

# Machine Learning Homework # 5

sakohl, milsen, jkirchner

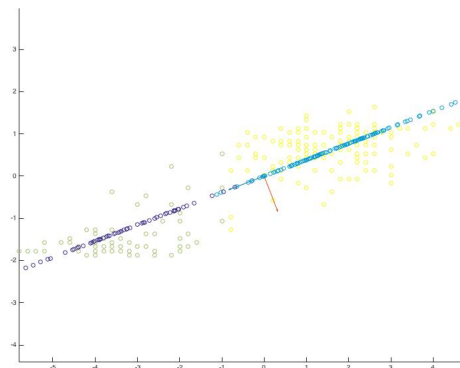
May 23, 2015

## 1 Exercise (Curse of Dimensionality (3p))

- Discuss why increasing number of features in feature space could lead to poorer performance? (3p)
- The term *Curse of Dimensionality* refers to the fact that when representing data in a high dimensional space the number of possible data points grows exponentially with the number of dimensions. This causes many computational problems, since an algorithm might not scale, but also conceptual problems, since high-dimensional spaces have many counter-intuitive properties. For example the volume of an  $n$ -dimensional cube is for large  $n$  mainly concentrated in the outermost shell of the cube. So while a higher number of dimensions of the data space generally allows for a better representation of the data, the curse of dimensionality makes it necessary that we apply techniques like *Dimension Reduction* to work with the data in practice.

## 2 Exercise (PCA (8p))

- 2.2 Plot this reconstructed data and briefly discuss whether the original data lost much information? Why? (2p)
- 



- Of course there is some loss of information since the reconstructed data has one dimension less than the original data. But the relevant portion of the information was preserved - the linear correlation between duration of eruption and intervals between eruptions and also the separation into distinct categories is still there. This worked well because the data was oriented and separated mainly along one axis to begin with.