Dimensionality Reduction

Jakramate Bootkrajang

10 Aug 2014

Outlines

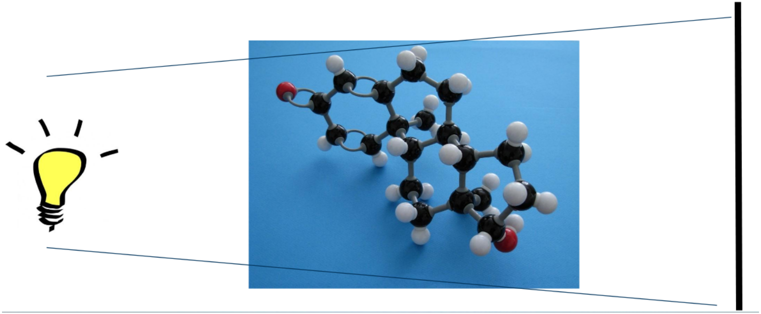
- Principle Component Analysis (PCA)
- Feature Subset Selection
- Random Projection

PCA: Goals

- Reduce dimensionality of the data while trying to preserve data structure.

PCA: Intuition

- Find low-dimensional projection with largest spread



Credit: Applied Multivariate Statistics: ETZ

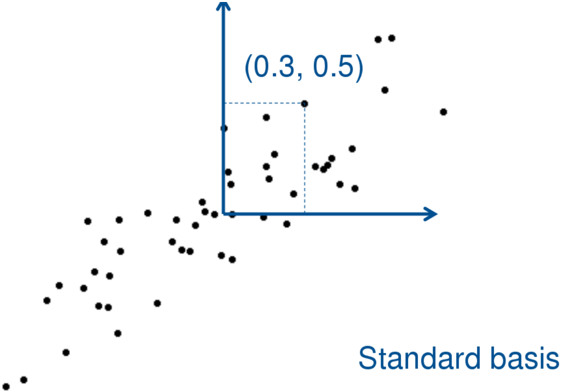
PCA: Intuition

- Our data

Credit: Applied Multivariate Statistics: ETZ

PCA: Intuition

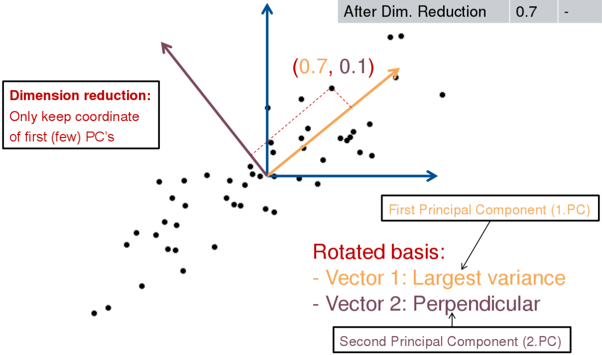
- In the standard basis $\{[0 \ \ 1], [1 \ \ 0]\}$
  - (A basis is a set of linearly independent vectors that, in a linear combination, can represent every vector in a given vector space.)



PCA: Intuition

- In a new basis produced by PCA

	X_1	X_2
Std. Basis	0.3	0.5
PC Basis	0.7	0.1
After Dim. Reduction	0.7	-



PCA: in practice

- Find projection directions that maximise variance
  - Subject to being uncorrelated with those already selected
- These directions are known as principle components
- They form a linear basis
- Hopefully we can select a few of these PCs and project the data onto this new basis.

PCA: What is a projection ?

- Projection of a vector $\mathbf{x} = (x_1, x_2, \ldots, x_m)^T$ on to a *unit vector* $\mathbf{a} = (a_1, a_2, \ldots, a_m)^T$ is the linear combination

$$\mathbf{a}^T\mathbf{x} = \sum_{i=1}^{m} a_i x_i$$

- To generalise the above, the projection of a data point $\mathbf{x}$ onto a *standard basis* A is then

$$A^T\mathbf{x} = \begin{bmatrix} a_1 & a_2 & \ldots & a_m \\ b_1 & b_2 & \ldots & b_m \\ c_1 & c_2 & \ldots & c_m \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix} = \begin{bmatrix} x'_1 \\ x'_2 \\ x'_3 \end{bmatrix}$$

which $\mathbf{x}'$ is an 3×1 vector, compared to the original $\mathbf{x}$: $m \times 1$ vector

PCA: Theory

- Goal: find projection directions that maximise variance
- Assumed that X and $\mathbf{a}X$ are mean-centred.

$$\sigma_{\mathbf{a}}^2 = (\mathbf{a}X)(\mathbf{a}X)^T \quad (1)$$
$$= \mathbf{a}XX^T\mathbf{a}^T \quad (2)$$
$$= \mathbf{a}V\mathbf{a}^T \quad (3)$$

where V is the $m \times m$ covariance matrix of the data

- We see that variance is a function of both the projection direction $\mathbf{a}$ and the covariance matrix V.

PCA: Maximising the variance

- Recall:

$$\sigma_{\mathbf{a}}^2 = \mathbf{a} V \mathbf{a}^T$$

- Maximising the above makes no sense, as we can increase the variance by increasing **a**.
- We have to impose normalisation constraint on the vector **a** such that $\mathbf{a}\mathbf{a}^T = 1$
- Optimisation objective is then $u = \mathbf{a} V \mathbf{a}^T - \lambda(\mathbf{a}\mathbf{a}^T - 1)$
- λ is a Langrange multiplier imposing the needed constraint.

PCA: Maximising the variance (cont.)

- Differentiating with respect to **a** and equating to zero yields

$$\frac{\partial u}{\partial \mathbf{a}} = 2V\mathbf{a} - 2\lambda\mathbf{a} = 0 \tag{4}$$

$$V\mathbf{a} = \lambda\mathbf{a} \tag{5}$$

Eq.(5) is one type of Linear system of equations

PCA: Characteristic Equations, Eigenvectors and Eigenvalues

- $V\mathbf{a} = \lambda\mathbf{a}$ is called the Characteristic Equations.
- For an $m \times m$, real and symmetric matrix V, there are m possible solution vectors to this system.
- In our case, V is surely real and symmetric since it's the covariance matrix.
- Each of the solutions $\mathbf{a}_i$ is also known as eigenvector $\mathbf{e}_i$ of the covariance matrix V.
- Each eigenvector is associated with an eigenvalue λ_i.

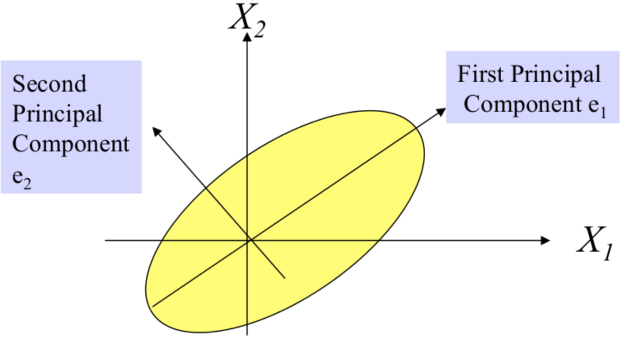
PCA: So...

- Principle Component is actually an eigenvector of the covariance matrix.
- The first principal component is the eigenvector associated with the largest eigenvalue.
- The second principal component is then the eigenvector associated with the second largest eigenvalue, and so on.
- Note that eigenvectors of V are all orthogonal to each other.
- All eigenvectors of V also form an eigen basis.

PCA: Practical usage

- Standardising **X**.
- Calculate **V**, a covariance matrix for **X**
- Find all the eigenvectors of **V**.
- Select k most important principle components put it in a matrix **E**.
- Project **X** onto **E** by calculating **XE**.

PCA: Example



Credit: Applied Multivariate Statistics: ETZ

PCA: How to choose dimensionality of the subspace?

- Recall that our eigenvector is a unit vector.
- Recall also that eigenvalue is the amount of stretching that $\mathbf{V}$ induced on $\mathbf{e}_i$.
- So from $\mathbf{X}\mathbf{e}_i = \lambda_i \mathbf{e}_i$, we see that λ_i is actually the variance of data after projecting on $\mathbf{e}_i$.
- The error of selecting k PCs and omitting the rest is then equal

$$\frac{\sum_{i=k+1}^{m} \lambda_i}{\sum_{i=1}^{m} \lambda_i}$$

- We can stop adding more PCs once the above error exceeds some predefined threshold.

PCA: Computational Issue

- $O(nm^2)+$ Complexity of solving eigenvector
- Calculating $\mathbf{V}$ takes $O(nm^2)$
- Can be applied to large dataset but *does not scale well* with dimensionality m.
- Useful SCILAB function `[d v] = eigs()`