582631 — 5 credits

Introduction to Machine Learning

Lecturer: Teemu Roos
Assistant: Ville Hyvönen

Department of Computer Science
University of Helsinki

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

November 1st–December 16th 2016

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 is a one-dimensional representation of set of related but different skills (of completing puzzles)
- Visualization requires one-, two- or three-dimensional representations

Principal Componenent 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 $\mathbf{x}'_1, \ldots, \mathbf{x}'_n$, we replace them by $\mathbf{x}_i = \mathbf{x}'_i - \bar{\mathbf{x}}'$ where $\bar{\mathbf{x}}' = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}'_i$
  - Then $\sum_i \mathbf{x}_i = 0$ and therefore $\bar{x}_j = 0$ for $j = 1, \ldots, d$

Principal Componenent Analysis (2)

- Pick now a unit vector $\phi \in \mathbb{R}^d$ and project all instances $\mathbf{x}_i$ along direction ϕ
- The projection of $\mathbf{x}_i$ is $\phi^{\mathrm{T}}\mathbf{x}_i$, and the variance of the projections is

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

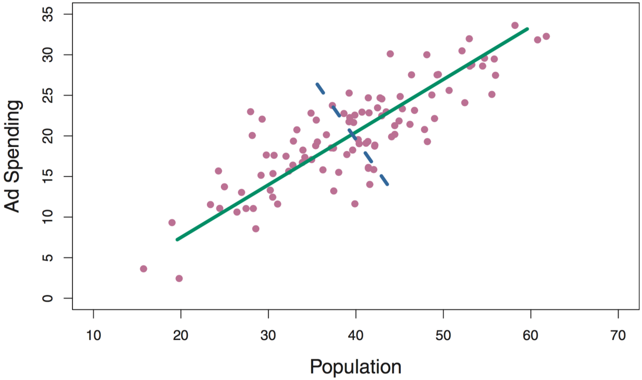
  (recall the data is zero-centered)
- The unit vector ϕ that maximises this is the Eigenvector of $X^{\mathrm{T}}X$ corresponding to the largest Eigenvalue: `eigen` in R
- The resulting vector ϕ is called the first principal component of the data

Principal Componenent Analysis (3)

- In Principal Component Analysis (PCA) we first find $\mathbf{v}_1, \ldots, \mathbf{v}_k$, the Eigenvectors corresponding to k largest Eigenvalues of $X^{\mathrm{T}}X$
- Then $\mathbf{x}_i$ is replaced by its projection $\mathbf{x}_i'$ onto subspace spanned by $\mathbf{v}_1, \ldots, \mathbf{v}_k$:
$$\mathbf{x}_i' = (\mathbf{v}_1^{\mathrm{T}}\mathbf{x}_i, \ldots, \mathbf{v}_k^{\mathrm{T}}\mathbf{x}_i)$$
- 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” $\sum_i \|\mathbf{x}_i - \mathbf{x}_i'\|_2^2$
- 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

Principal Componenent Analysis (3)

- Example



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

Principal Component Analysis (4)

- 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 forces to be equal by normalizing them to be one

Principal Components: Interpretation

- The principal components, i.e., the vectors ϕ, also have an interpretation
- Remember that each ϕ_i is a unit vector of length p
- Each element ϕ_{ij} is called the **loading** ("weight") of the ith 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)