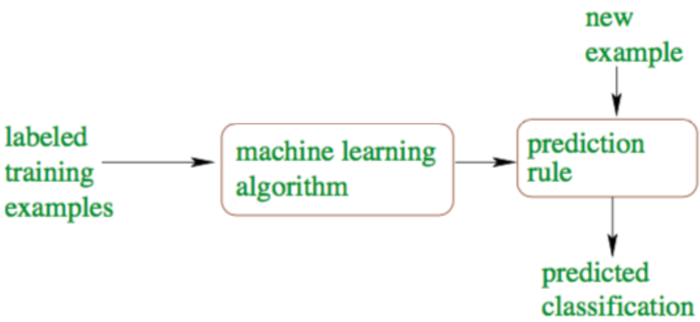
Mathematical techniques in data science

Lecture 14: Principal component analysis (PCA)

Topics

- By problems:
  - Classification
  - Regression
  - Manifold learning
  - Clustering
- By methods:
  - Classical regression-based methods
  - Tree-based methods
  - Network-based methods
- By meta-level techniques:
  - Regularization
  - Kernel methods
  - Boosting

Diagram of a typical supervised learning problem

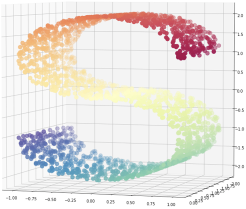
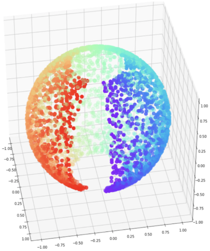


Supervised learning: learning a function that maps an input to an output based on example input-output pairs

Unsupervised learning

- Unsupervised learning
  - learning an unlabelled dataset: we observe a vector of measurements x_i but no associated response $Y^{(i)}$
  - searching for indirect hidden structures, patterns or features to analyze the data
- Problems:
  - Manifold learning
  - Clustering
  - Anomoly detection

Low dimensional structures in data

- high-dimensional data often has a low-rank structure
- Question: how can we discover low dimensional structures in data?

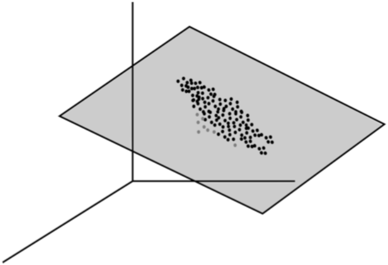
Manifold learning

- learning geometric and topological structures of high-dimensional manifolds (smooth surfaces)
- learning the low-dimensional approximation (or embedding) to visualize the dataset
- learning the mapping from high-dimensional manifold to its low-dimensional embedding

Manifold learning: methods

- Principal component analysis
- Multi-dimensional scaling (MDS)
- Locally linear embedding (LLE)
- Spectral embedding
- t-distributed Stochastic Neighbor Embedding (t-SNE)

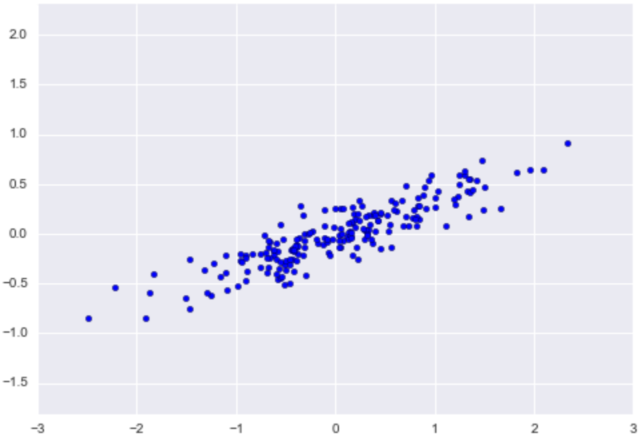
Principal component analysis



Problem: How can we discover low dimensional structures in data?

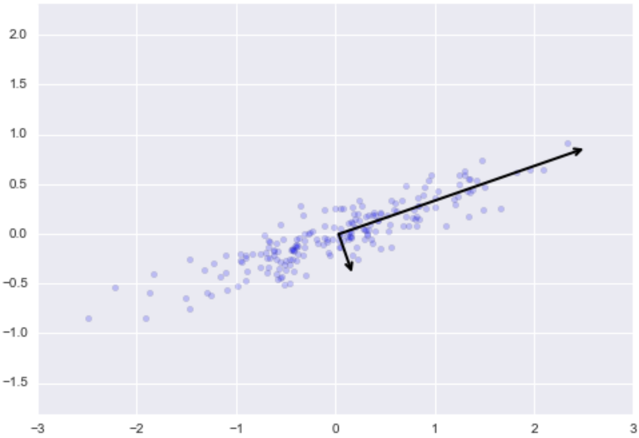
- Principal components analysis: construct projections of the data that capture most of the *variability* in the data.
- Provides a low-rank approximation to the data.
- Can lead to a significant dimensionality reduction.

PCA



PCA: first component

PCA: second component



PCA: formulation

We have a random vector X

$$X = \begin{pmatrix} X^{(1)} \\ X^{(2)} \\ \vdots \\ X^{(p)} \end{pmatrix}$$

with mean 0 and population variance-covariance matrix

$$\mathrm{var}(X) = \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_p^2 \end{pmatrix}$$

PCA: formulation

Consider the linear combinations

$$
\begin{aligned}
Y^{(1)} &= w_{11}X^{(1)} + w_{12}X^{(2)} + \cdots + w_{1p}X^{(p)} \\
Y^{(2)} &= w_{21}X^{(1)} + e_{22}X^{(2)} + \cdots + w_{2p}X^{(p)} \\
&\cdots \\
Y^{(p)} &= w_{p1}X^{(1)} + w_{p2}X^{(2)} + \cdots + w_{pp}X^{(p)}
\end{aligned}
$$

then

$$\mathrm{var}(Y^{(i)}) = \sum_{k=1}^{p}\sum_{l=1}^{p} w_{ik}w_{il}\sigma_{kl} = w_i \Sigma w_i^T$$

and

$$\mathrm{cov}(Y^{(i)}, Y^{(j)}) = \sum_{k=1}^{p}\sum_{l=1}^{p} w_{ik}w_{jl}\sigma_{kl} = w_i \Sigma w_j^T$$

PCA: formulation

- Let $X \in \mathbb{R}^{n \times p}$
- We think of X as n observations of a random vector $(X^{(1)}, X^{(2)}, \dots, X^{(p)}) \in \mathbb{R}^p$
- Suppose each column has mean 0
- We want to find a linear combination

$$\beta^{(1)} X^{(1)} + \beta^{(2)} X^{(2)} + \ldots + \beta^{(p)} X^{(p)}$$

with maximum variance.
(Intuition: we look for a direction where the data varies the most.)

PCA

- In practice, we don't know the covariance matrix $\Sigma = E[X^T X]$, and we need to approximate that by

$$\hat{\Sigma} = \mathbf{X}^\mathbf{T}\mathbf{X}$$

- We want to solve

$$w^{(1)} = \arg \max_{\|w\|=1} w\hat{\Sigma}w^T$$

- Note that

$$\sum_{i=1}^{n} |\langle x_i, w\rangle|^2 = \|\mathbf{X}w^T\|^2 = w\mathbf{X}^T\mathbf{X}w^T = w\hat{\Sigma}w^T$$

PCA: first component

- We solve
$$w^{(1)} = \arg \max_{\|w\|=1} w\hat{\Sigma}w^T$$
- Known result:
$$\max_{\|w\|=1} wAw^T = \lambda_{max}$$
where λ_{max} is the largest eigenvalue of A, and the equality is obtained if w is an eigenvector corresponding to λ_{max}

Proof

Let $A \in \mathbb{R}^{p\times p}$ be a symmetric (or Hermitian) matrix. The *Rayleigh quotient* is defined by

$$R(A,x) = \frac{x^T A x}{x^T x} = \frac{\langle Ax, x\rangle}{\langle x, x\rangle}, \qquad (x \in \mathbb{R}^p, x \neq \mathbf{0}_{p\times 1}).$$

Observations:

1. If $Ax = \lambda x$ with $\|x\|_2 = 1$, then $R(A,x) = \lambda$. Thus,
$$\sup_{x\neq \mathbf{0}} R(A,x) \geq \lambda_{\max}(A).$$

2. Let $\{\lambda_1, \dots, \lambda_p\}$ denote the eigenvalues of A, and let $\{v_1, \dots, v_p\} \subset \mathbb{R}^p$ be an orthonormal basis of eigenvectors of A. If $x = \sum_{i=1}^p \theta_i v_i$, then $R(A,x) = \frac{\sum_{i=1}^p \lambda_i \theta_i^2}{\sum_{i=1}^n \theta_i^2}$.
It follows that
$$\sup_{x\neq \mathbf{0}} R(A,x) \leq \lambda_{\max}(A).$$

Thus, $\sup_{x\neq\mathbf{0}} R(A,x) = \sup_{\|x\|_2=1} x^T A x = \lambda_{\max}(A)$.

PCA: second component

We look for a new linear combination of the Xi's that

- is orthogonal to the first principal component, and
- maximizes the variance.

In other words

$$w^{(2)} = \arg \max_{\|w\|=1; w \perp w^{(1)}} w \hat{\Sigma} w^T$$

Using a similar argument as before, we have

$$\hat{\Sigma} w^{(2)} = \lambda_2 w^{(2)}$$

where λ_2 is the second largest eigenvalue

PCA: high-order components

- We solve

$$w^{(k+1)} = \arg \max_{\|w\|=1; w \perp w^{(1)}, \ldots, w^{(k)}} w \hat{\Sigma} w^T$$

- Using the same arguments as before, we have

$$\hat{\Sigma} w^{(k+1)} = \lambda_{k+1} w^{(k+1)}$$

where λ_{k+1} is the $(k+1)^{th}$ largest eigenvalue

PCA: summary

In summary, suppose

$$X^T X = U \Lambda U^T$$

where $U \in \mathbb{R}^{p\times p}$ is an orthogonal matrix and $\Lambda \in \mathbb{R}^{p\times p}$ is diagonal. (Eigendecomposition of $X^T X$.)

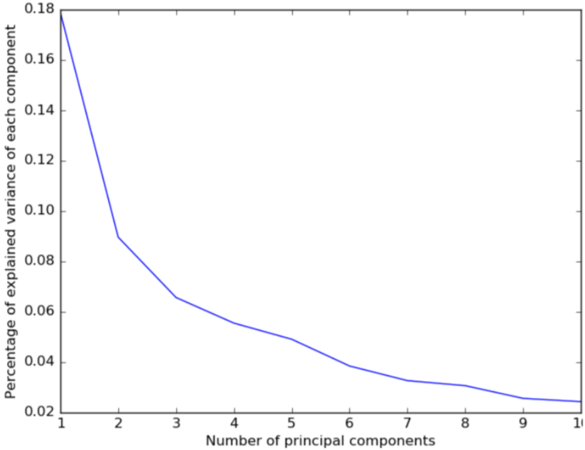
- Recall that the columns of U are the eigenvectors of $X^T X$ and the diagonal of Λ contains the eigenvalues of $X^T X$ (i.e., the (square of the) singular values of X).
- Then the *principal components* of X are the columns of XU.
- Write $U = (u_1, \dots, u_p)$. Then the variance of the i-th principal component is

$$(Xu_i)^T (Xu_i) = u_i^T X^T X u_i = (U^T X^T X U)_{ii} = \Lambda_{ii}.$$

Conclusion: The variance of the i-th principal component is the i-th eigenvalue of $X^T X$.

- We say that the first k PCs *explain* $(\sum_{i=1}^k \Lambda_{ii})/(\sum_{i=1}^p \Lambda_{ii}) \times 100$ percent of the variance.

PCA: summary


Percentage of explained variance of each component
0.18
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
1
2
3
4
5
6
7
8
9
Number of principal components