Wissenschaftliches Rechnen II/Scientific Computing II

Sommersemester 2016
Prof. Dr. Jochen Garcke
Dipl.-Math. Sebastian Mayer

universitätbonn

Exercise sheet 11 To be handed in on **Thursday, 07.06.2016**

Application of PCA and MDS

1 Group exercises

G 1. (MDS: embedding of out-of-sample data)

Assume you had training data in the form of a centered Gram matrix $G^c = (\langle y^i, y^j \rangle)_{i,j=1}^n = Y^T Y$ or in the form of a Euclidean distance matrix $D = (\|y^i - y^j\|)_{i,j=1}^n$ and learned a p-dimensional embedding of the training data using the CMDS algorithm. Now assume there is a new test point $x \in \mathbb{R}^d$, which is different from the y^i but stems from the same data generating source. You cannot observe x directly but only one of the following sets of features:

a) you either observe inner products $x_S = Y^T x$,

b) or you observe squared Euclidean distances $x_E = (\|x - y^i\|_2^2)_{i=1}^n$.

Use the components computed by the CMDS algorithm to construct a p-dimensional embedding $\hat{x}$ of x from the given feature representation. Give a geometric interpretation of the constructed embedding $\hat{x}$. Discuss what properties the training data $y^1, \ldots, y^n$ must have such that the obtain embedding $\hat{x}$ is reasonable.

Solution. I give a more detailled explanation, since some of you still struggled with how to interpret the embeddings generated by MDS/PCA. Let us first recapitulate how the training data is embedded. Consider the singular value decomposition $Y = U\Sigma V^T$. We have seen in the lecture that CMDS orthogonaly projects the data Y on the p-dimensional subspace spanned by the the columns of $W = UI_{d\times p}$, where W is such that the sum of squared errors (lengths of orthogonal complements)

$$\|Y - WW^T Y\|_F^2 = \sum_{i=1}^n \|y^i - WW^T y^i\|_2^2$$

becomes minimal. The embedding of the training data is given by $W^T Y = I_{p\times d} U^T Y = I_{p\times d}\Sigma V^T$. Note that $\Sigma = \Lambda^{1/2}$, where $Y^T Y = V\Lambda V^T$.
It seems natural to take

$$\hat{x} = W^T x = I_{p\times d} U^T x$$

as the embedding of the new data point x. How can we justify this? To this end, let us discuss another geometric interpretation of U and W, which can be considered as the non-probabilistic version of the PCA-perspective. The training data points $y^1, \ldots, y^n$ lie in an ellipsoid E_Y, which is given by the transformation of the the Euclidean unit ball $B_2^d = \{y \in \mathbb{R}^d : \|y\|_2 \leq 1\}$ under the matrix YY^T, $E_Y = YY^T B_2^d$. Now the nonzero eigenvalues on the diagonal of Λ and the columns of U describe the *principal axes* of this

ellipsoid. Concretely, the ith column of U contains the unit vector that is aligned with the ith principal axis of E_Y and the ith eigenvalues λ_i gives the length of this principal axis.
If the given training data are representative for the data generating source, future data points will also lie in E_Y (at least, they will be close to it). Hence, if we assume that the given data $y^1, \dots, y^n$ is representative, then $W^T x = I_{p\times d} U^T x$ will provide a good p-dimensional approximation of x in the sense that

$$\|x - WW^T x\|_2$$

is of the same magnitude as for the training data.
Now we are only left with the problem how to compute $I_{p\times d} U^T x$ as we only know $Y^T x$. But this is simple: since $Y^T = V\Sigma U^T$ we have $U^T = \Sigma^{-1} V^T Y^T$, such that

$$\hat{x} = I_{p\times d} \Sigma^{-1} V^T Y^T x.$$

In case that we only know Euclidean distances, then we use the same double centering trick which you have seen in the lecture:

$$\langle y^i, x\rangle = -\frac{1}{2}\left(\|y^i - x\|_2^2 - \frac{1}{n}\sum_{j=1}^n \|y^i - y^j\|_2^2 - \frac{1}{n}\sum_{j=1}^n \|x - y^j\|_2^2 + \frac{1}{n^2}\sum_{k,l=1}^n \|y^l - y^k\|_2^2 \right).$$

I leave it to you to write this in matrix form. □

G 2. (Kernel-MDS)

Discuss how MDS could be generalized to distances and inner products which are induced by a reproducing kernel $k : \Omega \times \Omega \to \mathbb{R}$. Concretely, assume that there are points $x_1, \dots, x_n \in \Omega$ of which you observe $(k(x_i, x_j))_{i,j=1}^n$ and you want to construct embeddings $\hat{x}^1, \dots, \hat{x}^n \in \mathbb{R}^p$ such that

$$k(x_i, x_j) \approx \langle \hat{x}^i, \hat{x}^j \rangle.$$

Solution. Consider the eigenvalue decomposition of K, $K = V\Lambda V^T$ or $\Lambda = V^T K V$. Let us denote by $V_{\cdot j}$ the jth column of V. Since $k(x_i, x_j) = \langle k(x_i \cdot), k(x_j, \cdot)\rangle_k$, we have

$$\lambda_i \delta_{ij} = V_{\cdot i}^T K V_{\cdot j} = \sum_{l_1, l_2 = 1}^n V_{l_1 i} k(x_{l_1}, x_{l_2}) V_{l_2 j} = \langle \sum_{l=1}^n V_{li} k(x_l, \cdot), \sum_{l=1}^n V_{lj} k(x^l, \cdot)\rangle_k,$$

where δ_{ij} is as usual the Kronecker delta. Assuming that K has rank n, the functions

$$f_i = \frac{1}{\sqrt{\lambda_i}} \sum_{l=1}^n V_{li} k(x_l, \cdot), \quad i = 1, \dots, n$$

are pairwise orthogonal and $\|f_i\|_k = 1$.
To understand what these functions mean, consider the *feature map*

$$\Phi : \Omega \to H, \quad x \mapsto k(x, \cdot).$$

Then Φ maps the given data points $X = \{x_1, \dots, x_n\}$ into a n-dimensional subspace of $\mathcal{H}$, namely $\mathcal{H}_X = \operatorname{span}\{k(x_i, \cdot), i = 1, \dots, n\}$. Since $f_i \in \mathcal{H}_X$, $(f_i)_{i=1}^n$ is an orthonormal basis of $\mathcal{H}_X$. As you will show in H1, this ONB has the property that $f_1, \dots, f_p$ span the optimal p-dimensional subspace of $\mathcal{H}_X$ such that

$$\sum_{i=1}^n \|k(x_i, \cdot) - \sum_{j=1}^p \langle k(x_i, \cdot), f_j\rangle f_j\|_k^2$$

becomes miminal. Note that $\langle k(x_i, \cdot), f_j \rangle = f_j(x_i)$. Hence, for sufficiently large p,

$$k(x_i, x_j) \approx \langle \sum_{l=1}^{p} \langle k(x_i, \cdot), f_l \rangle f_l, \sum_{l=1}^{p} \langle k(x_j, \cdot), f_l \rangle f_l \rangle = \sum_{l=1}^{p} f_l(x_i) f_l(x_j).$$

Thus, we choose as embedding

$$\hat{x}^i = \begin{pmatrix} f_1(x_i) \\ \vdots \\ f_p(x_i) \end{pmatrix}.$$

□

2 Homework

H 1. (Optimal p-dimensional subspace in a RKHS)

Let $k : \Omega \times \Omega \to \mathbb{R}$ be a reproducing kernel, $\mathcal{H}$ its native Hilbert space and $X = \{x_1, \dots, x_n\} \subset \Omega$. Consider the kernel matrix $K = (k(x_i, x_j))_{i,j=1}^n$ and the corresponding eigenvalue decomposition $K = V\Lambda V^T$ with $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_m, 0, \dots, 0)$, where we assume $m \leq n$. Consider for $i = 1, \dots, m$ the functions

$$f_i := \frac{1}{\sqrt{\lambda_i}} \sum_{j=1}^n V_{ji} k(x^j, \cdot) \in \mathcal{H}_X.$$

a) Show that $(f_i)_{i=1}^m$ forms an orthonormal basis of H_X. **Hint:** Use that $\Lambda = V^T K V$.

b) Let $p \in \{1, \dots, m\}$. Show that $(f_i)_{i=1}^p$ is the solution of

$$\min_{(g_i)_{i=1}^n \text{ ONB of } \mathcal{H}_X} \sum_{i=1}^n \|k(x_i, \cdot) - \sum_{j=1}^p g_j(x_i) g_j\|_k^2.$$

Hint: Argue analogously as in Sheet 10, H2 b).

(10 Punkte)

Solution.

a) Orthogonality has already been shown in G2, see the solution there. It only remains to note that since K has rank m, the subspace $\mathcal{H}_X$ is m-dimensional and hence the f_i form an ONB.

b) Let us first observe, using the symmetry of k and $\Lambda = V^T K V$,

$$\begin{aligned}\sum_{l=1}^n f_i(x_l) f_j(x_l) &= \frac{1}{\sqrt{\lambda_i \lambda_j}} \sum_{t_1,t_2=1}^n V_{t_1 i} V_{t_2 j} \sum_{l=1}^n k(x_{t_1}, x_l) k(x_l, x_{t_2}) \\ &= \sum_{t_1,t_2=1}^n V_{t_1 i} V_{t_2 j} (K^2)_{t_1 t_2} = V_{\cdot i}^T K^2 V_{\cdot j} = \sqrt{\lambda_i \lambda_j} V_{\cdot i}^T V_{\cdot j} = \lambda_i \delta_{ij}.\end{aligned}$$

Now, using the orthogonality of the g_j, we compute

$$\sum_{i=1}^n \|k(x_i, \cdot) - \sum_{j=1}^p g_j(x_i) g_j\|_k^2 = \sum_{i=1}^n k(x_i, x_i) - \sum_{i=1}^n \sum_{j=1}^p g_j(x_i)^2.$$

This sum becomes minimal when $\sum_{i=1}^n \sum_{j=1}^p g_j(x_i)^2$ is maximal. Since the f_i form an ONB, we have $k(x_i, \cdot) = \sum_{l=1}^m f_l(x_i) f_l$, which yields

$$\sum_{i=1}^n \sum_{j=1}^p g_j(x_i)^2 = \sum_{i=1}^n \sum_{j=1}^p |\langle g_j, k(x_i, \cdot)\rangle_l|^2 = \sum_{i=1}^n \sum_{j=1}^p \sum_{l_1,l_2=1}^m \langle g_j, f_{l_1}\rangle_k \langle g_j, f_{l_2}\rangle_k f_{l_1}(x_i) f_{l_2}(x_i)$$

Our first observation implies

$$\sum_{i=1}^n \sum_{j=1}^p \sum_{l_1,l_2=1}^m \langle g_j, f_{l_1}\rangle_k \langle g_j, f_{l_2}\rangle_k f_{l_1}(x_i) f_{l_2}(x_i) = \sum_{j=1}^p \sum_{l=1}^m |\langle g_j, f_l\rangle_k|^2 \lambda_l$$

Put $\alpha_l = \sum_{j=1}^p |\langle g_j, f_l\rangle_k|^2$. By Bessel's inequality, $\alpha_l \leq \|f_l\|_k^2 = 1$. Since $g_j = \sum_{l=1}^m \langle g_j, f_l\rangle_k f_l$, we further have

$$\sum_{l=1}^m \alpha_l = \sum_{l=1}^m \sum_{j=1}^p \|\langle g_j, f_l\rangle_k f_l\|_k^2 = \sum_{j=1}^p \|\sum_{l=1}^m \langle g_j, f_l\rangle_k f_l\|_k^2 = p.$$

It remains to apply Sheet 10 G2 to conclude

$$\sum_{i=1}^{n}\sum_{j=1}^{p} g_j(x_i)^2 = \sum_{l=1}^{m} \alpha_l \lambda_l \leq \sum_{l=1}^{p} \lambda_l.$$

We have equality if $g_i = f_i$ for $i = 1, \ldots, p$.

□

H 2. (Programming exercise: pedestrian classification)

See accompanying notebook.

(10 Punkte)