advertisement in their mailbox. To train a model, you are given a dataset, with millions of row, with a hundred different features. After cleaning

and preparing a training set, you discover that it is going to take multiple days to train a model, and your deadline is tomorrow! Therefore you decide to use PCA to train your model much faster, so now it only takes a few hours, while still keeping 96% accuracy. You train your model, reach your deadline, and buy a celebratory cake on your way home.