Introduction 17/10/2007

- In today's class we will introduce principal components analysis.
- First, we will quickly introduce the sample mean and covariance matrix.
- We will re-cap on methods for constrained optimization using Lagrange multipliers.
- We will then introduce principal components analysis.

Mean and Covariance of a set of vectors I

Consider that we have a set of vectors $\{\mathbf{x}_i\}_{i=1\cdots N}$ in $\mathbb{R}^d$. We can define

- the mean $\overline{\mathbf{x}}$ such that

$$\overline{\mathbf{x}} = \frac{\sum_{i=1}^{N} \mathbf{x}_i}{N}$$

spatially the mean can be understood as the center of gravity of the clouds of points $\{\mathbf{x}_i\}_{i=1\cdots N}$.

- the covariance matrix:

$$\mathrm{C} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T$$

Mean and Covariance of a set of vectors II

with defining $\tilde{\mathbf{x}}_i = \mathbf{x} - \overline{\mathbf{x}}$, $\forall i$ then

$$\mathrm{C} = \frac{1}{N}\sum_{i=1}^{N}\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^T$$

or

$$\mathrm{C} = \frac{1}{N}\begin{bmatrix} \sum_{i=1}^{N}\tilde{x}_{1i}^2 & \sum_{i=1}^{N}\tilde{x}_{1i}\tilde{x}_{2i} & \cdots & \sum_{i=1}^{N}\tilde{x}_{1i}\tilde{x}_{di} \\ \sum_{i=1}^{N}\tilde{x}_{1i}\tilde{x}_{2i} & \sum_{i=1}^{N}\tilde{x}_{2i}^2 & & \\ & & \ddots & \\ & & \ddots & \\ & & & \sum_{i=1}^{N}\tilde{x}_{di}^2 \end{bmatrix}$$

The Lagrangian I

Definition

We consider the optimization problem:

$$\begin{array}{lll} \text{minimize} & f_0(\mathbf{x}) & \\ \text{subject to} & f_i(\mathbf{x}) = 0 & i = 1,\cdots,m \\ & h_j(\mathbf{x}) \leq 0 & j = 1,\cdots,p \end{array}$$

with $\mathbf{x} \in \mathbf{R}^d$.

The Lagrangian $\mathscr{L} : \mathbb{R}^d \times \mathbb{R}^m \times \mathbb{R}^p$ associated with the problem is defined as:

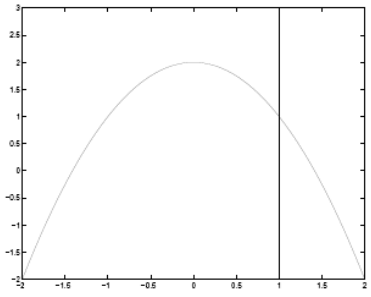
$$\mathscr{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = f_0(\mathbf{x}) + \sum_{i=1}^{m}\lambda_i f_i(\mathbf{x}) + \sum_{j=1}^{p}\nu_j h_j(\mathbf{x})$$

The vectors $\boldsymbol{\lambda}$ and $\boldsymbol{\nu}$ are called the Lagrange multiplier vectors.

The Lagrangian II

Example

Maximize the value $2 - x^2$ while satisfying $x - 1 = 0$



The Lagrangian III

We are mainly interested in minimizing functions with equality constraints.
Differentiating the Lagrangian $\mathscr{L}(\mathbf{x}, \boldsymbol{\lambda})$ in $\mathbf{x}$ gives d equations and differentiating the Lagrangian $\mathscr{L}(\mathbf{x}, \boldsymbol{\lambda})$ in $\boldsymbol{\lambda}$ gives m equations.
Solving the optimization problem then become solving

$$\begin{cases} \frac{\partial \mathscr{L}}{\partial \mathbf{x}} = 0 \\ \frac{\partial \mathscr{L}}{\partial \boldsymbol{\lambda}} = 0 \end{cases}$$

Determining Principal components I

Consider that we have a set of vectors $\{\mathbf{x}_i\}_{i=1\cdots N}$ in $\mathbb{R}^d$ arranged in a matrix X:

$$\mathrm{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,N} \\ \vdots & & & \vdots \\ x_{d,1} & x_{d,2} & \cdots & x_{d,N} \end{bmatrix}$$

we are looking in a direction or vector $\mathbf{v} \in \mathbb{R}^d$ such that the projections of $\{\mathbf{x}_i\}_{i=1\cdots N}$ on $\mathbf{v}$ leads to the scatter of N points with the highest dispersion.

Determining Principal components II

- The projection of $\mathbf{x}_i$ on to $\mathbf{v}$ is $\mathbf{v}\mathbf{v}^T\mathbf{x}_i$.
- The distance between two projections is

$$\begin{aligned} \|\mathbf{v}\mathbf{v}^T\mathbf{x}_i - \mathbf{v}\mathbf{v}^T\mathbf{x}_j\|^2 &= (\mathbf{v}\mathbf{v}^T\mathbf{x}_i - \mathbf{v}\mathbf{v}^T\mathbf{x}_j)^T(\mathbf{v}\mathbf{v}^T\mathbf{x}_i - \mathbf{v}\mathbf{v}^T\mathbf{x}_j) \\ &= \mathbf{v}^T(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T\mathbf{v} \end{aligned}$$

- Considering all the vectors $\{\mathbf{x}_i\}_{i=1\cdots N}$, the criterion to be maximized is:

$$\begin{aligned} \mathcal{J}(\mathbf{v}) &= \frac{1}{2N^2}\sum_{i=1}^N\sum_{j=1}^N \mathbf{v}^T(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T\mathbf{v} \\ &= \mathbf{v}^T\mathrm{C}\mathbf{v} \end{aligned}$$

Determining Principal components III

so the problem can be summarized as finding $\mathbf{v}$ such that:

$$\begin{cases} \max_{\mathbf{v}} \mathcal{J}(\mathbf{v}) \\ \text{subject to } \mathbf{v}^T\mathbf{v} = 1 \end{cases}$$

Using the Lagrange multipliers, the equivalent problem is:

$$\max \mathcal{J}(\mathbf{v}) - \lambda(\mathbf{v}^T\mathbf{v} - 1)$$

Determining Principal components IV

The solution is found by solving $\frac{\partial \mathcal{J}(\mathbf{v})}{\partial \mathbf{v}} = 0$ and $\frac{\partial \mathcal{J}(\mathbf{v})}{\partial \lambda} = 0$:

$$\begin{cases} \mathrm{C}\mathbf{v} - \lambda\mathbf{v} = 0 \\ \mathbf{v}^T\mathbf{v} = 1 \end{cases}$$

So $\mathbf{v}$ is an eigenvector of C, and $\mathcal{J}(\mathbf{v}) = \lambda$. Then to get the biggest dispersion we should choose the eigenvector associated with the highest eigenvalue. Hence the name of this method principal component analysis.

This result can be generalized: the k principal directions are the k eigendirections of the highest eigenvalues.

Determining Principal components V

Theorem (Principal Component Analysis)

From a set of vectors $\{\mathbf{x}_i\}$

1. *Compute the mean* $\bar{\mathbf{x}}$
2. *Center each observations* $\tilde{\mathbf{x}}_i = \mathbf{x}_i - \bar{\mathbf{x}}$
3. *Compute the covariance matrix*

$$\mathrm{C} = \frac{1}{N}\sum_{i=1}^N \tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^T$$

4. *Compute the eigenvectors of* C *and sort them from the one associated with the highest eigenvalue , to the one associated with the lowest eigenvalue.*

Using PCA

Each vector in the set can be written as a linear combination of the mean and the eigenvectors:

$$\mathbf{x} = \bar{\mathbf{x}} + \sum_j \alpha_j \mathbf{v}_j$$

How many eigenvectors really needed?

Applications:

- Compression / Dimensionality reduction
- Visualisation of data distribution
- etc.