KAPITEL 6

Dimensionsreduktion: Mit dem Hauptkomponentenverfahren die Datenwolke flach drücken

Mithilfe automatischer Verfahren zur Datensammlung und Merkmalserzeugung kann man im Handumdrehen eine große Zahl von Merkmalen bekommen. Aber nicht alle davon sind sinnvoll. In den Kapiteln 3 und 4 besprachen wir das Filtern nach Häufigkeit und die Merkmalsskalierung als Möglichkeiten, um aussagelose Merkmale zu beseitigen. Nun werfen wir einen genauen Blick auf die Reduktion der Merkmalsdimensionen mithilfe des *Hauptkomponentenverfahrens* (PCA, engl. *Principal Component Analysis*).

Dieses Kapitel stellt einen Einstieg in die modellbezogenen Verfahren zur Merkmalskonstruktion dar. Die meisten bisher gezeigten Verfahren funktionieren unabhängig von den Daten. Beispielsweise könnte die Vorschrift beim Filtern nach Häufigkeit lauten: »Verwirf alle Zähler, die kleiner als n sind.« Dafür sind keine weiteren Informationen aus den Daten selbst nötig.

Modellbezogene Verfahren benötigen dagegen Informationen aus den Daten. Das Hauptkomponentenverfahren ist beispielsweise über die Hauptachsen der Daten definiert. In den früheren Kapiteln gab es immer eine klare Trennung zwischen Daten, Merkmalen und Modellen. Ab jetzt wird diese Unterscheidung immer mehr verschwimmen. Und genau darin liegt das Spannende in der aktuellen Forschung zum Erlernen von Merkmalen.

Die Grundidee

Dimensionsreduktion bedeutet, »aussagelose Informationen« loszuwerden, dabei aber die entscheidenden Teile zu behalten. Es gibt viele mögliche Definitionen von »aussagelos«. Das Hauptkomponentenverfahren bezieht sich auf das Konzept der linearen Abhängigkeit. Im Abschnitt »Die Anatomie einer Matrix« auf Seite 180 beschreiben wir den Spaltenraum einer Datenmatrix als die lineare Hülle aller Merkmalsvektoren. Ist der Spaltenraum klein gegenüber der Gesamtzahl von Merkmalen, sind die meisten dieser Merkmale Linearkombinationen einiger weniger Schlüsselmerkmale. Linear abhängige Merkmale verschwenden Speicherplatz und Rechenleistung, da ihr Informationsgehalt in viel weniger Merkmalen hätte kodiert

werden können. Um das zu vermeiden, versucht man mittels des Hauptkomponentenverfahrens, diese »aufgeblähten« Daten in einem linearen Unterraum mit viel weniger Dimensionen flach zu drücken.

Stellen Sie sich die Menge der Datenpunkte im Merkmalsraum vor. Jeder Datenpunkt ist ein Punkt, und die gesamte Menge von Datenpunkten bildet eine Wolke. In Abbildung 6-1 (a) sind die Datenpunkte gleichmäßig über beide Merkmalsdimensionen verteilt, und die Wolke füllt den Raum aus. In diesem Beispiel hat der Spaltenraum den vollen Rang. Sind jedoch einige dieser Merkmale Linearkombinationen von anderen, sieht die Wolke nicht mehr so prall aus, sondern mehr wie in Abbildung 6-1 (b): eine flache Wolke, wobei Merkmal 1 ein Duplikat (oder ein skalares Vielfaches) von Merkmal 2 ist. In diesem Fall sagen wir, dass die *intrinsische Dimension* der Wolke 1 ist, obwohl sie in einem zweidimensionalen Merkmalsraum liegt.

In der Praxis hat man es selten mit genau gleichen Größen zu tun. Viel eher werden wir Merkmalen begegnen, die einander beinahe gleichen, aber eben nicht ganz. In so einem Fall könnte die Datenwolke aussehen wie in Abbildung 6-1 (c), eine recht magere Wolke. Wenn wir die Anzahl der Merkmale, die in das Modell eingehen, verkleinern wollen, könnten wir Merkmal 1 und Merkmal 2 durch ein neues Merkmal ersetzen, das wir vielleicht Merkmal 1,5 nennen und das auf der Diagonalen zwischen den beiden ursprünglichen Merkmalen liegt. Der Ausgangsdatensatz kann dann gut durch eine Zahl – den Ort entlang der Richtung von Merkmal 1,5 – dargestellt werden anstatt durch zwei Zahlen *f1* und *f2*.

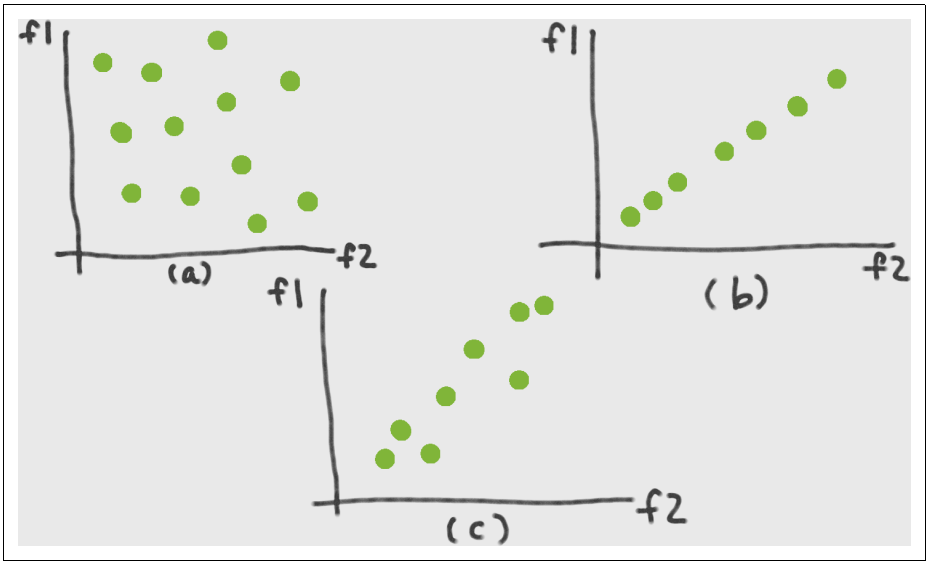

Abbildung 6-1: Datenwolken im Merkmalsraum: Datenwolke mit vollem Rang (a), niedrigdimensionale Datenwolke (b) und annähernd niedrigdimensionale Datenwolke (c)

Der Schlüsselgedanke dabei ist, *redundante Merkmale durch wenige neue Merkmale zu ersetzen, die die im ursprünglichen Merkmalsraum enthaltene Information geeignet zusammenfassen*. Wenn es nur zwei Merkmale gibt, ist es leicht, das neue Merkmal zu formulieren. Viel schwerer ist es, wenn der ursprüngliche Merkmalsraum Hunderte oder Tausende von Dimensionen besitzt. Wir brauchen eine mathematische Beschreibung der gesuchten neuen Merkmale, um sie dann mithilfe von Optimierungsverfahren zu finden.

Eine mögliche mathematische Definition des »geeigneten Zusammenfassens von Informationen« besteht darin, zu fordern, dass die neue Datenwolke möglichst viel ihres ursprünglichen Volumens behalten soll. Wir drücken die Datenwolke zu einem Eierkuchen flach, aber der Eierkuchen soll in den richtigen Richtungen so dick wie möglich sein. Wir brauchen also ein Maß für das Volumen.

Volumen hat mit Abständen zu tun. Der Begriff des Abstands in einer Wolke von Datenpunkten ist aber gar nicht so offensichtlich. Man könnte die größte Entfernung zwischen jeweils zwei beliebigen Punkten messen, aber das erweist sich als eine mathematisch sehr schwer optimierbare Funktion. Eine Alternative besteht in der Messung des mittleren Abstands zwischen Punktepaaren oder, äquivalent dazu, der mittleren Entfernung jedes Punkts vom Mittelwert, also der Varianz. Diese lässt sich viel leichter optimieren. (Das Leben ist schwer. Statistiker haben gelernt, bequeme Abkürzungen zu nehmen.) Mathematisch ausgedrückt, handelt es sich um die Maximierung der Varianz der Datenpunkte im neuen Merkmalsraum.

Wie Sie sich in Formeln der linearen Algebra zurechtfinden

Um in der Welt der linearen Algebra die Orientierung zu behalten, merken Sie sich, welche Größen Skalare bzw. Vektoren sind und ob die Vektoren vertikal oder horizontal ausgerichtet sind. Behalten Sie die Dimensionen Ihrer Matrizen im Blick, da sie oft verraten, ob es um die Zeilen- oder die Spaltenvektoren geht. Zeichnen Sie die Matrizen und Vektoren als Rechtecke auf ein Blatt und stellen Sie sicher, dass die Abmessungen zusammenpassen. Ebenso, wie man in der Algebra schon weit kommt, wenn man auf die Maßeinheiten achtet (Entfernungen in Kilometern, Geschwindigkeiten in Kilometern pro Stunde), braucht man in der linearen Algebra nur die Dimensionen.

Herleitung

$\boldsymbol{X}$ sei wieder die $n \times d$-Datenmatrix, wobei n die Anzahl der Datenpunkte und d die Anzahl der Merkmale ist. $\mathbf{x}$ sei ein Spaltenvektor, der einen einzelnen Datenpunkt enthält. ($\mathbf{x}$ ist also die Transponierte einer der Zeilen von $\boldsymbol{X}$.) Weiterhin sei $\mathbf{v}$ einer der neuen Merkmalsvektoren oder auch Hauptkomponenten, die wir bestimmen wollen.

Singulärwertzerlegung einer Matrix

Jede rechteckige Matrix kann in drei Matrizen mit bestimmten Abmessungen und Eigenschaften zerlegt werden:

$$\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$$

Dabei sind $\boldsymbol{U}$ und $\boldsymbol{V}$ orthogonale Matrizen (das heißt, $\boldsymbol{U}^T\boldsymbol{U} = I$ und $\boldsymbol{V}^T\boldsymbol{V} = I$). $\boldsymbol{\Sigma}$ ist eine Diagonalmatrix, die die Singulärwerte von $\boldsymbol{X}$ enthält, die wiederum positiv, null oder negativ sein können. Angenommen, $\boldsymbol{X}$ hat n Zeilen und d Spalten und es gilt $n \geq d$. Dann hat $\boldsymbol{U}$ die Abmessungen $n \times d$, und $\boldsymbol{\Sigma}$ und $\boldsymbol{V}$ haben die Abmessungen $d \times d$. (der Abschnitt »Singulärwertzerlegung« auf Seite 183 behandelt die Singulär- und Eigenwertzerlegung einer Matrix ausführlich.)

Lineare Projektion

Schauen wir uns die Herleitung des Hauptkomponentenverfahrens Schritt für Schritt an. Abbildung 6-2 veranschaulicht den gesamten Vorgang.

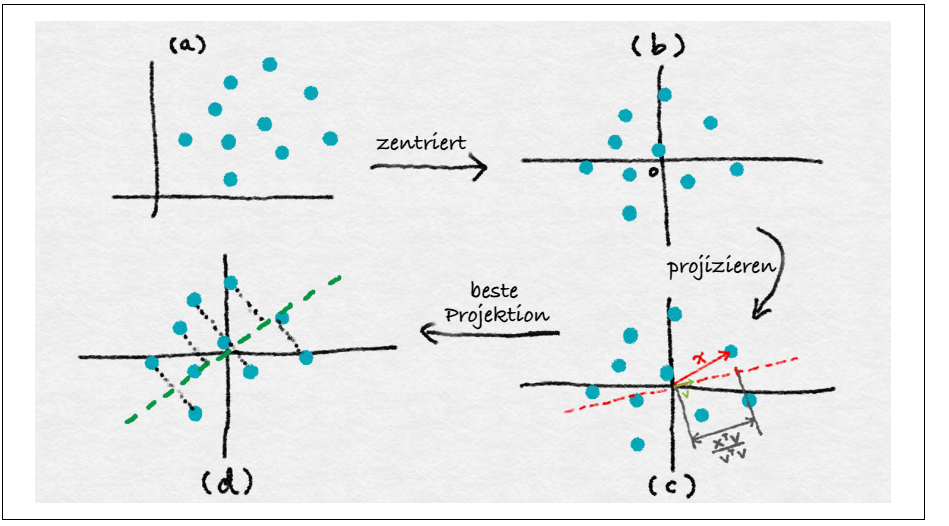

Abbildung 6-2: Veranschaulichung des Hauptkomponentenverfahrens: Ausgangsdaten im Merkmalsraum (a), zentrierte Daten (b), Projektion eines Datenvektors x auf einen anderen Vektor v (c), Richtung der maximalen Varianz der projizierten Koordinaten (d), gleich dem Haupteigenvektor von X^TX

Beim Hauptkomponentenverfahren werden die Daten durch eine lineare Projektion in den neuen Merkmalsraum überführt. Abbildung 6-2 (c) zeigt eine lineare Projektion. Wenn wir **x** auf **v** projizieren, ist die Länge der Projektion proportional zum inneren Produkt der zwei Vektoren, normiert auf die Norm von **v** (sein inneres Produkt mit sich selbst). Später werden wir für **v** die Einheitsnorm verlangen. Der einzige relevante Teil ist also der Zähler – nennen wir ihn **v** (siehe Formel 6-1).

Formel 6-1: Projektionskoordinate

$$z = \mathbf{x}^T\mathbf{v}$$

Die Größe z ist ein Skalar, während $\mathbf{x}$ und $\mathbf{v}$ Spaltenvektoren sind. Da es eine Menge Datenpunkte gibt, können wir den Vektor $\mathbf{z}$ von deren Projektionskoordinaten bezüglich des neuen Merkmals $\mathbf{v}$ bilden (siehe Formel 6-2). Dabei ist $\boldsymbol{X}$ die bekannte Datenmatrix, deren Zeilenvektoren jeweils einen Datenpunkt enthalten. Der so entstandene Vektor $\mathbf{z}$ ist ein Spaltenvektor.

Formel 6-2: Vektor der Projektionskoordinaten

$$\mathbf{z} = \boldsymbol{X}\mathbf{v}$$

Varianz und empirische Varianz

Als Nächstes berechnen wir die Varianz der Projektionen. Die Varianz ist definiert als der Erwartungswert des Quadrats des Abstands vom Mittelwert (siehe Formel 6-3).

Formel 6-3: Varianz einer Zufallsvariablen Z

$$\mathrm{Var}(Z) = \mathrm{E}[Z - \mathrm{E}(Z)]^2$$

Es gibt dabei ein kleines Problem: Unsere Aufgabenstellung sagt nichts über den Mittelwert *E(Z)* aus; er ist eine freie Variable. Eine Lösung dafür wäre, ihn von jedem Datenpunkt zu subtrahieren und so aus der Gleichung zu entfernen. Der dadurch entstehende Datensatz hat den Mittelwert null, sodass die Varianz einfach der Erwartungswert von Z^2 ist. Geometrisch ausgedrückt, führt die Subtraktion des Mittelwerts zur Zentrierung der Daten (siehe Abbildung 6-2 (a-b)).

Eine damit nah verwandte Größe ist die Kovarianz zwischen zwei Zufallsvariablen Z^1 und Z^2 (siehe Formel 6-4). Stellen Sie sie sich als die Erweiterung der Varianz (einer einzelnen Zufallsvariablen) auf zwei Zufallsvariablen vor.

Formel 6-4: Kovarianz zwischen zwei Zufallsvariablen Z^1 und Z^2

$$\mathrm{Cov}(Z^1, Z^2) = \mathrm{E}[(Z^1 - \mathrm{E}(Z^1)(Z^2 - \mathrm{E}(Z^2)]$$

Wenn die Zufallsvariablen den Mittelwert null haben, fällt ihre Kovarianz mit ihrer *linearen Korrelation* $E[Z_1Z_2]$ zusammen. Wir werden diesen Begriff an späterer Stelle erläutern.

Statistische Größen wie Varianz und Erwartungswert sind auf einer Datenverteilung definiert. In der Praxis kennen wir die wahre Verteilung nicht, sondern nur eine Menge beobachteter Datenpunkte z_1, ..., z_n. Wir sprechen von einer *empirischen Verteilung*, die uns eine empirische Abschätzung der Varianz ermöglicht (siehe Formel 6-5).

Formel 6-5: Empirische Varianz von Z anhand der Beobachtungsdaten z

$$\mathrm{Var}_{\mathrm{emp}}(Z) = \frac{1}{n-1}\sum_{i=1}^{n} z_i^2$$

Hauptkomponenten: Erste Schreibweise

Zusammen mit der Definition von z_i in Formel 6-1 können wir die Maximierung der Varianz der projizierten Daten durch Formel 6-6 ausdrücken. (Wir lassen den Nenner n–1 aus der Definition der empirischen Varianz weg, da er eine globale Konstante ist und keinen Einfluss darauf hat, wo das Maximum auftritt.)

Formel 6-6: Zielfunktion der Hauptkomponenten

$$\max_{\mathbf{w}} \sum_{i=1}^{n} (\mathbf{x}_i^T \mathbf{w})^2, \text{ wobei } \mathbf{w}^T \mathbf{w} = 1$$

Die Bedingung legt dabei das innere Produkt von **w** mit sich selbst auf 1 fest, was gleichbedeutend damit ist, dass der Vektor die Einheitslänge haben muss. Wir verlangen das, weil wir uns nur für die Richtung, nicht aber den Betrag von **w** interessieren. Der Betrag von **w** ist ein unnötiger Freiheitsgrad, also werden wir ihn los, indem wir ihn auf einen willkürlichen Wert festlegen.

Hauptkomponenten: Matrix-Vektor-Schreibweise

Nun kommt der interessante Schritt. Der Term mit der Summe der Quadrate in Formel 6-6 ist recht umständlich; in einem Matrix-Vektor-Format wäre er viel klarer. Können wir das erreichen? Ja, der Schlüssel liegt in der Quadratsummenidentität: Die Summe einer Menge von Quadraten von Termen ist gleich dem Quadrat der Norm eines Vektors, dessen Elemente gerade diese Terme sind – was wiederum äquivalent zum inneren Produkt des Vektors mit sich selbst ist. Mithilfe dieser Identität können wir Formel 6-6 in Matrix-Vektor-Schreibweise umformen (siehe Formel 6-7).

Formel 6-7: Zielfunktion für Hauptkomponenten, Matrix-Vektor-Schreibweise

$$\max_{\mathbf{w}} \mathbf{w}^T \mathbf{w}, \text{ wobei } \mathbf{w}^T \mathbf{w} = 1$$

Diese Schreibweise der Hauptkomponentenzerlegung macht das Ziel deutlicher erkennbar: Wir suchen eine Ausrichtung der Eingabe, die die Norm der Ausgabe maximiert. Kommt Ihnen das bekannt vor? Die Lösung liegt in der *Singulärwertzerlegung* (SVD, engl. Singular Value Decomposition) von $\boldsymbol{X}$. Der optimale Vektor **w** ergibt sich als der linksseitige Hauptsingulärvektor von $\boldsymbol{X}$, der zugleich der Haupteigenvektor von $\boldsymbol{X}^T\boldsymbol{X}$ ist. Die projizierten Daten werden als Hauptkomponente der Ausgangsdaten bezeichnet.

Allgemeine Lösung für die Hauptkomponenten

Dieser Vorgang kann wiederholt werden. Sobald wir die erste Hauptkomponente bestimmt haben, können wir Formel 6-7 erneut anwenden unter der zusätzlichen

Bedingung, dass der neue Vektor orthogonal zu den zuvor gefundenen Vektoren sein soll (siehe Formel 6-8).

Formel 6-8: Zielfunktion für die (k+1)te Hauptkomponente

$\max_{\mathbf{w}} \mathbf{w}^T\mathbf{w}$, wobei $\mathbf{w}^T\mathbf{w} = 1$ und $\mathbf{w}^T\mathbf{w}_1 = \ldots = \mathbf{w}^T\mathrm{w}_k = 0$

Die Lösung ist der (k+1)te linksseitige Singulärvektor von $\boldsymbol{X}$, geordnet nach absteigenden Singulärwerten. Somit entsprechen die ersten k Hauptkomponenten den ersten k linksseitigen Singulärvektoren von $\boldsymbol{X}$.

Transformation der Merkmale

Sind die Hauptkomponenten bestimmt, können wir die Merkmale mittels linearer Projektion transformieren. $\boldsymbol{X} = \boldsymbol{U\Sigma V}^T$ sei die Singulärwertzerlegung von $\boldsymbol{X}$ und $\boldsymbol{V}_k$ die Matrix, deren Spalten die ersten k linksseitigen Singulärvektoren enthalten. $\boldsymbol{X}$ hat die Abmessungen $n \times d$, wobei d die Anzahl der ursprünglichen Merkmale ist und $\boldsymbol{V}_k$ die Abmessungen $d \times k$ besitzt. Statt auf einen einzelnen Projektionsvektor wie in Formel 6-2 können wir gleichzeitig auf mehrere Vektoren in einer Projektionsmatrix projizieren (siehe Formel 6-9).

Formel 6-9: Projektionsmatrix für das Hauptkomponentenverfahren

$\boldsymbol{W} = \boldsymbol{V}_k$

Die Matrix der projizierten Koordinaten ist leicht zu berechnen und kann noch weiter vereinfacht werden, da die Singulärvektoren paarweise orthogonal sind (siehe Formel 6-10).

Formel 6-10: Einfache Hauptkomponententransformation

$\boldsymbol{Z} = \boldsymbol{XW} = \boldsymbol{XV}_k = \boldsymbol{U\Sigma V}^T\boldsymbol{V}_k = \boldsymbol{U}_k\boldsymbol{\Sigma}_k$

Die projizierten Werte sind einfach die ersten k rechtsseitigen Singulärvektoren, skaliert um die ersten k Singulärwerte. Somit können wir die gesamte Lösung nach dem Hauptkomponentenverfahren, Komponenten wie auch Projektionen, einfach durch die Singulärwertzerlegung von $\boldsymbol{X}$ erhalten.

Implementierung des Hauptkomponentenverfahrens

Bei vielen Herleitungen des Hauptkomponentenverfahrens werden zuerst die Daten zentriert, und dann wird die Eigenwertzerlegung der Kovarianzmatrix durchgeführt. Der einfachste Weg, das Hauptkomponentenverfahren zu implementieren, führt jedoch über die Singulärwertzerlegung der zentrierten Datenmatrix.

Implementierungsschritte für das Hauptkomponentenverfahren

1. Zentriere die Datenmatrix:

   $$\boldsymbol{C} = \boldsymbol{X} - \mathbf{1}\boldsymbol{\mu}^T$$

   wobei $\mathbf{1}$ ein Spaltenvektor aus lauter Einsen und $\boldsymbol{\mu}$ ein Spaltenvektor aus den Mittelwerten der Zeilen von $\boldsymbol{X}$ ist.
2. Berechne die Singulärwertzerlegung:

   $$\boldsymbol{C} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$$

3. Bestimme die Hauptkomponenten. Die ersten k Hauptkomponenten sind die ersten k Spalten von $\boldsymbol{V}$, das heißt die rechtsseitigen Singulärvektoren, die zu den k größten Singulärwerten gehören.
4. Transformiere die Daten. Die transformierten Daten sind einfach die ersten k Spalten von $\boldsymbol{U}$. (Wenn das Ergebnis geweißt werden soll, skaliere die Vektoren um die inversen Singulärwerte. Dazu dürfen die ausgewählten Singulärwerte nicht null sein; siehe den Abschnitt »Weißen und Nullphasenverfahren« auf Seite 108.)

Das Hauptkomponentenverfahren am Werk

Wir wollen ein besseres Gespür dafür bekommen, wie das Hauptkomponentenverfahren funktioniert, indem wir es auf Bilddaten anwenden. Der MNIST-Datensatz (*http://yann.lecun.com/exdb/mnist/*) enthält Bilder handschriftlicher Ziffern von 0 bis 9. Die Originalbilder sind 28 × 28 Pixel groß. Eine Untermenge der Bilder in niedrigerer Auflösung wird mit scikit-learn (*http://bit.ly/2G3A3dA*) ausgeliefert, wozu jedes Bild auf 8 × 8 Pixel verkleinert wurde. Die Ausgangsdaten in scikit-learn haben 64 Dimensionen. In Beispiel 6-1 wenden wir das Hauptkomponentenverfahren an und visualisieren den Datensatz anhand der ersten drei Hauptkomponenten.

Beispiel 6-1: Hauptkomponentenverfahren anhand des Zifferndatensatzes von scikit-learn (einer Untermenge des MNIST-Datensatzes)

```
>>> from sklearn import datasets
>>> from sklearn.decomposition import PCA

# Lade die Daten.
>>> digits_data = datasets.load_digits()
>>> n = len(digits_data.images)

# Jedes Bild wird als 8-mal-8-Feld dargestellt.
# Linearisiere das Feld zur Eingabe ins Hauptkomponentenverfahren.
>>> image_data = digits_data.images.reshape((n, -1))
```

```
>>> image_data.shape
(1797, 64)

# Wahre Markierungen der Ziffern in den Bildern.
>>> labels = digits_data.target
>>> labels
array([0, 1, 2, ..., 8, 9, 8])

# Passe einen Hauptkomponententransformator an den Datensatz an.
# Die Anzahl der Komponenten wird automatisch so gewählt, dass mindestens
# 80 % der Gesamtvarianz erreicht werden.
>>> pca_transformer = PCA(n_components=0.8)
>>> pca_images = pca_transformer.fit_transform(image_data)
>>> pca_transformer.explained_variance_ratio_
array([ 0.14890594,  0.13618771,  0.11794594,  0.08409979,  0.05782415,
        0.0491691 ,  0.04315987,  0.03661373,  0.03353248,  0.03078806,
        0.02372341,  0.02272697,  0.01821863])
>>> pca_transformer.explained_variance_ratio_[:3].sum()
0.40303958587675121

# Visualisiere die Ergebnisse.
>>> import matplotlib.pyplot as plt
>>> from mpl_toolkits.mplot3d import Axes3D
>>> %matplotlib notebook
>>> fig = plt.figure()
>>> ax = fig.add_subplot(111, projection='3d')
>>> for i in range(100):
...     ax.scatter(pca_images[i,0], pca_images[i,1], pca_images[i,2],
...                marker=r'${}$'.format(labels[i]), s=64)

>>> ax.set_xlabel('Principal component 1')
>>> ax.set_ylabel('Principal component 2')
>>> ax.set_zlabel('Principal component 3')
```

Die ersten 100 projizierten Bilder sind in einem 3-D-Diagramm in Abbildung 6-3 zu sehen, wo sie durch ihre jeweiligen Markierungen dargestellt werden. Die ersten drei Hauptkomponenten sind für etwa 40 % der gesamten Varianz des Datensatzes verantwortlich. Das ist bei Weitem nicht perfekt, aber es erlaubt eine griffige niedrigdimensionale grafische Darstellung. Wir können sehen, dass das Hauptkomponentenverfahren ähnliche Zahlen nahe beieinander anordnet. Die Zahlen 0 und 6 liegen in derselben Gegend, ebenso 1 und 7 sowie 3 und 9. Der Raum wird grob unterteilt in 0, 4 und 6 auf der einen Seite und alle übrigen Zahlen auf der anderen.

Da die Zahlen einander zu einem guten Teil überlappen, könnte sie ein linearer Klassifikator im projizierten Raum schwer auseinanderhalten. Sollen also die handschriftlichen Ziffern klassifiziert werden und wurde als Modell ein linearer Klassifikator gewählt, reichen die ersten drei Hauptkomponenten nicht als Merkmale aus. Dennoch ist es aufschlussreich, zu sehen, wie gut ein 64-dimensionaler Datensatz in nur drei Dimensionen abgebildet werden kann.

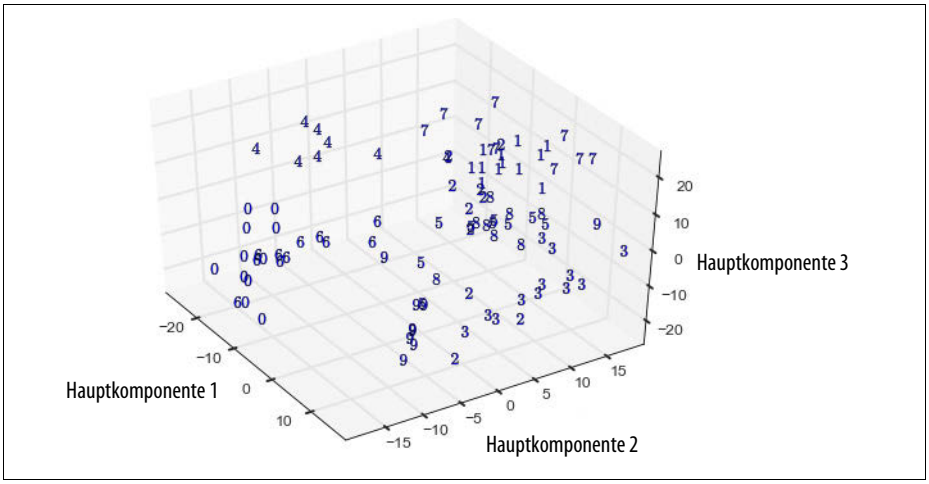

Abbildung 6-3: Hauptkomponentenprojektionen einer Untermenge der MNIST-Daten – dargestellt anhand der Bildmarkierungen

Weißen und Nullphasenverfahren

Infolge der Orthogonalitätsbedingung an die Zielfunktion hat die Hauptkomponententransformation eine hübsche Nebenwirkung: Die transformierten Merkmale korrelieren nicht mehr miteinander; die inneren Produkte zwischen Paaren von Merkmalsvektoren sind null. Das kann man leicht anhand der Orthogonalität der Singulärvektoren zeigen:

$$Z^T Z = \boldsymbol{\Sigma}_k \boldsymbol{U}_k{}^T \boldsymbol{U}_k \boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}_k{}^2$$

Das Ergebnis ist eine Diagonalmatrix, die die Quadrate der Singulärwerte enthält, die wiederum die Korrelation jedes Merkmalsvektors mit sich selbst darstellen, auch bekannt als seine ℓ^2-Norm.

Manchmal ist es hilfreich, auch die Größe der Merkmale auf 1 zu normieren. In der Sprache der Signalverarbeitung wird das als *Weißen* bezeichnet. Das Ergebnis ist ein Satz von Merkmalen mit einer Eigenkorrelation von 1 und völlig ohne Korrelation untereinander. Mathematisch ausgedrückt, bedeutet Weißen, die Hauptkomponententransformation mit den inversen Singulärwerten zu multiplizieren (siehe Formel 6-11).

Formel 6-11: Hauptkomponentenverfahren und Weißen

$$\boldsymbol{W}_{white} = \boldsymbol{V}_k \boldsymbol{\Sigma}_k{}^{-1}$$

$$\boldsymbol{Z}_{white} = \boldsymbol{X}\boldsymbol{V}_k \boldsymbol{\Sigma}_k{}^{-1} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T \boldsymbol{V}_k \boldsymbol{\Sigma}_k{}^{-1} = \boldsymbol{U}_k$$

Das Weißen ist unabhängig von der Dimensionsreduktion; jedes der beiden kann ohne das andere vorgenommen werden. Beispielsweise ist das *Nullphasenverfahren*

(ZCA, engl. Zero-phase Component Analysis, Bell and Sejnowski, 1996) eine Weißtransformation, die zwar eng mit dem Hauptkomponentenverfahren verwandt ist, aber die Anzahl der Merkmale nicht verringert. Das Nullphasen-Weißen verwendet den vollständigen Satz von Hauptkomponenten $\boldsymbol{V}$ ohne Reduktion und führt eine zusätzliche Multiplikation mit $\boldsymbol{V}^T$ durch (siehe Formel 6-12).

Formel 6-12: Nullphasen-Weißen

$$\boldsymbol{W}_{ZCA} = \boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{V}^T$$

$$\boldsymbol{Z}_{zca} = \boldsymbol{X}\boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{V}^T = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{\Sigma}^{-1} = \boldsymbol{U}$$

Eine einfache Hauptkomponentenprojektion (siehe Formel 6-10) erzeugt Koordinaten im neuen Merkmalsraum, wobei die Hauptkomponenten als Basis dienen. Diese Koordinaten bilden nur die Länge des projizierten Vektors ab, nicht aber seine Richtung. Die Multiplikation mit den Hauptkomponenten liefert uns sowohl die Länge als auch die Richtung. Eine andere zutreffende Interpretation lautet, dass die zusätzliche Multiplikation die Koordinaten zurück in den ursprünglichen Merkmalsraum rotiert. ($\boldsymbol{V}$ ist eine Orthogonalmatrix, und Orthogonalmatrizen rotieren einen Vektor, ohne ihn zu strecken oder zu stauchen.) Das Nullphasenverfahren erzeugt also geweißte Daten, die (gemessen im euklidischen Abstand) so nah wie möglich bei den ursprünglichen Daten liegen.

Bedingungen und Grenzen des Hauptkomponentenverfahrens

Setzt man das Hauptkomponentenverfahren zur Dimensionsreduktion ein, muss man sich die Frage stellen, wie viele Hauptkomponenten (k) verwendet werden sollen. Wie alle Hyperparameter kann diese Zahl anhand der Qualität des zu erzeugenden Modells gewählt werden. Es gibt aber auch Heuristiken, die ohne großen Rechenaufwand auskommen.

Eine Möglichkeit besteht darin, k so zu wählen, dass der gewünschte Anteil der Gesamtvarianz erreicht wird. (Diese Option ist im Paket `PCA` von scikit-learn verfügbar.) Die Varianz der Projektion auf die k-te Komponente beträgt:

$$\left\| \boldsymbol{X}\boldsymbol{v}_k \right\|^2 = \left\| \boldsymbol{u}_k \boldsymbol{\sigma}_k \right\|^2 = \boldsymbol{\sigma}_k{}^2$$

was gleich dem Quadrat des k-größten Singulärwerts von $\boldsymbol{X}$ ist. Die geordnete Liste der Singulärwerte einer Matrix wird als ihr *Spektrum* bezeichnet. Um festzulegen, wie viele Komponenten verwendet werden sollen, kann man also eine Spektralanalyse der Datenmatrix vornehmen und einen Schwellenwert wählen, der genügend Varianz bewahrt.

Wahl von k anhand der erreichten Varianz

Um genügend Komponenten zu behalten, sodass wir 80% der Gesamtvarianz der Daten erreichen, wählen wir k so, dass gilt:

$$\frac{\sum_{i=1}^{k} \sigma_i^2}{\sum_{i=1}^{d} \sigma_i^2} \geq 0.8$$

Eine weitere Möglichkeit der Wahl von k verwendet die intrinsische Dimension des Datensatzes. Diese ist konzeptionell etwas undurchsichtiger, kann aber auch aus dem Spektrum abgeleitet werden. Die Grundidee ist, dass man bei einem Spektrum mit ein paar großen und vielen kleinen Singulärwerten wahrscheinlich einfach die größten verwenden und den Rest verwerfen kann. Es kann vorkommen, dass der Rest des Spektrums gar nicht so klein ist, dass aber eine große Lücke zwischen den großen und den kleinen Werten klafft. Dort könnte man auch sinnvoll abschneiden. Diese Methode erfordert, dass man das Spektrum anschaut und beurteilt, und kann daher nicht als Teil einer automatisierten Pipeline ausgeführt werden.

Ein wesentlicher Kritikpunkt am Hauptkomponentenverfahren lautet, dass die Transformation ziemlich komplex ist und die Ergebnisse daher schwer zu interpretieren sind. Die Hauptkomponenten und die projizierten Vektoren sind reellwertig und können positiv oder negativ sein. Die Hauptkomponenten sind im Wesentlichen Linearkombinationen der (zentrierten) Zeilen, und die Projektionen sind Linearkombinationen der Spalten. Hat man es beispielsweise mit Aktienerträgen zu tun, ist jeder Faktor eine Linearkombination von Momentaufnahmen der Aktienerträge. Was bedeutet das aber? Die erlernten Faktoren lassen sich kaum für einen Menschen verständlich begründen. Deshalb tun sich Analysten schwer damit, den Ergebnissen zu trauen. Aber wenn Sie nicht erklären können, warum man anderer Leute Milliarden in diese oder jene Aktien investieren soll, werden Sie so ein Modell wahrscheinlich nicht verwenden.

Das Hauptkomponentenverfahren ist rechenaufwendig, da es sich auf die teure Singulärwertzerlegung stützt. Die vollständige Singulärwertzerlegung einer Matrix benötigt $O(nd^2 + d^3)$ Rechenoperationen (Golub and Van Loan, 2012), wenn $n \geq d$ ist – wenn es also mehr Datenpunkte als Merkmale gibt. Selbst wenn wir nur k Hauptkomponenten brauchen, erfordert die abgebrochene Singulärwertzerlegung (die k größten Singulärwerte und zugehörigen Vektoren) immer noch $O((n+d)^2\, k)$ = $O(n^2k)$ Schritte. Für eine große Zahl von Datenpunkten oder Merkmalen verbietet sich der Ansatz also.

Es ist schwierig, das Hauptkomponentenverfahren auf Datenströme, aufeinanderfolgende Teildatensätze oder Stichproben aus den vollständigen Daten anzuwenden. Die Berechnung der Singulärwertzerlegung im Datenstrom, ihre Aktualisierung

mit neuen Daten und ihre Berechnung anhand einer Stichprobe sind jeweils eigene Forschungsthemen. Es gibt Algorithmen dafür, aber sie führen zu verminderter Genauigkeit. Man muss also mit einer geringeren Abbildungsgenauigkeit rechnen, wenn man Testdaten auf Hauptkomponenten projiziert, die aus den Anlerndaten herrühren. Wenn sich die Verteilung der Daten ändert, müsste man stattdessen die Hauptkomponenten aus dem jeweils neuen Datensatz neu bestimmen.

Schließlich sollte das Hauptkomponentenverfahren nicht auf rohe Zählerwerte (Wortzahlen, Abspielzähler für Musik oder Filme usw.) angewandt werden. Der Grund liegt in den großen Ausreißern, die bei solchen Zählern häufig vorkommen. (Es ist ziemlich wahrscheinlich, dass es einen Fan gibt, der den *Herrn der Ringe* 314.582 Mal angesehen hat, wogegen alle übrigen Zählerwerte kümmerlich dastehen.) Wie wir wissen, sucht das Hauptkomponentenverfahren nach linearen Korrelationen zwischen den Merkmalen. Korrelationen und Varianzen sind sehr empfindlich für große Ausreißer; eine einzige große Zahl könnte die Statistik stark beeinflussen. Es empfiehlt sich also, zunächst die Daten um große Werte zu bereinigen (siehe den Abschnitt »Filtern nach Häufigkeit« auf Seite 46) oder eine Skalierung wie TF-IDF (siehe Kapitel 4) oder die Logarithmustransformation (siehe den Abschnitt »Die Logarithmustransformation« auf Seite 15) anzuwenden.

Anwendungsfälle

Das Hauptkomponentenverfahren reduziert die Dimension des Merkmalsraums, indem es nach linearen Korrelationsmustern zwischen Merkmalen sucht. Da es eine Singulärwertzerlegung vornimmt, ist es für mehr als ein paar Tausend Merkmale teuer zu berechnen. Für wenige reellwertige Merkmale ist es jedoch gewiss einen Versuch wert.

Die Hauptkomponententransformation verwirft einen Teil der Informationen in den Daten. Das so entstehende Modell mag also billiger anzulernen sein, ist aber weniger genau. Am MNIST-Datensatz wurde beobachtet, dass dimensionsreduzierte Daten aus dem Hauptkomponentenverfahren zu ungenaueren Klassifikationsmodellen führen. In solchen Fällen hat das Hauptkomponentenverfahren also Vor- und Nachteile.

Eine der geschicktesten Anwendungen des Hauptkomponentenverfahrens ist die Anomalieerkennung bei Zeitreihen. Lakhina u.a. (2004) haben damit Anomalien des Datenverkehrs im Internet aufgespürt und untersucht. Dabei haben sie sich auf Anomalien im Verkehrsaufkommen konzentriert, also auf Spitzen oder Flauten in der Menge an Datenverkehr zwischen verschiedenen Netzbereichen. Solche plötzlichen Änderungen können auf ein fehlkonfiguriertes Netz oder koordinierte Denial-of-Service-Angriffe hinweisen. In jedem Fall ist das Wissen darum, wann und wo solche Änderungen auftreten, von großem Wert für die Betreiber der Netze.

Da das Gesamtaufkommen an Datenverkehr im Internet so hoch ist, sind einzelne Spitzen in kleinen Netzbereichen schwer zu erkennen. Ein großer Teil des Daten-

verkehrs läuft über eine relativ kleine Zahl von Backbone-Knoten. Die wesentliche Erkenntnis der Forscher besteht darin, dass Anomalien im Verkehrsaufkommen immer mehrere Knoten gleichzeitig betreffen (da Datenpakete auf dem Weg zu ihrem Ziel mehrere Knoten passieren müssen). Wir fassen nun jeden der Knoten als ein Merkmal auf und die Menge an Datenverkehr in jedem Zeitschritt als den Messwert. Ein Datenpunkt ist eine Momentaufnahme der Messungen des Verkehrsaufkommens an allen Knoten des Netzes. Die Hauptkomponenten dieser Matrix zeigen allgemeine Entwicklungen des Verkehrsaufkommens im Netz an, während die übrigen Komponenten das restliche Signal darstellen, in dem die Anomalien stecken.

Das Hauptkomponentenverfahren wird auch häufig zur Modellierung im Finanzwesen herangezogen. In diesen Anwendungsfällen stellt es eine Art der *Faktoranalyse* dar – einer Familie statistischer Verfahren, um in Daten beobachtete Strukturen durch einige wenige unbeobachtete Faktoren zu beschreiben. Bei der Faktoranalyse geht es also darum, anstelle der transformierten Daten erklärungstaugliche Komponenten zu finden.

Finanzwirtschaftliche Größen wie Aktienerträge korrelieren oft miteinander. Aktien können zur selben Zeit steigen und fallen (eine positive Korrelation) oder sich in entgegengesetzte Richtungen bewegen (eine negative Korrelation). Um die Volatilität auszugleichen und das Risiko zu verringern, muss ein Investment-Portfolio eine Auswahl von verschiedenen Aktien enthalten, die nicht miteinander korrelieren. (Man sollte nicht alles auf eine Karte setzen, wenn diese Karte verlieren kann.) Korrelationsmuster zwischen Aktien zu erkennen, hilft bei der Entscheidung für eine Investitionsstrategie.

Korrelationsmuster von Aktien können sich auf ganze Branchen beziehen. Technikaktien können beispielsweise gemeinsam steigen und fallen, während Aktien von Fluglinien meist sinken, wenn Öl teuer ist. Die Branche ist aber möglicherweise nicht die beste Erklärung für ein solches Verhalten. Analysten suchen auch nach unerwarteten Korrelationen in den beobachteten statistischen Daten. So wendet das *statistische Faktormodell* (Connor, 1995) das Hauptkomponentenverfahren auf die Matrix der Zeitreihen einzelner Aktienerträge an, um Aktien zu finden, die gewöhnlich eine gewisse Kovarianz aufweisen. In diesem Anwendungsfall werden die Hauptkomponenten selbst gesucht, nicht die transformierten Daten.

Das Nullphasenverfahren kann ein nützlicher Vorverarbeitungsschritt beim Lernen aus Bildern sein. Bei natürlichen Bildern haben benachbarte Pixel oft ähnliche Farben. Beim Nullphasen-Weißen kann man diese Korrelation beseitigen, um sich beim nachfolgenden Modellieren auf interessantere Bildstrukturen zu konzentrieren. Krizhevskys Dissertation (2009) »Learning Multiple Layers of Features from Images« (*http://bit.ly/2ts42tc*) enthält schöne Beispiele, die die Auswirkung des Nullphasen-Weißens auf natürliche Bilder zeigen.

Viele Deep-Learning-Modelle verwenden das Hauptkomponenten- oder das Nullphasenverfahren als Vorverarbeitungsschritt, obwohl das nicht immer nötig ist. In »Factored 3-Way Restricted Boltzmann Machines for Modeling Natural Images« (*http://bit.ly/2D7hKkK*) bemerken Ranzato u.a. (2010), dass das Weißen zwar nicht notwendig ist, aber die Konvergenz des Algorithmus beschleunigt. Coates u.a. (2011) stellen in »An Analysis of Single-Layer Networks in Unsupervised Feature Learning« (*http://stanford.io/2oVhBvu*) fest, dass das Nullphasen-Weißen bei einigen Modellen nützlich ist, jedoch nicht bei allen. (Die Modelle in diesem Aufsatz verwenden unüberwachtes Erlernen von Merkmalen, das Nullphasenverfahren wird also zur Merkmalskonstruktion im Dienst anderer Verfahren der Merkmalskonstruktion eingesetzt. Mehrere Verfahren zu stapeln und nacheinander anzuwenden, ist gang und gäbe bei Machine-Learning-Pipelines.)

Zusammenfassung

Hier endet unsere Besprechung des Hauptkomponentenverfahrens. Wenn Sie sich zwei Dinge davon merken, dann den Mechanismus (die lineare Projektion) und das Ziel (die Varianz der projizierten Daten zu maximieren). Das Verfahren stützt sich auf die Eigenwertzerlegung der Kovarianzmatrix, die eng mit der Singulärwertzerlegung der Datenmatrix verwandt ist. Man kann sich das Hauptkomponentenverfahren auch anhand der Vorstellung einprägen, die Daten zu einem Eierkuchen flach zu drücken, der jedoch so ausgedehnt wie möglich ist.

Das Hauptkomponentenverfahren ist ein Beispiel für modellgetriebene Merkmalskonstruktion. (Man sollte immer sofort ein Modell im Hintergrund vermuten, wenn eine Zielfunktion auftaucht.) Die Modellannahme lautet dabei, dass die Varianz eine geeignete Darstellung der in den Daten enthaltenen Information ist oder – gleichbedeutend damit – dass das Modell lineare Korrelationen zwischen Merkmalen aufspürt. Dieser Umstand wird in verschiedenen Anwendungen genutzt, um die Korrelation zu verringern oder gemeinsame Faktoren in den Eingangsdaten zu finden.

Beim Hauptkomponentenverfahren handelt es sich um eine verbreitete Methode der Dimensionsreduktion. Es hat aber seine Beschränkungen, etwa der hohe Rechenaufwand und das nicht interpretierbare Ergebnis. Als Vorverarbeitungsschritt ist es jedoch nützlich, insbesondere dann, wenn lineare Korrelationen zwischen Merkmalen bestehen.

Als Technik zur Beseitigung linearer Korrelationen steht das Hauptkomponentenverfahren in Beziehung mit dem Weißen. Seine Verwandte, die Singulärwertzerlegung, weißt die Daten auf interpretierbare Weise, reduziert dabei aber nicht ihre Dimension.

Literatur

Bell, Anthony J., und Terrence J. Sejnowski. »Edges Are the ›Independent Components‹ of Natural Scenes.« *Advances in Neural Information Processing Systems* 9 (1996): 831–837.

Coates, Adam, Andrew Y. Ng und Honglak Lee. »An Analysis of Single-Layer Networks in Unsupervised Feature Learning.« *Proceedings of the 14th International conference on Artificial Intelligence and Statistics* (2011): 215–223.

Connor, Gregory. »The Three Types of Factor Models: A Comparison of Their Explanatory Power.« *Financial Analysts Journal* 51:3 (1995) 42–46.

Golub, Gene H., und Charles F. Van Loan. *Matrix Computations*. 4th ed. Baltimore, MD: Johns Hopkins University Press, 2012.

Krizhevsky, Alex. »Learning Multiple Layers of Features from Tiny Images.« MSc thesis, University of Toronto, 2009.

Lakhina, Anukool, Mark Crovella und Christophe Diot. »Diagnosing Network-wide Traffic Anomalies.« *Proceedings of the 2004 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications* (2004): 219–230.

Ranzato, Marc'Aurelio, Alex Krizhevsky und Geoffrey E. Hinton. »Factored 3-Way Restricted Boltzmann Machines for Modeling Natural Images.« *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics* (2010): 621–628.