Statistical Machine Learning

Christian Walder

College of Engineering and Computer Science
The Australian National University

Canberra
Semester One, 2020.

(Many figures from C. M. Bishop, "Pattern Recognition and Machine Learning")

Outlines

Part IX

Motivation: Pre-training Deep Neural Networks

Empirical observations - pre 2006:

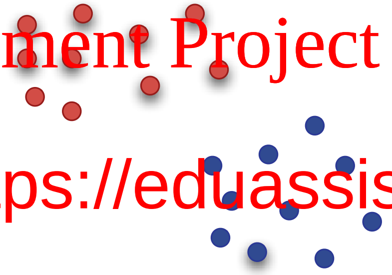
- Deep architectures get stuck in local minima or plateaus
- As a
  gen
- Har

- 1 or 2 hidden layers seem to perform better
- 2006: Unsupervised pre-training of each laye
  models possible
  - Usually based on auto-encoders (tomorro
  - Similar in spirit to PCA (today's lecture)

Given a dataset of numerical features:

- Lo
- Hig
- Dim
  -
  - Enables visualisation
  - The new basis may yield insights
  - Aside: can simplify/speed up subsequent an regression

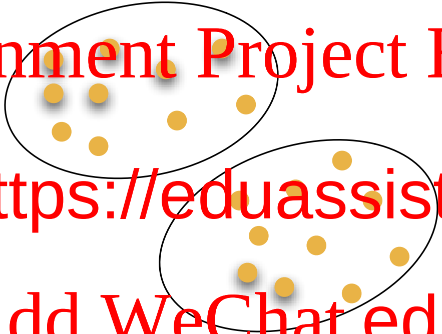
Supervised Learning



- Given are pairs of data $\mathrm{x}_\mathrm{i} \in \mathcal{X}$ and targe form $(\mathrm{x}_\mathrm{i}, t_i)$, where $i = 1 \dots N$.
- Learn a mapping between the data X and the target $\mathbf{t}$ which generalises well to new data.

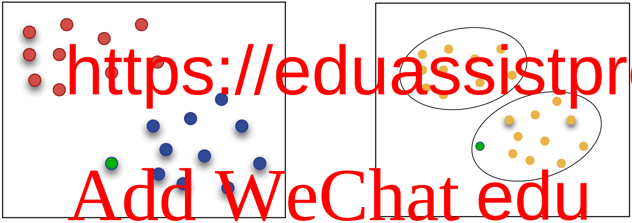
Unsupervised Learning

- Given only the data $x_i \in \mathcal{X}$.
- Discover (=learn) some interesting structure inherent in the data X.



- Given only the data $x_i \in \mathcal{X}$.
- Discover (=learn) some interesting structure inherent in the data.

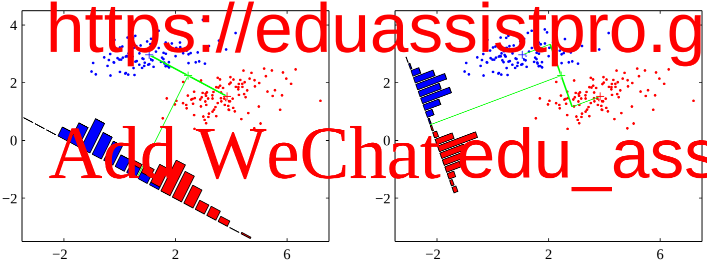
Testing - Supervised versus Unsupervised Learning



Recall: Fisher's Linear Discriminant

©2020
Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

Samples from two classes in a two-dimensional input space and their histogram when projected to two different one-dimensional spaces.



Eigenvectors

- Every square matrix $A \in \mathbb{R}^{n \times n}$ has an Eigenvector decomposition

$$Ax = \lambda x$$

wh

- Exa

$$-1 \quad 0$$

$$\lambda = \{-\imath, \imath\}$$

$$x = \left\{ \begin{bmatrix} \imath \\ 1 \end{bmatrix}, \begin{bmatrix} -\imath \\ 1 \end{bmatrix} \right\}$$

Eigenvectors

- How many eigenvalue/eigenvector pairs?
-

is eq

$$(A - \lambda I)x = 0$$

- Has only non-trivial solution for $\det \{A$
- polynom of nth order; at most n distinct

Real Eigenvalues

- How can we enforce real eigenvalues?
- Let' n n
- Tra

$$5 + \imath 6 \quad 7 + \imath 8 \quad = \quad 3 - \imath 4 \quad 7$$

- Denote the complex conjugate of a complex nu $\overline{\lambda}$.

Real Eigenvalues

- How can we enforce real eigenvalues?
- Let's assume $A \in \mathbb{C}^{n \times n}$, Hermitian ($A^H = A$).
- Calculate

$$x^H A x = \lambda x^H x$$

for an eigenvector x $\mathbb{C}^n$ of A.

- An

$$= (x^H A x)^H \qquad \text{(reverse order)}$$

$$= (\lambda x^H x)^H \qquad ($$

$$= \overline{\lambda} x^H x$$

- and therefore

$$\lambda = \overline{\lambda} \quad (\lambda \text{ is real}).$$

- If A is Hermitian, then all eigenvalues are real.
- Special case: If A has only real entries and is symmetric, then all eigenvalues are real.

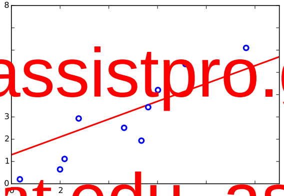
Singular Value Decomposition

Every ma n p
three mat

where U
$U^T U = I$ and $V^T V = I$), and $\Sigma \in \mathbb{R}^{n \times p}$ has n
numbers on the diagonal.

Linear Curve Fitting - Least Squares

$$N = 10$$



$$t_i \in \mathbb{R} \quad i = 1, \dots, N$$

$$y(x, \mathbf{w}) = w_1 x + w_0$$

$$X \equiv [\mathbf{x} \quad 1]$$

$$w^* = (X^T X)^{-1} X^T \mathbf{t}$$

Inverse of a matrix

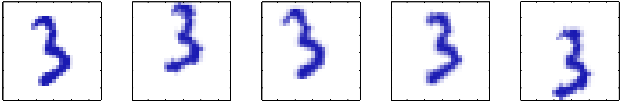
- Assume a full rank symmetric real matrix A.
- Then $A = U^T \Lambda U$ where
- Λ is a diagonal matrix with real eigenvalues
- U c

$$
\begin{aligned}
A^{-1} &= (U^T \Lambda U)^{-1} \\
&= U^{-1} \Lambda^{-1} U^{-T} \qquad \text{inverse ch} \\
&= U^T \Lambda^{-1} U \qquad U
\end{aligned}
$$

- The inverse of a diagonal matrix is the inverse of its elements.

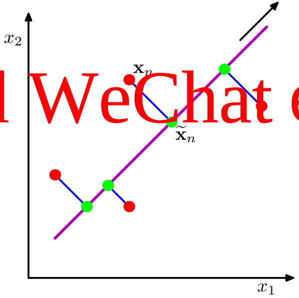
Dimensionality Reduction

- Main goal of Principal Component Analysis: dimensionality reduction
- Many applications in visualisation, feature extraction, signal processing, data compression . . .
- Example: Use hand-written digits (binary data) and place the
  and t
- Dat
- But
  rotation angle).
- FYI only: this manifold is not linear and requires bl
  edge models like capsule networks (Hinton 20
  can locally approximate with PCA.



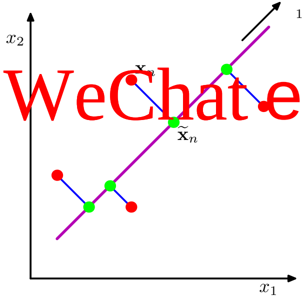
Principal Component Analysis (PCA)

- Idea: Linearly project the data points onto a lower dimensional subspace such that
  - the variance of the projected data is maximised, or
  - the distortion error from the projection is minimised.
- Both formulation lead to the same result.
- Ne
  prin



- Given N observations $\mathbf{x}_n \in \mathbb{R}^D$, $n = 1, \dots, N$.
- Project onto a space with dimensionality $M < D$ while maximising the variance.

- More advanced : How to calculate M from the data. The
- Co vec



PCA - Maximise Variance

- Each data point $\mathbf{x}_n$ is then projected onto a scalar value $\mathbf{u}_1^T \mathbf{x}_n$.
- The mean of the projected data is $\mathbf{u}_1^T \bar{\mathbf{x}}$ where $\bar{\mathbf{x}}$ is the sample mean

$$N$$

- The

$$\frac{1}{N} \sum_{n=1}^{N} \{\mathbf{u}_1^T \mathbf{x}_n - \mathbf{u}_1^T \bar{\mathbf{x}}\}^2 =$$

with the covariance matrix

$$\mathbf{S} = \frac{1}{N} \sum_{n=1}^{N} (\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})^T.$$

PCA - Maximise Variance

- Maximising $\mathbf{u}_1^T \mathbf{S} \mathbf{u}_1$ under the constraint $\mathbf{u}_1^T \mathbf{u}_1 = 1$ (why do we need to bound $\mathbf{u}_1$?) leads to the Lagrange equation

PCA - Maximise Variance

- Maximising $\mathbf{u}_1^T \mathbf{S} \mathbf{u}_1$ under the constraint $\mathbf{u}_1^T \mathbf{u}_1 = 1$ (why do we need to bound $\mathbf{u}_1$?) leads to the Lagrange equation

whi

PCA - Maximise Variance

- Maximising $\mathbf{u}_1^T \mathbf{S} \mathbf{u}_1$ under the constraint $\mathbf{u}_1^T \mathbf{u}_1 = 1$ (why do we need to bound $\mathbf{u}_1$?) leads to the Lagrange equation

whi
wit λ_1

$$\mathbf{S}\mathbf{u}_1 = \lambda_1 \mathbf{u}_1 .$$

- The variance is then $\mathbf{u}_1^T \mathbf{S} \mathbf{u}_1 = \lambda_1$.
- Variance is maximised if $\mathbf{u}_1$ is the eigenv
covariance $\mathbf{S}$ with the largest eigenvalue.

PCA - Maximise Variance

- Continue maximising the variance amongst all possible directions orthogonal to those already considered.
- The
  for w
  eig
  corresponding to the M largest eigenvalues $\lambda_1, \ldots, \lambda_M$.

- Is this subspace always uniquely defined?

- Not if $\lambda_M = \lambda_{M+1}$.

- The distortion between data points $\mathbf{x}_n$ and their projection $\widetilde{\mathbf{x}}_n$

$$J = \frac{1}{N} \sum_{n=1}^{N} \left\| \mathbf{x}_n - \widetilde{\mathbf{x}}_n \right\|^2$$

  is mi

- The

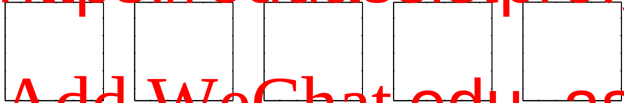
$$J = \sum_{i=M+1} \lambda_i$$

  where λ_i, $i = M+1, \dots, D$ are the smalle
  the covariance matrix $\mathbf{S}$.

- In signal processing we speak of the signal space (principal subspace) and the noise space (orthogonal to the principal subspace).

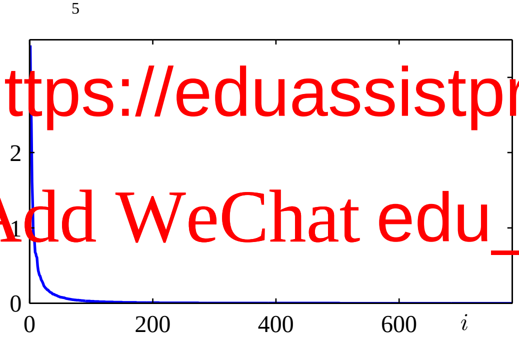
PCA - Applications

- The eigenvectors of the covariance matrix are elements of the original vector space $u_i \in \mathbb{R}^D$.
- If the input data are images, the eigenvectors are also images.



The mean and the first four eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_4$ of a set of handwritten digits of 'three'.
Blue corresponds to positive values, white is zero and yellow corresponds to negative values.

- The eigenvalues of the covariance matrix express the variance of the data set in the direction of the corresponding eigenvectors.

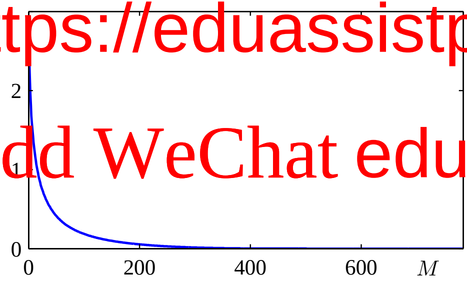


(a)

Plot of the eigenvalue spectrum for the digits of three data set.

- The sum of the eigenvalues of the covariance matrix of the discarded directions express the distortion error.

$$J = \frac{1}{N} \sum_{n=1}^{N} \| \mathbf{x}_n - \tilde{\mathbf{x}}_n \|^2$$



(b)

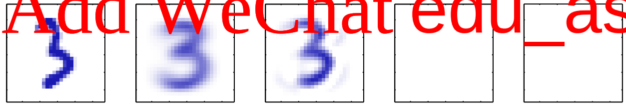
Plot of the distortion error versus the number of dimension of the subspace considered for projection.

PCA - Compression

- The approximated data vector $\widetilde{\mathbf{x}}_n$ can be written in the form

$$\widetilde{\mathbf{x}}_n = \bar{\mathbf{x}} + \sum_{i=1}^{M} \left(\mathbf{u}_i^T (\mathbf{x}_n - \bar{\mathbf{x}}) \right) \mathbf{u}_i$$

- Co
- Co



Reconstruction of an image retaining M principal components.

- Standardise certain features of a data set (for instance as a preprocessing step to subsequent algorithms expecting these features).
- Usually, individual standardisation: each variable (dimension) has zero mean and unit variance. But vari
- PC the i

- Wri

$$\mathbf{S}\mathbf{U} = \mathbf{U}\mathbf{L}$$

where $\mathbf{L}$ is the diagonal matrix of (positive!) ei
- Transform the original data by

$$\mathbf{y}_n = \mathbf{L}^{-1/2}\,\mathbf{U}^T(\mathbf{x}_n - \bar{\mathbf{x}})$$

- The set $\{\mathbf{y}_n\}$ has mean zero and covariance given by the identity.

PCA - Data Preprocessing

- Transform the original data by

$$\mathbf{y}_n = \mathbf{L}^{-1/2}\,\mathbf{U}^T(\mathbf{x}_n - \bar{\mathbf{x}})$$

- Mean of the set $\{\mathbf{y}_n\}$

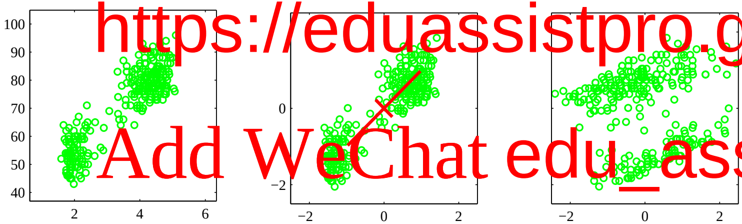
$$\frac{1}{N}\sum^{N} \quad \frac{1}{N}\sum^{N} \quad {}^{1/2} \quad {}^{T} \quad \sum_{n=1}$$

- Covariance of the set $\{\mathbf{y}_n\}$

$$\begin{aligned}\frac{1}{N}\sum_{n=1}^{N}\mathbf{y}_n\mathbf{y}_n^T &= \frac{1}{N}\sum_{n=1}^{N}\mathbf{L}^{-1/2}\,\mathbf{U}^T(\mathbf{x}_n - \bar{\mathbf{x}} \quad {}_n \\ &= \mathbf{L}^{-1/2}\,\mathbf{U}^T\mathbf{S}\mathbf{U}\mathbf{L}^{-1/2} \\ &= \mathbf{L}^{-1/2}\,\mathbf{U}^T\mathbf{U}\mathbf{L}\mathbf{L}^{-1/2} \\ &= \mathbf{I}\end{aligned}$$

PCA - The Effect of Whitening

- Compare standardising and whitening of a data set.
- (b) also shows the principal axis of the normalised data set plotted as red lines over the range $\pm\lambda_i^{1/2}$.



Original data (note the different axis).

Standardising to zero mean and unit variance.

Whitening to achieve unit covariance.

Extensions of PCA

- Kernel PCA
  - Use $\Phi(x)$ as features, and express in terms of kernel matrix
  -

- Pro
  - Explicitly model latent variable $\mathbf{z} \sim \mathcal{N}($
  - Mean value of observed variable is given by
  - Conditional distribution of observed variabl

$$\mathbf{x} \sim \mathcal{N}\left(\mathbf{x}|\mathbf{W}\mathbf{z} + \mu, \sigma\right.$$

Independence versus Uncorrelatedness

- Independence

$$p(x_1, x_2) = p(x_1)\, p(x_2)$$

- Uncorrelatedness (defined via a zero covariance)

$$\mathbb{E}\left[x_1 x_2\right] - \mathbb{E}\left[x_1\right]\, \mathbb{E}\left[x_2\right] = 0$$

- Ind
- BU

- Exa $\quad 1 \quad 2$
  the set $\{(0,1), (0,-1), (1,0), (-1,0)\}$.
- Then x_1 and x_2 are uncorrelated because
  $\mathbb{E}\left[x_1\right] = \mathbb{E}\left[x_2\right] = \mathbb{E}\left[x_1 x_2\right] = 0$.

- But x_1 and x_2 are NOT independent

$$p(x_1 = 0, x_2 = -1) = \frac{1}{4}$$

$$p(x_1 = 0)\, p(x_2 = -1) = \frac{1}{2} \times \frac{1}{4}$$

Independent Component Analysis - Overview

- Assume we have K signals and K recordings, each recording containing a mixture of the signals.

- 'Cocktail party' problem : K people speak at the same time in a room, and K microphones pickup a mixture of what they say.
- Giv
mix

$$=$$

- Can we recover the original signals (Blind Sour Separation)?

- Yes, under the assumption that
    - at most one of the signals is Gaussian distributed.
    - we don't care for the amplitude (including the sign).
    - we don't care for the order of the recovered signals.
    - we have at least as many observed mixtures as signals, the matrix $\mathbf{A}$ has full rank and can be inverted.

Independent Component Analysis - Overview

- Uncorrelated variables are not necessarily independent.
- ICA maximises the statistical independence of the estimated components.
- Fin

  are maximally independent.
- Several definitions for statistical independe
- Central Limit Theorem: The distribution of a su independent random variables tends towar distribution (under certain conditions).
- FastICA algorithm.