Case Studies: Nonlinear Optimization

Fin, Stan, Stefan, Jakob, Roland

21. April 2015

Inhaltsverzeichnis

Notation
Math environment in aligned mode:

$$\begin{aligned} f(x) &= x \\ &= e^{\log(x)} \end{aligned}$$

The numbered versions of this is

$$\begin{aligned} f(x) &= x \\ &= e^{\log(x)} \end{aligned} \tag{0.1}$$

Citations work this way: (0.1)
Preprocessing
Curse of Dimensionality: ?
Postulates: 1. Samples: Representative Samples 2. Features: A classifier is as good as its features 3. Compactness: Features belonging to the same class occupy a compact area in the feature space, and the different classes are reasonably separable
-> Low intra-class-distance, high inter-class-distance
4. Decomposition: Complex patterns can be decomposed into smaller parts, their combined presence makes up the pattern 5. Sturcture: 6. Similarity: Two representations are similar if a proper distance measure is small for them

1 Principal Component Analysis (PCA)

The PCA stems from Linear Algebra. It is relevant to know that if I have got two different vector spaces U and W with basis $\{u\}_j \in \mathrm{R}^m$ and $\{w\}_i \in \mathrm{R}^n$, respectively, I there is a matrix A, such that

$$w_i = \sum_j a_{i,j} v_j.$$

If $m = n$, by writing a vector x in both bases, we get that

$$\sum_j x_j v_j = x = \sum_i \sum_j a_{i,j} y_i v_j$$

with new coordinates y_i. So, we can say: old coordinates $= A$ new coordinates

$$\text{new coordinates } = A^{-1} \text{ old coordinates.}$$

If the old basis is in the standard coordinate system with basis vectors $e_i \in \mathrm{R}^n$ such that $e_{i,j} = \delta_{i,j}$, then the new basis w_i is given as the rows of the matrix A. Conversely, if I want to have a transform into the basis w_i, all I need to do, is to put the new basisvectors as rowvectors of the transformation matrix A. If, furthermore, the new basis is orthonormal, then I can also write the basis vectors as columns of A and do the transformation by applying $A^{-1} = A^T$.
There is a theorem which says that to every real valued matrix $A \in \mathrm{R}^{m \times n}$ there exists a Singular Value Decomposition (SVD) with matrices $U \in \mathrm{R}^{m \times m}$, $V \in \mathrm{R}^{n \times n}$ and $\Sigma \in \mathrm{R}^{m \times n}$, such that

$$A = U \Sigma V^T$$

The two matrices U and V are orthogonal matrices containing the eigenvectors as columns of A. Both, the old, and the new basis vectors. Furthermore, the matrix $\Sigma = diag(\sigma_1, \ldots, \sigma_m, 0, \ldots 0)$ consists of the so called "Singular Values" on its diagonal.
For a set of feature vectors $\{X_i\}_{i=1}^M$, the SVD is now applied to the covariance matrix $C \in \mathrm{R}^{N \times N}$ with mean value of the training data $\bar{\mu}$.

$$C = \frac{1}{M-1} \sum_{i=1}^{M} (X_i - \bar{\mu})(X_i - \bar{\mu})^T, \quad \bar{\mu} = \frac{1}{M} \sum_{i=1}^{M} X_i,$$

Since the matrix C is symmetric and real, according to the theorem of the "Hauptachsentransfromation", there exists a orthogonal matrix $U \in \mathrm{R}^{N \times N}$ such that

$C = UDU^T$, with diagnonal matrix Σ containing the eigenvalues of C.

This means, that the covariance matrix is preserving the coordinate system. Moreover, the matrix $U^{-1} = U^T$ transformes any vector $c \in \mathrm{R}^N$ into the basis of the eigenvectors.
If we use the SVD in order to compute the Hauptachsentransfromation of C, we get a basis of eigenvectors such that the eigenvalues are a set of ordered non-negative real numbers.

$$C = UDU^T = \left(UD^{\frac{1}{2}}\right)\left(UD^{\frac{1}{2}}\right)^T, \text{ and}$$
$$C^{-1} = \left(UD^{\frac{1}{2}}\right)^{-T}\left(UD^{\frac{1}{2}}\right)^{-1} = \left(D^{-\frac{1}{2}}U^T\right)^T\left(D^{-\frac{1}{2}}U^T\right)$$

In other words: By multyplying each sample $\{X_i \in \mathrm{R}^N\}_{i=1}^M$ with the matrix $B := D^{-\frac{1}{2}}U^T$, we get a normalized set of vectors with zero mean and uniform covariance.

1.1 Derivation of Kernel PCA

Assume the data set $\{X_i\}_{i=1}^M$ is of zero mean. Let $\phi : \mathrm{R}^N -> \mathrm{R}^H$ be a nonlinear lifting transform.

Thus, the covarianve matrix simplifies to $C = \frac{1}{M-1} \sum_{k=1}^{M} X_k X_k^T$. Let $\{v_i\}_{i=1}^N$ be the set of eigenvectors of C with corresponding eigenvalues $\{\lambda_i\}_{i=1}^N$. For each i, we can write v_i as a linear combination of samples:

$$v_i = \sum_{j=1}^{M} \alpha_{i,j} X_j.$$

Now we rewrite the eigenvalue problem using this equality:

$$\sum_{k,j} \alpha_{i,j} X_k X_k^T X_j = (M-1)\lambda_i \sum_j \alpha_{i,j} X_j$$

Considering the sum of projections of all samples onto the eigenvector v_i, this equation plugs into

$$\sum_{l}^{M} X_l^T v_i = \sum_{l,k,j} X_l^T X_k X_k^T X_j \alpha_{i,j} = (M-1)\lambda_i \sum_{l,j} X_l^T X_j \alpha_{i,j}$$

It is now time to introduce the **kernel matrix** K with $K_{i,j} = k(X_i, X_j) = \phi(X_i)^T \phi(X_j)$ with kernel k and feature transform ϕ.

$$\sum_{l,j} \sum_{k} K_{l,k} K_{k,j} \alpha_{i,j} = \sum_{l,j} K_{l,j}^2 \alpha_{i,j} = (M-1)\lambda_i \sum_{l,j} K_{l,k} \alpha_{i,j}$$

In matrix notation, we get the **kernel eigenvalue problem**

$$K^2 \alpha_i = (M-1)\lambda_i K \alpha_i \Leftrightarrow K\left[K\alpha_i - (M-1)\lambda_i \alpha_i\right] = 0 \Leftrightarrow K\alpha_i = (M-1)\lambda_i \alpha_i$$

which tells us that all α_i can be obtained as an eigenvector of the kernel matrix K. Writing the transformed samples $\{X_k\}_k^M$ in the basis of the v_i, we can get the transformed values by means of

$$\phi(X) = \sum_{i}^{N} \phi(X)^T v_i = \sum_{i}^{N} \sum_{j}^{M} \alpha_{i,j} \phi(X)^T \phi(X_j) = \sum_{i}^{N} \sum_{j}^{M} \alpha_{i,j} k(X, X_j)$$

2 Optimal Feature Transform

We want to consider feature transforms Φ which operate on the digital signal f. The resulting feature vector $c = \Phi f$ is later used for classification.
One of the aims in this section is to find an optimal feature transform. This can be a very challenging task, however. Therefore we first consider feature transforms which reduce the dimension of the feature space by projection on one or more axes, which is also called Linear Discriminant Analysis.

3 PCA

Input: A set of K vectors $X_i \in \mathrm{R}^n$, $i \in \{1, \dots K\}$ with is normalized, i.e.

$$\frac{1}{K} \sum_{i=1}^{K} X_i = 0.$$

Task: Find a linear orthogonal transformation $\Phi : \mathrm{R}^n -> \mathrm{R}^m$ such that the spread of the features ΦX_i around the zero mean is maximized:

$$\max \left\{ \sum_{i=1}^{K} \|\Phi X_i\| - \sum_{k=1}^{m} \lambda_k (\|\phi_k\|_2 - 1) \right\}$$

where $\|\Phi\|_2$ is the Frobenius norm. The first part in the objective function is the spread of the transformed features, the second part poses a boundary condition on the length of the matrix vectors.
We can rearrange the first part:

$$\begin{aligned} \sum_{i=1}^{K} \|\Phi X_i\| &= \sum_{i=1}^{K} X_i^T \Phi^T \Phi X_i = \sum_{i=1}^{K} X_i^T \left(\sum_{k=1}^{m} \phi_k \phi_k^T \right) X_i \\ &= \sum_{k=1}^{m} \phi_k^T \left(\sum_{i=1}^{K} X_i X_i^T \right) \phi_k =: \sum_{k=1}^{m} \phi_k^T \Sigma \phi_k \end{aligned}$$

with $\Sigma = \sum_{i=1}^{K} X_i X_i^T$, K-times the covariance matrix.
At the maximum point we have that the derivitives with repect to ϕ are zero, i.e. for all k between 1 and M

$$\begin{aligned}
&\frac{\partial}{\partial \phi_k} \sum_{k=1}^{m} \phi_k^T \Sigma \phi_k - \lambda_k (\|\phi_k\|_2 - 1) = 0 \\
&\Leftrightarrow 2\Sigma\phi_k - 2\lambda_k \phi_k = 0 \\
&\Leftrightarrow \Sigma\phi_k = \lambda_k \phi_k
\end{aligned}$$

PCA

1. Normalize the data to zero mean.
2. Compute the covariance matrix Σ of $\{X_i\}_{i=1}^{K}$.
3. Determine the eigenvalues and eigenvectors of Σ using SVD
4. Order the eigenvalues and eigenvectors in descending order.
5. The higher the eingenvalue, the more important is the eigenface.
6. For classification represent an normalized image as a linear combination of eigenvectors.
7. Use the weights as feature vectors for classification.
8. Use any classification algorithm to learn.

4 LDA

Let N be the number of classes and m be total number of feature vectors available. Then we define the absolute inter-class-distance and the absolute intra-class-distance as follows:

$$\Sigma_{\text{inter}} = \sum_{k}^{N} \sum_{l \neq k}^{N} \sum_{i}^{m} \sum_{j}^{m} \left\| c_i^k - c_j^l \right\|^2, \quad \Sigma_{\text{intra}} = \sum_{k}^{N} \sum_{i}^{m} \sum_{j}^{m} \left\| c_i^k - c_j^k \right\|^2$$

Rayleigh Quotient

$$a^* = \arg\max_a \frac{a^T \Sigma_{\text{inter}} a}{a^T \Sigma_{\text{intra}} a}$$

A Appendix