T-61.183 Support Vector Machines and Kernel Methods

Kernel Principal Component Analysis

Anne Patrikainen

7.4.2003

Classical PCA and Kernel PCA

- Principal Component Analysis (PCA) is a technique for extracting structure from data sets
- PCA projects the N-dimensional data into a N'-dimensional subspace such that the variance of the projected data is maximized
- This subspace is spanned by so called *principal components*
- Kernel PCA does the same to a data that is non-linearly mapped into a high-dimensional feature space
- As usual, we do not explicitly map the data into the feature space, but employ the kernel trick instead

How to perform classical PCA

- Consider a set of observations $\{x_j\}_{j=1}^m$ in R^N
- The observations have to be centered: $\sum_{j=1}^m x_j = 0$
- Sample covariance matrix $C = \frac{1}{m}\sum_{j=1}^m x_j x_j^T$
- The principal components v^i are the eigenvectors of the sample covariance matrix C
- That is, $Cv^i = \lambda_i v^i$, where λ_i is the corresponding eigenvalue
- Finally, a point x_j is projected onto the principal component v^i as $< x_j, v^i >$
- Note that all solutions v^i lie in the span of the data points $x_1, \ldots, x_m$: $\lambda v^i = Cv^i = \frac{1}{m}\sum_{j=1}^m < x_j, v^i > x_j$

How to perform Kernel PCA

- Let us consider a (non-linear) mapping Φ of the (high-dimensional) input space X into the feature space H
- In the feature space, the sample covariance matrix takes the form $C = \frac{1}{m} \sum_{j=1}^{m} \Phi(x_j)\Phi(x_j)^T$
- In principle, we proceed analogously to classical PCA: we calculate the eigenvectors v^i of C and project the mapped data points onto them
- In practice, the calculations are a bit different: we solve the dual eigenvalue problem

Kernel PCA, technical details

- As in the linear case, the eigenvectors lie in the span of the mapped data points $\Phi(x_1), \ldots, \Phi(x_m)$
- That is, $v = \sum_{j=1}^{m} \alpha_j \Phi(x_j)$
- We collect the coefficients up into a vector $\alpha = (\alpha_1, \ldots, \alpha_m)^T$
- The dual eigenvalue problem: $m\lambda\alpha = K\alpha$
- K is the gram matrix with elements $K_{ij} =< \Phi(x_i), \Phi(x_j) >$
- We solve the dual eigenvalue problem and obtain m eigenvectors α^i and m eigenvalues λ_i
- A data point $\Phi(x)$ is projected onto the principal component v^i as $< \Phi(x), v^i >= \sum_{j=1}^{m} a_j^i k(x_j, x)$

Comparison of some properties of classical PCA and Kernel PCA

- Assume that the dimension of the input space is N and that the dimension of the feature space is P, where $N < P$
- In classical PCA, the size of the covariance matrix C is $N \times N$, and we get N eigenvectors (principal components) of size $N \times 1$.
- In Kernel PCA, C is a $P \times P$ matrix and we get P eigenvectors (more) of size $P \times 1$ (bigger)
- We obtain a better reconstruction of the data points by using more principal components
- However, there is no guarantee that we are able to find the pre-images of partially constructed patterns in the original space (See Chapter 18)

Comparison of Kernel PCA with related methods

- Other generalizations of linear PCA: Hebbian Networks, Autoassociative Multi-Layer Perceptrons, Principal Curves
- Advantages of Kernel PCA:
  - No nonlinear optimization is involved
  - The number of principal components does not need to be specified in advance
- Disadvantages of Kernel PCA:
  - Compared with neural approaches, processing very large number of observations is much more difficult (matrix K is large)
  - Compared with the Principal Curves approach, the results are harder to interpret in the input space

Other related methods

- Kernel PCA with RBF (Radial Basis Function) kernels is related to metric MDS (MultiDimensional Scaling)
- The solution of LLE (Locally Linear Embedding) is obtained by performing kernel PCA with the locally linear embedding kernel
- Kernel PCA is also related to Orthogonal Series Density Estimation

Results on real-world data: USPS character recognition

- Handwritten character recognition problem using the US postal service database
- The dimension of the input space is 256
- Kernel PCA with the polynomial kernel $k(x, x') =< x, x' >^d$ was used to extract features from the training set
- A linear soft margin hyperplane classifier was trained on these features

USPS character recognition, continued

- Advantages of Kernel PCA over classical (linear) PCA:
  - For the same number of principal components, Kernel PCA performs better
  - There are more principal components (features) in the non-linear case
- The best results are identical to those obtained with non-linear SVM classifiers
- This is because using all principal components is equivalent to running a non-linear SVM with the same kernel

Conclusions

- Kernel PCA is a generalization of classical PCA
- Kernel PCA can be applied in the same numerous domains as classical PCA, often yielding better results
- Kernel PCA can be used to cluster data, unlike classical PCA
- As the number of data points increases, computational difficulties are encountered in diagonalizing the kernel matrix
- A solution is to approximate Kernel PCA in a sparse greedy way (see Section 10.2)
- Another solution is to modify the algorithm to make it more efficient, resulting in Kernel Feature Analysis (KFA)

Problems

- **14.1**: Positive Definiteness of the Covariance Matrix (•)
- **14.3**: Pre-image problem (•)