3F3: Signal and Pattern Processing

Lecture 5: Dimensionality Reduction

Zoubin Ghahramani
zoubin@eng.cam.ac.uk

Department of Engineering
University of Cambridge

Lent Term
Dimensionality Reduction

Given some data, the goal is to discover and model the intrinsic dimensions of the data, and/or to project high dimensional data onto a lower number of dimensions that preserve the relevant information.
Principal Components Analysis (PCA)

Data Set $\mathcal{D} = \{x_1, \ldots, x_N\}$ where $x_n \in \mathbb{R}^D$

Assume that the data is zero mean, $\frac{1}{N} \sum_n x_n = 0$.

Principal Components Analysis (PCA) is a linear dimensionality reduction method which finds the linear projection(s) of the data which:

- maximise variance
- minimise squared reconstruction error
- have highest mutual information with the data under a Gaussian model
- are maximum likelihood parameters under a linear Gaussian factor model of the data
PCA: Direction of Maximum Variance

Let \( y = w^\top x \). Find \( w \) such that \( \text{var}(y) \) is maximised for the data set \( D = \{x_1, \ldots, x_N\} \). Since \( D \) is assumed zero mean, \( \mathbb{E}_D(y) = 0 \). Using \( y_n = w^\top x_n \) we optimise:

\[
\mathbf{w}^* = \arg \max_w \text{var}(y) = \arg \max_w \mathbb{E}_D(y^2) = \arg \max_w \frac{1}{N} \sum_n y_n^2
\]

\[
\frac{1}{N} \sum_n y_n^2 = \frac{1}{N} \sum_n (w^\top x_n)^2 = \frac{1}{N} \sum_n w^\top x_n x_n^\top w = w^\top \left( \frac{1}{N} \sum_n x_n x_n^\top \right) w = w^\top C w
\]

where \( C = \frac{1}{N} \sum_n x_n x_n^\top \) is the data covariance matrix. Clearly arbitrarily increasing the magnitude of \( w \) will increase \( \text{var}(y) \), so we will restrict ourselves to directions \( w \) with unit norm, \( \|w\|^2 = w^\top w = 1 \). Using a Lagrange multiplier \( \lambda \) to enforce this constraint:

\[
\mathbf{w}^* = \arg \max_w w^\top C w - \lambda(w^\top w - 1)
\]

Solution \( w^* \) is the eigenvector with maximal eigenvalue of covariance matrix \( C \).
Eigenvalues and Eigenvectors

$\lambda$ is an eigenvalue and $z$ is an eigenvector of $A$ if:

$$Az = \lambda z$$

and $z$ is a unit vector ($z^Tz = 1$).

**Interpretation:** the operation of $A$ in direction $z$ is a scaling by $\lambda$.

The $K$ Principal Components are the $K$ eigenvectors with the largest eigenvalues of the data covariance matrix (i.e. $K$ directions with the largest variance).

Note: $C$ can be decomposed:

$$C = USU^T$$

where $S$ is diag$(\sigma_1^2, \ldots, \sigma_D^2)$ and $U$ is a an orthonormal matrix.
PCA: Minimising Squared Reconstruction Error

Solve the following **minimum reconstruction error** problem:

\[
\begin{align*}
\min_{\{\alpha_n\}, w} & \quad \|x_n - \alpha_n w\|^2 \\
\end{align*}
\]

Solving for \(\alpha_n\) holding \(w\) fixed gives:

\[
\alpha_n = \frac{w^\top x_n}{w^\top w}
\]

Note if we rescale \(w\) to \(\beta w\) and \(\alpha_n\) to \(\alpha_n/\beta\) we get equivalent solutions, so there won’t be a unique minimum. Let’s constrain \(\|w\| = 1\) which implies \(w^\top w = 1\). Plugging \(\alpha_n\) into the original cost we get:

\[
\min_w \sum_n \|x_n - (w^\top x_n)w\|^2
\]

Expanding the quadratic, and adding the Lagrange multiplier, the solution is again:

\[
w^* = \arg \max_w \quad w^\top Cw - \lambda(w^\top w - 1)
\]
PCA: Maximising Mutual Information

Problem: Given \( x \) and assuming that \( P(x) \) is zero mean Gaussian, find \( y = w^\top x \), with \( w \) a unit vector, such that the mutual information \( I(x; y) \) is maximised.

\[
I(x; y) = H(x) + H(y) - H(x, y) = H(y)
\]

So we want to maximise the entropy of \( y \). What is the entropy of a Gaussian?

Let \( z \sim \mathcal{N}(\mu, \Sigma) \), then:

\[
H(z) = - \int p(z) \ln p(z) \, dz = \frac{1}{2} \ln |\Sigma| + \frac{D}{2} (1 + \ln 2\pi)
\]

Therefore we want the distribution of \( y \) to have largest variance (in the multidimensional case, largest volume —i.e. det of covariance matrix).

\[
w^* = \arg \max_w \text{var}(y) \quad \text{subject to} \quad \|w\| = 1
\]
Principal Components Analysis

The full multivariate case of PCA finds a sequence of $K$ orthogonal directions $w_1, w_2, \ldots, w_K$.

Here $w_1$ is the eigenvector with largest eigenvalue of $C$, $w_2$ is the eigenvector with second largest eigenvalue and orthogonal to $w_1$ (i.e. $w_2^\top w_1 = 0$), etc.
Example of PCA: Eigenfaces

from www-white.media.mit.edu/vismod/demos/facerec/basic.html
There are many different ways of generalising PCA to find *nonlinear* directions of variation in the data.

A simple example (very similar to what we did with regression and classification!) is to map the data in some nonlinear way,

$$x \rightarrow \phi(x)$$

and then do PCA on the \{\phi(x_1) \ldots \phi(x_N)\} vectors.

This is sometimes called “kernel PCA” since it can be completely defined in terms of the kernel functions $K(x_n, x_m) = \phi(x_n)^\top \phi(x_m)$, or alternatively in terms of a similarity metric on the inputs.
Summary

We have covered four key topics in machine learning and pattern recognition:

- Classification
- Regression
- Clustering
- Dimensionality Reduction

In each case, we see that these methods can be viewed as building probabilistic models of the data. We can start from simple linear models and build up to nonlinear models.
Appendix: Information, Probability and Entropy

Information is the reduction of uncertainty. How do we measure uncertainty?

Some axioms (informal):

- if something is certain its uncertainty $= 0$
- uncertainty should be maximum if all choices are equally probable
- uncertainty (information) should add for independent sources

This leads to a discrete random variable $X$ having uncertainty equal to the entropy function:

$$H(X) = - \sum_{x \in \mathcal{X}} P(X = x) \log P(X = x)$$

measured in bits (binary digits) if the base 2 logarithm is used or nats (natural digits) if the natural (base $e$) logarithm is used.
Appendix: Information, Probability and Entropy

- **Surprise** (for event \( X = x \)): \(- \log P(X = x)\)

- **Entropy** = average surprise: \( H(X) = - \sum_{x \in X} P(X = x) \log_2 P(X = x) \)

- **Conditional entropy**

\[
H(X|Y) = - \sum_x \sum_y P(x, y) \log_2 P(x|y)
\]

- **Mutual information**

\[
I(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y)
\]

- **Independent random variables**: \( P(x, y) = P(x)P(y) \forall x \forall y \)