

DATA11002

# Introduction to Machine Learning

Lecturer: Teemu Roos

TAs: Ville Hyvönen and Janne Leppä-aho

Department of Computer Science

University of Helsinki

(based in part on material by Patrik Hoyer and Jyrki Kivinen)

November 2nd–December 15th 2017

# Principal Component Analysis

# Dimension Reduction

- ▶ When the dimension of the data  $p$  is large, it is often hard to visualize and process the data
- ▶ Classification and regression models easily overfit
- ▶ However, in many cases, the data can be approximately summarized by a lower-dimensional representation
  - ▶ For example, large questionnaire data sets can often be reduced to a few dimensions (cf. psychological scales such as Myers–Briggs, Keirsey)
  - ▶ Intelligence quotient (IQ) is a one-dimensional representation of set of related but different skills (of completing puzzles)
- ▶ Visualization requires one-, two- or three-dimensional representations

# Principal Component Analysis

- ▶ Principal Component Analysis is a common dimensionality reduction technique
- ▶ The basic idea is to project the data onto a lower dimensional subspace so that as much variance as possible is retained
- ▶ Assume from now on that data is zero-centered
  - ▶ If the original instances are  $\mathbf{z}_1, \dots, \mathbf{z}_n$ , we replace them by  $\mathbf{x}_i = \mathbf{z}_i - \bar{\mathbf{z}}$  where  $\bar{\mathbf{z}} = \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i$  denotes the overall mean
  - ▶ Then  $\sum_i \mathbf{x}_i = 0$  and therefore the mean of each feature,  $j$ , is also zero  $\bar{x}_j = 0$  for  $j = 1, \dots, p$

## Principal Component Analysis (2)

- ▶ Pick now a unit vector  $\phi \in \mathbb{R}^p$  and project all instances  $\mathbf{x}_i$  along direction  $\phi$
- ▶ The scalar projection of  $\mathbf{x}_i$  is  $\phi^T \mathbf{x}_i$ , and its variance is

$$\sum_{i=1}^n (\phi^T \mathbf{x}_i)^2 = \sum_{i=1}^n \left( \sum_{j=1}^p \phi_j x_{ij} \right)^2$$

(recall that the data is zero-centered)

- ▶ The unit vector  $\phi$  that maximises this is the Eigenvector of  $X^T X$  corresponding to the largest Eigenvalue: eigen in R
- ▶ The resulting vector  $\phi$  is called the first principal component of the data

## Principal Component Analysis (3)

- ▶ In Principal Component Analysis (PCA) we first find  $\mathbf{v}_1, \dots, \mathbf{v}_k$ , the Eigenvectors corresponding to  $k$  largest Eigenvalues of  $X^T X$
- ▶ Then  $\mathbf{x}_i$  is replaced by its projection  $\mathbf{x}'_i$  onto subspace spanned by  $\mathbf{v}_1, \dots, \mathbf{v}_k$ :

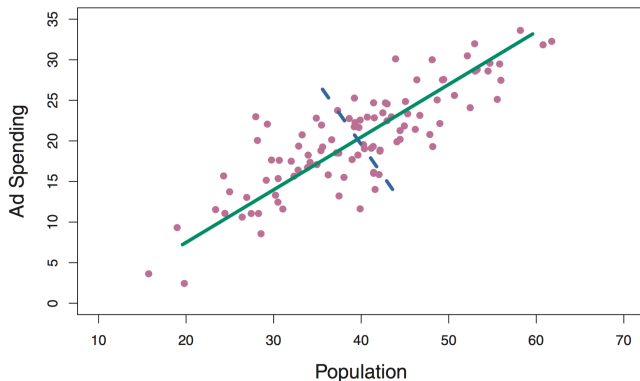
$$\mathbf{x}'_i = \mathbf{v}_1 (\mathbf{v}_1^T \mathbf{x}_i) + \dots + \mathbf{v}_k (\mathbf{v}_k^T \mathbf{x}_i),$$

where  $\mathbf{v}_i^T \mathbf{x}_i$  is the scalar projection, and  $\mathbf{v}_i (\mathbf{v}_i^T \mathbf{x}_i)$  is the corresponding vector projection.

- ▶ Among all possible linear projections of the data onto a  $k$  dimensional subspace, this method
  - ▶ maximises the variance of  $\mathbf{x}'_i$
  - ▶ minimises the “squared error”  $\sum_i \|\mathbf{x}_i - \mathbf{x}'_i\|_2^2$

# Principal Component Analysis (3)

## ► Example



Source: (James et al., 2013), p. 230

## Principal Component Analysis (4)

- ▶ Other linear dimensionality reduction techniques include Independent Component Analysis and Factor Analysis
- ▶ Furthermore, non-linear techniques such as Isomap (and other manifold methods) and kernel PCA allow non-linear mappings
- ▶ If the variables are independent, the principal components are simply the  $k$  variables with the highest variance: then PCA would simply do *feature selection*
- ▶ This also makes it clear that the *scale* of the variables (e.g., grams vs kilograms) is important
- ▶ Often the variances are forced to be equal by normalizing them to be one



# Principal Components: Interpretation

- ▶ The principal components, i.e., the vectors  $\phi$ , also have an interpretation
- ▶ Remember that each  $\phi_i$  is a unit vector of dimension  $p$
- ▶ Each element  $\phi_{ij}$  is called the **loading** (“weight”) of the  $i$ th principal component on variable  $j$
- ▶ If variables  $j$  and  $j'$  have similar loadings, they are usually correlated, for example:
  - ▶ companies in the same business sector
  - ▶ genes regulated by same factors
  - ▶ users' preferences on music or movie genres  
(*Four Weddings and a Funeral* & *Bridget Jones' Baby* vs *Prometheus* & *Rogue One*)
- ▶ Read the textbook Sec. 10.2 for a more thorough explanation of this

# Why dimensionality reduction?

- ▶ Understanding data: see where the variance comes from
- ▶ Visualisation: reduce to 2 or 3 dimensions and plot
- ▶ Whitening: it can be shown that the components are uncorrelated
- ▶ Lossy image compression: keeping only some of the principal components (with suitable pre-processing) may still give adequate quality
- ▶ Image denoising: dropping the lower components may even improve the quality of a noisy image
- ▶ Preprocessing for supervised learning (**but** directions with large variance may not be the ones that matter for a classification task)