Proseminar: Machine Learning

Hauptkomponentenanalyse (PCA) und Kernel-PCA

Claudia Broelemann
Betreuer: Christian Spieth, Andreas Dräger

18. Juli 2006

Inhaltsverzeichnis

1 Einleitung **3**

2 Hauptkomponentenanalyse (PCA) **3**
- 2.1 Allgemeines zu PCA . . . 3
- 2.2 Auswahl der Hauptkomponenten . . . 3
- 2.3 PCA Algorithmus . . . 4
  - 2.3.1 Beispiel Schuhkauf . . . 5
- 2.4 Multivariate Verteilung . . . 6
  - 2.4.1 Beispiel Apfelbaumplantage . . . 7
- 2.5 Bivariat verteilte Eingabedaten . . . 8

3 Kernel-PCA **9**
- 3.1 Kurze Wiederholung von Kernels . . . 9
- 3.2 Allgemeines zu Kernel-PCA . . . 10
  - 3.2.1 Die Gram-Matrix . . . 11
  - 3.2.2 Normieren der Eigenvektor-Koeffizienten . . . 12
  - 3.2.3 Berechnung der Abbildungen auf die Eigenvektoren . . . 12
- 3.3 Kernel-PCA Experiment . . . 13

4 Unterschied zwischen PCA und Kernel-PCA **14**

5 Zusammenfassung **15**
- 5.1 Anwendungsgebiete . . . 15

6 Literatur **15**

1 Einleitung

Das erste Kapitel der Ausarbeitung dient dazu, die lineare Hauptkomponentenanalyse (englisch: Principal Component Analysis (PCA)) zu erklären. Die nichtlineare Hauptkomponentenanalyse (Kernel-PCA) wird im zweiten Kapitel erläutert und das dritte Kapitel hebt die Unterschiede zwischen PCA und Kernel-PCA hervor.

2 Hauptkomponentenanalyse (PCA)

2.1 Allgemeines zu PCA

Die Hauptkomponentenanalyse ist eine variablenorientierte Methode, die bei Variablen mit vielen Eigenschaften versucht, wenige latente Faktoren zu extrahieren. Dazu werden Hauptkomponenten in absteigender Bedeutung gebildet, d. h. dass die erste Hauptkomponente für den größten Teil der Variationen verantwortlich ist.
Mathematisch betrachtet heißt das, dass eine Hautachsentransformation durchgeführt wird: Die Korrelation mehrdimensionaler Merkmale wird durch Überführung in einen Vektorraum mit neuer Basis minimiert. Aus den Eigenvektoren der Kovarianzmatrix lässt sich eine neue Matrix bilden, welche die Hauptachsentransformation angibt.
Diese Matrix muss für jeden Datensatz neu berechnet werden, wodurch die Hauptkomponentenanalyse problemabhängig wird.

2.2 Auswahl der Hauptkomponenten

Angenommen es sind m Daten – also eine Punktwolke mit m Punkten – in einem P-dimensionalen Raum gegeben. Um die Hauptkomponenten zu bilden, wird folgendes Verfahren angewandt:
Zunächst wird der Ursprung des Koordinatensystems in den Schwerpunkt der Punkt-wolke gesetzt. Als nächstes wird das Koordinatensystem gedreht, sodass die erste Koordinate in Richtung der größten Varianz der Punktwolke zeigt. Die erste Koordinate stellt somit die erste Hauptachse, die Varianz die erste Hauptkomponente dar.
Für die zweite Hauptkomponente wird das Koordinatensystem so weitergedreht, dass nun die zweite Hauptachse in Richtung der noch verbleibenden größten Vari-

anz zeigt. Somit stehen die zweite Hauptachse und die zweite Hauptkomponente fest.
Dieser Vorgang wird solange wiederholt, bis eine neue Basis geschaffen ist.

2.3 PCA Algorithmus

Auf die Mathematik bezogen bedeutet das folgendes:
Die Daten $x_i \in \mathbb{R}^N, i = 1, \ldots, m$, entsprechen den m Punkten der Punktwolke. Diese Daten sind zentriert, d. h. dass $\sum_{i=1}^{m} x_i = 0$ gilt. PCA findet die Hauptachsen durch das Diagonalisieren der Kovarianzmatrix,

$$C = \frac{1}{m} \sum_{j=1}^{m} x_j x_j^T \tag{1}$$

Diese kann mit nichtnegativen Eigenwerten λ diagonalisiert werden, da sie positiv definit ist. Die Eigenwerte werden ermittelt, indem für die Eigenwerte $\lambda \geq 0$ und die Eigenvektoren $v \in \mathbb{R}^N \setminus \{0\}$ die Gleichung

$$\lambda v = Cv \tag{2}$$

berechnet wird. Durch Einsetzen der Gleichung (1) in (2), ergibt sich:

$$\lambda v = Cv = \frac{1}{m} \sum_{j=1}^{m} \langle x_j, v \rangle x_j$$

Somit liegen alle Lösungen v mit $\lambda \neq 0$ in dem Bereich von $x_1, \ldots, x_m$. Folglich ist die Gleichung (2) äquivalent zu:

$$\lambda \langle x_i, v \rangle = \langle x_i, Cv \rangle$$

für alle $i = 1, \ldots, m$.

2.3.1 Beispiel Schuhkauf

Als ein Besipiel für PCA dient das Problem der Schuhgrößenbestimmung beim Schuhkauf. Um die Schuhgröße zu ermitteln, gibt es viele Möglichkeiten und somit viele Messdaten. Werden jedoch nur die beiden Daten Fußbreite und Fußlänge betrachtet, so fällt auf, dass sie in einem Zusammenhang stehen.

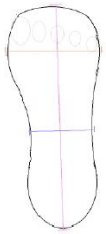

Abbildung 1: Messdaten bei einem Fuß [4]

Eingetragen in ein Koordiantensystem ergeben die Daten für Fußbreite und Fußlänge folgende Abbildung:

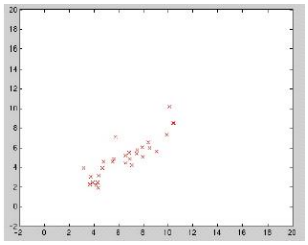

Abbildung 2: Messdaten im Koordinatensystem [4]

PCA wird angewandt, um die Schuhgröße durch nur eine Komponente zu beschreiben:

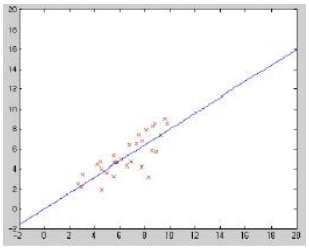

Abbildung 3: PCA angewandt auf die Messdaten im Koordinatensystem [4]

Es fällt auf, dass die meisten Informationen durch Übertragung der Abbildung des Fußes auf die Linie übermittelt werden. Die restlichen Informationen befinden sich entlang der orthogonalen Mittellinie.

2.4 Multivariate Verteilung

Bisher wurde keine Annamhe über die Verteilung der Daten gemacht. Diese können z. B. multivariat verteilt sein. Bevor darauf näher eingegangen wird, sollte erst einmal erklärt werden, was eine multivariate Verteilung ist:
Eine multivariate Verteilung ist eine gemeinsame Wahrscheinlichkeitsverteilung mehrerer Zufallsvariablen $X_j, j = 1,\ldots,p$, jeweils mit einem Erwartungswert $E(X_j)$ und einer Varianz $V(X_j)$. Außerdem sind die Zufallsvariablen paarweise korreliert, sie haben die Kovarianz $Cov(X_j,X_k), (j,k = 1,\ldots,p, j \neq k)$.
Interessant ist die Wahrscheinlichkeit, dass alle X_j gleich einer jeweiligen Konstante x_j sind. Das heißt:

$$P(X_1 \leq x_1; X_2 \leq x_2; \ldots; X_p \leq x_p) = F_x(x_1; x_2; \ldots; x_p)$$

Die Varianzen und Kovarianzen werden in der Kovarianzmatrix aufgeführt:

$$C = \begin{pmatrix} V(X_1) & Cov(X_1,X_2) & \ldots & Cov(X_1,X_p) \\ Cov(X_2,X_1) & V(X_2) & \ldots & Cov(X_2,X_p) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(X_p,X_1) & Cov(X_p,X_2) & \ldots & V(X_p) \end{pmatrix}$$

Durch Umformen zum Korrelationskoeffizienten

$$Kor(X_j,X_k) = \frac{Cov(X_j,X_k)}{\sqrt{V(X_j) \cdot V(X_k)}}$$

ergibt sich die Korrelationsmatrix

$$\underline{R} = \begin{pmatrix} 1 & Kor(X_1,X_2) & \dots & Kor(X_1,X_p) \\ Kor(X_2,X_1) & 1 & \dots & Kor(X_2,X_p) \\ \vdots & \vdots & \ddots & \vdots \\ Kor(X_p,X_1) & Kor(X_p,X_2) & \dots & 1 \end{pmatrix}$$

Die multivariate Verteilung ist von Bedeutung für die multivariate Normalverteilung:
Geben sei ein Vektor $\underline{x}$ aus p normalverteilten Zufallsvariablen mit dem Erwartungswertvektor $E(\underline{x})$ und der Korvarianzmatrix C:

$$\underline{x} \sim N_p(E(\underline{x}); C)$$

2.4.1 Beispiel Apfelbaumplantage

Als Beispiel für eine multivariate Normalverteilung dient eine Apfelbaumplantage:
Die Apfelbäume dieser Plantage sind alle ungefähr gleich alt. Interessant ist die Höhe der Bäume, der Ertrag und die Zahl der Blätter. Hierfür werden also folgende Zufallsvariabeln definiert:
X_1 : Höhe eines Baumes[m]; X_2 : Etrag[100kg]; X_3 : Zahl der Blätter[1000 Stück]
Die Variablen sind wie folgt normalverteilt:

$$X_1 \sim N(4;1); X_2 \sim N(20;100); X_3 \sim N(20;225);$$

Es wird beobachtet, dass Ertrag und Höhe eines Baumes korreliert sind. Somit ergibt sich für die Variablen Höhe und Ertrag die Kovarianz $Cov(X_1,X_2) = 9$ und den Korrelationskoeffizienten $Kor(X_1,X_2) = 0,9$. Für die restlichen Daten ergibt sich:
$Cov(X_1,X_3) = 12,75$ mit $Kor(X_1,X_3) = 0,85$ und $Cov(X_2,X_3) = 120$ mit $Kor(X_2,X_3) = 0,8$.
Fasst man diese Ergebnisse zusammen, ergibt sich Kovarianzmatrix

$$C = \begin{pmatrix} 1 & 9 & 12,75 \\ 9 & 100 & 120 \\ 12,75 & 120 & 225 \end{pmatrix}$$

sowie die Korrelationsmatrix

$$\underline{R} = \begin{pmatrix} 1 & 0,9 & 0,85 \\ 0,9 & 1 & 0,8 \\ 0,85 & 0,8 & 1 \end{pmatrix}.$$

2.5 Bivariat verteilte Eingabedaten

In diesem Abschnitt wird die multivariate Verteilung im Zusammenhang mit der Hauptkomponentenanalyse betrachtet. Um den Unterschied zwischen verschiedenen Verteilungen der Daten zu verdeutlichen, wird angenommen, dass die Daten normal bzw. bivariat verteilt sind.
An dem Algorithmus der Hauptkomponentenanalyse ändert sich nichts, allerdings haben die Hauptkomponenten im Falle einer gemeinsamen Normalverteilung eine anschauliche Interpretation: Sie stimmen mit den Hauptachsen eines Ellipsoids überein.

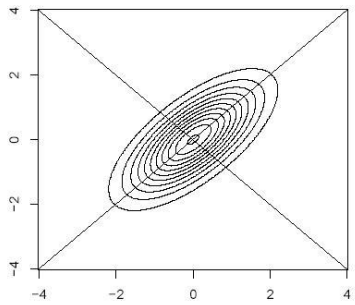

Abbildung 4: Normalverteilung der Hauptkomponenten [5]

Im Falle einer bivariaten Normalverteilung werden die Hauptachsen so gedreht, dass sie parallel zu den Koordinatenachsen verlaufen.

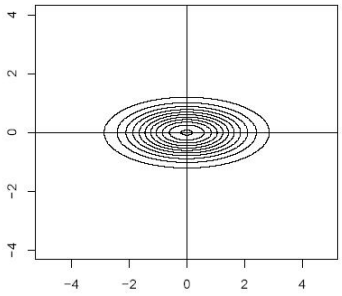

Abbildung 5: Bivariate Normalverteilung [5]

3 Kernel-PCA

3.1 Kurze Wiederholung von Kernels

In diesem Kapitel soll die nichtlineare Hauptkomponentenanalyse (Kernel-PCA) beschrieben werden. Diese baut auf den sogenannten Kernels auf, die zuerst kurz erklärt werden sollen:
Kernels transfomieren einen Datensatz, der nicht direkt linear trennbar ist, von dem Eingaberaum (englisch: Input Space (IS)) in einen neuen Raum mit höherer Dimension, dem Merkmalsraum (englisch: Feature Space (FS)). Dies erfolgt durch die (nicht-)lineare Abbildung $\Phi : IS \rightarrow FS$. Es gibt verschiedene Arten von Kerneln, z. B. :

- Lineare Kernel: $k(x_i, x_j) = \langle x_i, x_j \rangle$
- Polynomielle Kernel: $k(x_i, x_j) = \langle x_i, x_j \rangle^d$
- RBF-Kernel (Radial-Basisfunktionen-Kernel): $k(x_i, x_j) = \exp(-\frac{\|x_i - x_j\|^2}{2\sigma^2})$

Zur Verdeutlichung dient folgendes Beispiel:
Gegeben sei die Funktion:

$$\Phi : \mathbb{R}^2 \rightarrow \mathbb{R}^3$$

$$(x_1, x_2) \rightarrow (x_1, x_2, x_1^2 + x_2^2)$$

Die Daten aus dem zweidimensionalen Eingaberaum ergeben:

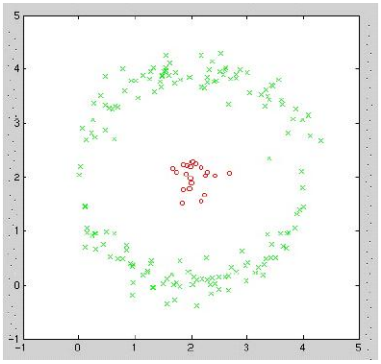

Abbildung 6: Eingaberaum für ein zwei Klassen Problem in $\mathbb{R}^2$ [4]

Es ist deutlich sichtbar, dass die Daten im Eingaberaum nicht linear trennbar sind. Durch die Abbildung in einen dreidimensionalen Raum kann dieses Problem jedoch gelöst werden:

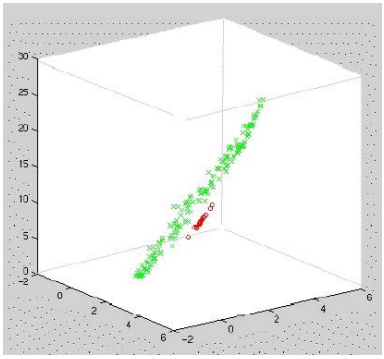

Abbildung 7: Merkmalsraum [4]

3.2 Allgemeines zu Kernel-PCA

Bei der Hauptkomponentenanalyse waren lineare Probleme im Eingaberaum interessant, nun sollen jedoch Hauptkomponenten von Variablen bestimmt werden, die nichtlinear mit dem Eingaberaum zusammenhängen. Dazu zählen auch Variablen, die durch willkürliche Wechselbeziehungen höherer Ordnung zwischen Eingabevariablen erreicht werden.
Der Algorithmus der nichtlinearen Hauptkomponentenanalyse ist folgender:
Als erstes muss die Gram-Matrix

$$K_{ij} = k(x_i, x_j)$$

berechnet und diagonalisiert werden. Als nächstes werden die Eigenvektor-Koeffizienten α^n normiert, es gilt also

$$\lambda_n \langle \alpha^n, \alpha^n \rangle = 1$$

Zuletzt werden die Hauptbestandteile eines Testpunktes x durch Berechnung von Abbildungen auf die Eigenvektoren extrahiert:

$$\langle v^n, \Phi(x) \rangle = \sum_{i=1}^{m} \alpha_i^n k(x_i, x), n = 1, \ldots, p$$

Angenommen X sei der Eingaberaum und H sei ein Merkmalsraum mit der Funktion

$$\Phi : X \to H, x \to \Phi(x)$$

Es gilt $\sum_{i=1}^{m} \Phi(x_i) = 0$, die Daten sind also zentriert. Die Kovarianzmatrix in H ist

$$C = \frac{1}{m} \sum_{j=1}^{m} \Phi(x_j)\Phi(x_j)^T$$

Genau wie bei der linearen Hauptkomponentenanalyse werden die Eigenvektoren $v \in H \setminus \{0\}$ mittels der Gleichung

$$\lambda v = Cv$$

ermittelt, wobei $\lambda \leq 0$ die Eigenwerte sind. Die Lösungen für v mit $\lambda \neq 0$ liegen im Bereich $\Phi(x_1), \ldots, \Phi(x_m)$. Dies hat zwei Konsequenzen:
Erstens gilt

$$\lambda \langle \Phi(x_n), v \rangle = \langle \Phi(x_n), Cv \rangle \tag{3}$$

für alle $n = 1, \ldots, m$ und zweitens existieren die Koeffizienten $\alpha_i, i = 1, \ldots, m$, sodass

$$v = \sum_{i=1}^{m} \alpha_i \Phi(x_i). \tag{4}$$

3.2.1 Die Gram-Matrix

Durch Kombinieren der Gleichungen (3) und (4) gilt

$$\begin{aligned}
\lambda \langle \Phi(x_n), v \rangle &=^4 \lambda \left\langle \Phi(x_n), \sum_{i=1}^{m} \alpha_i \Phi(x_i) \right\rangle = \lambda \sum_{i=1}^{m} \alpha_i \langle \Phi(x_n), \Phi(x_i) \rangle \\
&= \langle \Phi(x_n), Cv \rangle = \left\langle \Phi(x_n), \frac{1}{m} \sum_{j=1}^{m} \Phi(x_j)\Phi(x_j)^T \sum_{i=1}^{m} \alpha_i \Phi(x_i) \right\rangle \\
&= \frac{1}{m} \sum_{i=1}^{m} \alpha_i \left\langle \Phi(x_n), \sum_{j=1}^{m} \Phi(x_j) \left\langle \Phi(x_j), \Phi(x_i) \right\rangle \right\rangle
\end{aligned}$$

für alle $n = 1, \ldots, m$. Ausgedrückt in der $m \times m$-Gram-Matrix $K_{ij} = \left\langle \Phi(x_i), \Phi(x_j) \right\rangle$ ergibt das

$$m\lambda K\alpha = K^2\alpha \Leftrightarrow m\lambda\alpha = K\alpha, \tag{5}$$

wobei α der Spaltenvektor mit $\alpha_1, \ldots, \alpha_m$ ist.

3.2.2 Normieren der Eigenvektor-Koeffizienten

Seien $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m$ die Eigenwerte und $\alpha^1, \ldots, \alpha^m$ die Eigenvektoren von K. Angenommen die entsprechenden Vektoren in H sind normiert, also $\langle v^n, v^n \rangle = 1$ für alle $n = 1, \ldots, p$. Dies führt zu folgender Normierung von $\alpha^1, \ldots, \alpha^p$:

$$1 =^4 \sum_{i,j=1}^{m} \alpha_i^n \alpha_j^n \left\langle \Phi(x_i), \Phi(x_j) \right\rangle = \sum_{i,j=1}^{m} \alpha_i^n \alpha_j^n K_{ij}$$

$$= \langle \alpha^n, K\alpha^n \rangle =^5 \lambda_n \langle \alpha^n, \alpha^n \rangle$$

3.2.3 Berechnung der Abbildungen auf die Eigenvektoren

Die Eigenvektoren in H werden dazu genutzt, die Hauptkomponenten eines Testpunktes zu extrahieren. Angenommen es existiert ein Testpunkt x mit dem Bild $\Phi(x) \in H$, dann sind

$$\langle v^n, \Phi(x) \rangle = \sum_{i=1}^{m} \alpha_i^n \langle \Phi(x_i), \Phi(x) \rangle$$

die nichtlinearen Hauptkomponenten. Bis jetzt wurde immer die Annahme gemacht, dass die Daten zentriert sind. Das Problem dabei ist, dass dies zwar einfach im Eingaberaum, in H jedoch schwieriger ist. In H kann das Mittel der abgebildeten Beobachtungen nicht eindeutig berechnet werden. Deshalb muss die Matrix

$$\tilde{K}_{ij} = (K - 1_m K - K 1_m + 1_m K 1_m)_{ij}$$

diagonalisiert werden, wobei $(1_m)_{ij} := \frac{1}{m}$ für alle i, j.

3.3 Kernel-PCA Experiment

In diesem Abschnitt wird ein Spielzeugexperiment (Toy example) vorgestellt.

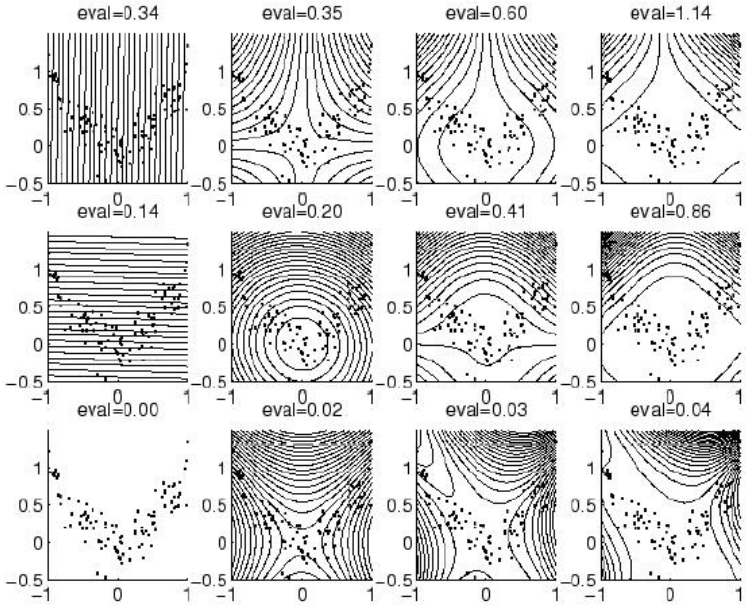

Abbildung 8: Zweidimensionales Spielzeugexperiment [3]

Zunächst wurde eine Parabel verrauscht. Dann wurde PCA und Kernel-PCA angewandt, wobei polynomielle Kernel benutzt wurden.Von links nach rechts steigt der Grad der Kernels von 1 bis 4. Die drei Zeilen zeigen jeweils die ersten 3 Eigenvektoren.
Lineare PCA (linke Spalte) führt nur zu zwei Eigenwerten ungleich null, da der Eingaberaum zweidimensional ist. Kernel-PCA hingegen erlaubt es, mehr Komponenten zu extrahieren. Außerdem ist in der Abbildung zu sehen, dass Kernel-PCA die Struktur der Daten besser beschreibt als lineare PCA. In allen Fällen schwankt die erste Hauptkomponente monoton der Parabel entlang, die den Daten zu Grunde liegt.

4 Unterschied zwischen PCA und Kernel-PCA

Wie in Abbildung 8 zu sehen, muss es einen Unterschied zwischen PCA und Kernel-PCA geben. Zunächst kann gesagt werden, dass Kernel-PCA PCA in einem hoch-dimensionalen Merkmalsraum entspricht. Die mathematischen und statistischen Eigenschaften von PCA werden auf Kernel-PCA übertragen. Allerdings gibt es dazu Modifikationen:
Bei Kernel-PCA werden Aussagen über Punkte $\Phi(X_i), i = 1, \ldots, m$ im Merkmalsraum gemacht und nicht über die Punkte x_i im Eingaberaum wie es bei PCA der Fall ist. Somit ist Kernel-PCA auch nicht abhängig von der Dimension des Eingaberaumes. Angenommen die Anzahl der Eingabedaten n übersteigt die Eingabedimension (m):
PCA kann höchstens m Eigenwerte $\lambda \neq 0$ finden, Kernel-PCA kann jedoch bis zu n Eigenwerte $\lambda \neq 0$ finden.
Arbeitstechnisch gesehen ist Kernel-PCA aber aufwändiger:
Für jede extrahierte Hauptkomponente muss die Kernel Funktion ausgewertet werden, bei PCA muss nur ein Produkt ausgewertet werden. Wie unterschiedlich die Ergebnisse bei einem verrauschten Bild sein können, zeigen folgende Abbildungen:

Abbildung 9: Eingabe [3]

Abbildung 10: Verrauschte Eingabe [3]

Abbildung 11: Lineare PCA [3]

Abbildung 12: Kernel-PCA [3]

Bei diesem Beispiel wurden handschriftliche Ziffern verrauscht. Daraufhin wurden lineare PCA und Kernel-PCA angewendet. Bei PCA ist die Struktur der

Daten erkennbar, jedoch ist z. B. die Ziffer 5 nur schwer identifizierbar. Die Ergebnisse von Kernel-PCA sind deutlich besser als die von PCA. Die Ziffern können alle eindeutig identifiziert werden.

5 Zusammenfassung

Eine Hauptkomponentenanalyse ist eine orthogonale Transformation im p-dimensionalen Raum der Originalvariablen in eine neue Variablenmenge, die Hauptkomponenten genannt werden. PCA hat den Nachteil, dass sie bei einem nichtlinearen Problem nutzlos ist. Dennoch ist PCA in der Lage, eine Einsicht in die Struktur der Daten zu bringen. Kernel-PCA führt PCA mittels Kernels in einem höher-dimensionalen Merkmalsraum aus. Obwohl Kernel-PCA wesentlich bessere Ergebnisse als PCA liefert, hat es auch Nachteile: Wenn die Mustergröße zu groß ist, kann die Kernel-Matrix nicht diagonalisiert werden.

5.1 Anwendungsgebiete

Lineare PCA findet in zahlreichen technischen und wissenschaftlichen Gebieten Anwendung, z. B. bei der Lärmverminderung, Dichteschätzung und Bildbearbeitung.
Kernel-PCA kann in allen Gebieten angewendet werden, in denen traditionelle PCA für die Merkmalsextraktion verwendet wird.

6 Literatur

[1] J. Bortz, Statistik für Sozialwissenschaftler, Springer-Verlag, 1999

[2] R. Duda, P. Hart, D. Stork, Pattern Classification, Wiley Interscience, 2000

[3] B. Schölkopf, A. Smola, Learning with Kernels, MIT Press, 2002

[4]`http://www.cs.ucsd.edu/classes/fa01/cse291/kernelPCA\_article.pdf`

[5]`http://www.statoek.wiso.uni-goettingen.de/veranstaltungen/Multivariate/Daten/mvsec4.pdf`