Maschinelles Lernen 08

Prof. Dr. David Spieler – david.spieler@hm.edu

Hochschule München

1. Oktober 2019

Unüberwachte Lernmethoden

Unüberwachte Lernmethoden

Unüberwachte Lernmethoden

Bisher haben wir uns mit überwachten Lernmethoden beschäftigt, d.h. die Aufgabe war für ein gegebenes $\mathcal{D} \subseteq \mathcal{X} \times \mathcal{Y}$ die Funktion $f : \mathcal{X} \to \mathcal{Y}$ möglichst gut zu beschreiben.
Beim unüberwachten Lernen geht es uns um die Struktur einer Menge

$$\mathcal{D} = \{\mathbf{x}^{(i)} \mid 1 \leq i \leq n\} \subseteq \mathcal{X}$$

und Fragestellungen wie

- Gibt es Zusammenhänge zwischen den einzelnen Dimensionen $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_d$?
- Gibt es Gruppen innerhalb der Datenpunkte $\mathbf{x}^{(i)}$?

Unüberwachte Lernmethoden

Unüberwachte Lernmethoden sind oft herausfordernder, da aufgrund der im Kontrast zu überwachten Lernmethoden fehlenden Referenz Ergebisse und Erkenntnisse händisch überprüft bzw. interpretiert werden müssen. Zum Beispiel:

- Präzise und formal definierte Metriken wie die Genauigkeit oder Fehlerrate fehlen.
- Methoden wie die Kreuzvalidierung sind nicht anwendbar.

Hauptkomponentenanalyse

Hauptkomponentenanalyse

Hauptkomponentenanalyse

Ziel der Hauptkomponentenanalyse (Principal Component Analysis, PCA) ist die Dimensionsreduktion. Das bedeutet, gegeben $\mathcal{D} \subseteq \mathbb{R}^d$ suchen wir eine Funktion

$$g : \mathbb{R}^d \to \mathbb{R}^p$$

mit $p < d$, sodass jedes $g(\mathbf{x}^{(i)})$ den Punkt $\mathbf{x}^{(i)} \in \mathcal{D}$ möglichst gut beschreibt. Bei der PCA nimmt man an, dass g eine lineare Funktion ist.

Hauptkomponentenanalyse

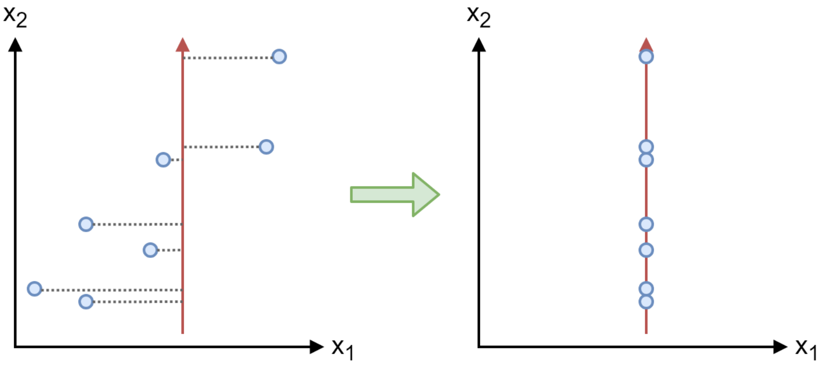

Abbildung 1: Erster Versuch einer Dimensionsreduktion $g : \mathbb{R}^2 \to \mathbb{R}$, aber es geht noch besser, die einzelnen Punkte zu separieren.

Hauptkomponentenanalyse

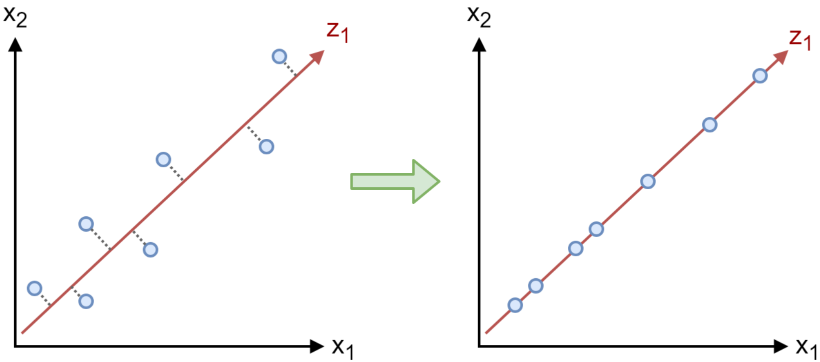

Abbildung 2: PCA mit einer Hauptkomponente, welche die Varianz maximiert und somit die Punkte im gegebenes Setting maximal separiert.

Hauptkomponentenanalyse

Die erste Hauptkomponente im Beispiel erhalten wir durch die lineare Abbildung

$$g(\mathbf{x}) = \mathbf{z}_1(\mathbf{x}) = \phi_{11}\mathbf{x}_1 + \phi_{12}\mathbf{x}_2$$

wobei $\mathbf{z}_1$ die größte Varianz haben soll und ϕ normiert ist, d.h. $\phi_{11}^2 + \phi_{12}^2 = 1$. Den Wert $\mathbf{z}_1(\mathbf{x})$ nennt man Score von Punkt $\mathbf{x}$ bzgl. der ersten Hauptkomponente und $\phi_1 = (\phi_{11}, \phi_{12})^T$ ist deren Gewichtungsvektor.

Hauptkomponentenanalyse

Um diesen Problem mathematisch fassen zu können, nehmen wir an, dass für den Mittelwert der Datenpunkte gilt

$$\mu = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}^{(i)} = \mathbf{0}.$$

Dies erreichen wir einfach durch die Berechnung des tatsächlichen Mittelwerts μ und Subtraktion von μ von jedem Datenpunkt. Wenn dies gilt, gilt auch

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{z}_1(\mathbf{x}) = 0$$

und die Varianz von $\mathbf{z}_1$ ist somit

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{z}_1(\mathbf{x}^{(i)})^2 = \frac{1}{n}\sum_{i=1}^{n} (\phi_{11}\mathbf{x}_1^{(i)} + \phi_{12}\mathbf{x}_2^{(i)})^2.$$

Hauptkomponentenanalyse

Für den Gewichtungsvektor der ersten Hauptkomponente erhalten wir somit das Optimierungsproblem

$$\arg\max_{\phi_{11},\phi_{12}} \frac{1}{n}\sum_{i=1}^{n}(\phi_{11}\mathbf{x}_1^{(i)} + \phi_{12}\mathbf{x}_2^{(i)})^2$$

unter der Bedingung

$$\phi_{11}^2 + \phi_{12}^2 = 1.$$

Hauptkomponentenanalyse

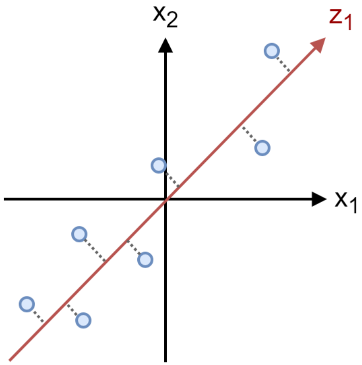

Abbildung 3: Im Beispiel erhalten wir $(\phi_{11}, \phi_{12})^T = \frac{1}{\sqrt{2}}(1,1)^T$

. Ein Punkt $(2,1)^T$ wird demnach bei einer Hauptkomponente auf den Punkt $z_1 = \frac{1}{\sqrt{2}}2 + \frac{1}{\sqrt{2}}1 = \frac{3}{\sqrt{2}}$ abgebildet.

Hauptkomponentenanalyse

Wir können nun auch das Optimierungsproblem im allgemeinen Fall für $g : \mathbb{R}^d \to \mathbb{R}^p$ formulieren als

$$\arg\max_{\phi_{11},\dots,\phi_{1p}} \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{d}\phi_{1j}\mathbf{x}_j^{(i)}\right)^2$$

unter der Bedingung

$$\sum_{j=1}^{d}\phi_{1j}^2 = 1.$$

Solche Probleme kann man mit Hilfe von *Eigenwertzerlegungen* lösen. Wir werden uns jedoch nicht explizit damit beschäftigen.

Hauptkomponentenanalyse

Die weiteren Hauptkomponenten $\mathbf{z}_2, \ldots, \mathbf{z}_p$ mit Gewichtungsvektoren $\phi_2, \ldots, \phi_p$ erhalten wir sukzessive, indem wir weiter nach den varianzmaximierenden Richtungen suchen, die jeweils unkorreliert bzgl. den vorherigen Richtungen – also orthogonal – sind.

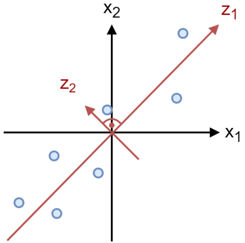

Abbildung 4: Die zweite Hauptkomponente steht seknkrecht auf der ersten Hauptkomponente. Falls wir weitere Dimensionen hätten, müsste $\mathbf{z}_2$ auch die verbliebene Varianz maximieren.

Hauptkomponentenanalyse
Beispiel

Beispiel: Verhaftungen in den USA

Wir folgen nun dem Beispiel aus [JWHT14] und betrachten den (wichtig: *normierten*) *USArrests* Datensatz, welcher für das Jahr 1973 und die $n = 50$ Bundesstaaten der USA jeweils $d = 4$ Features enthält:

- `Murder`: Anzahl der Verhaftungen wegen *Mordes*,
- `Assault`: *Überfall*,
- `Rape`: und *Vergewaltigung*, gerechnet auf 100.000 Einwohner, sowie
- `UrbanPop`: die prozentuale Anteil der Bewohner in *Städten*.

Hauptkomponentenanalyse
Beispiel

Mit Hilfe der PCA lassen sich die ersten beiden Hauptkomponenten bestimmen als

	ϕ_1	ϕ_2
`Murder`	0.5358995	-0.4181809
`Assault`	0.5831836	-0.1879856
`UrbanPop`	0.2781909	0.8728062
`Rape`	0.5434321	0.1673186

Tabelle 1: Die ersten beiden Hauptkomponenten des USArrests Datensatzes.

Hauptkomponentenanalyse

Beispiel

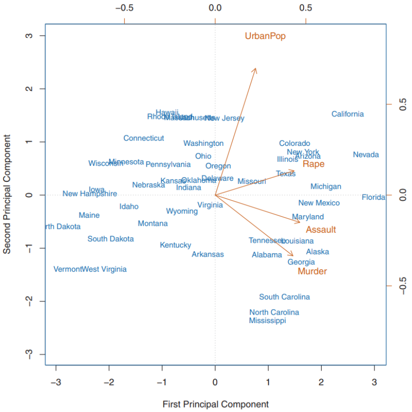

Abbildung 5: PCA ermöglicht die Darstellung des 4-dimensionalen *USArrests* Datensatzes durch die Projektion auf die ersten beiden Hauptkomponenten [JWHT14].

Hauptkomponentenanalyse

Beispiel

Interpretation:

- Die erste Hauptkomponente mit $\phi_1 \approx (0.6, 0.6, 0.3, 0.5)^T$ korrespondiert sehr stark mit den Arten von Verbechen (`Murder`, `Assault`, `Rape`) und weniger mit dem Anteil der Stadtbewohner (`UrbanPop`).
- Die zweite Hauptkomponente mit $\phi_2 \approx (-0.4, -0.2, 0.9, 0.2)^T$ korrespondiert sehr stark mit dem Anteil der Stadtbewohner und weniger mit den Verbrechen.
- Die Verbrechensarten korrelieren daher sehr stark miteinander, d.h. Staaten mit vielen Morden haben meist auch viele Überfälle bzw. Vergewaltigungen und umgekehrt.
- Die Anzahl an Verbrechen sind nicht sehr stark abhängig davon, ob es viele Stadtbewohner gibt.

Hauptkomponentenanalyse
Hinweise

Je nach Bibliothek und Optimierer, welcher verwendet wird, können sich die Hauptkomponenten im Vorzeichen unterscheiden. Die PCA ist demnach nicht eindeutig, da wenn zwei Vektoren $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$ orthogonal zu einander sind, d.h.

$$\mathbf{u} \circ \mathbf{v} = 0$$

automatisch auch

$$\mathbf{u} \circ (-\mathbf{v}) = (-\mathbf{u}) \circ \mathbf{v} = (-\mathbf{u}) \circ (-\mathbf{v}) = 0$$

gilt.

Hauptkomponentenanalyse
Hinweise

Die PCA sollte zu einem guten Verstädnis der Daten führen. Zu entscheiden wie viele Hauptkomponenten verwendet werden sollen, ist jedoch eine Kunst. Ein Hilfsmittel dafür ist der Anteil der erklärten Varianz (*proportion of variance explained*, PVE) der m-ten Hauptkomponente im Verhältnis zur kompletten Varianz in den Daten, d.h.

$$PVE_m = \frac{\sum_{i=1}^{n} \mathbf{z}_m(\mathbf{x}^{(i)})^2}{\sum_{i=1}^{n} \sum_{j=1}^{p} \left(\mathbf{x}_j^{(i)}\right)^2}.$$

Hauptkomponentenanalyse

Hinweise

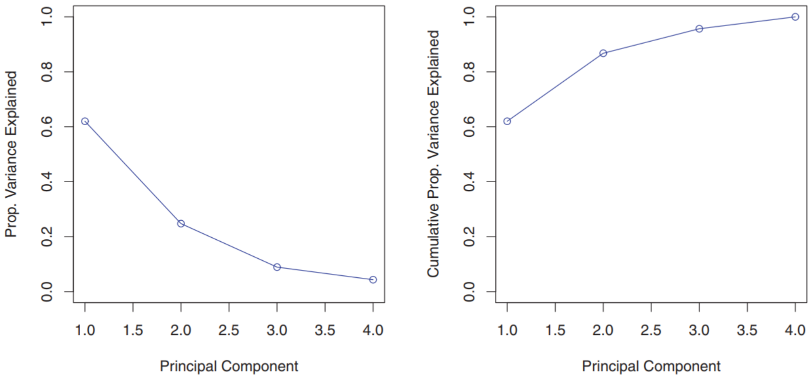

Abbildung 6: Ein *Scree-Plot* der PVE im *USArrests* Beispiel [JWHT14]. Die ersten beiden Hauptkomponenten erklären einen Großteil, d.h. mehr als 80%, der Varianz in den Daten.

Hauptkomponentenanalyse

Hinweise

Die PCA kann als Hilfsmittel für die Dimensionsreduktion auch *vorgeschaltet* für überwachte Lernmethoden verwendet werden. Soll im eigentlichen überwachten Problem eine Funktion

$$f : \mathbb{R}^d \to \mathcal{Y}$$

auf Daten $\mathcal{D} \subseteq \mathbb{R}^d \times \mathcal{Y}$ gelernt werden, wobei sowohl Klassifikation als auch Regression möglich sind, so kann evtl. mit Hilfe der PCA $g : \mathbb{R}^d \to \mathbb{R}^p$ mit $p < d$ das Problem auf

$$f' : \mathbb{R}^p \to \mathcal{Y}$$

welches auf den Daten $\{(g(\mathbf{x}^{(i)}), \mathbf{y}^{(i)}) \mid (\mathbf{x}^{(i)}, \mathbf{y}^{(i)}) \in \mathcal{D}\}$ trainiert wird reduziert werden.

Hauptkomponentenanalyse

References

G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning: With applications in r*, Springer Publishing Company, Incorporated, 2014.