Blatt 03 - Principal Components Analysis

Abgabe: Bis Mittwoch, 12.November 2014, 10:00 Uhr
An: Tobias Rothenberger, <rothenb@informatik.uni-frankfurt.de>

Aufgabe 3.1

a. Betrachten Sie eine Menge von 400 Punkten, gegeben durch:

$$p^{(i)} = \begin{bmatrix} p_x^{(i)} \\ p_y^{(i)} \end{bmatrix} = 3 \cdot r_i^3 \cdot \begin{bmatrix} \cos(2\pi \cdot w_i) \\ \sin(2\pi \cdot w_i) \end{bmatrix} + v^{(i)}, \qquad v^{(i)} = \begin{cases} [4,4]^T & ,0 \\ [6,7]^T & ,20 \end{cases}$$

Die w_i und die r_i werden dabei uniform verteilt aus dem Intervall $[0,1]$

Verwenden Sie das Sangernetzes, um eine PCA durchzuführen und Koordinaten dieser Punkte zu dekorrelieren. Stellen Sie anschließend ursprüngliche Punktwolke mit den gelernten neuen Koordinatenachse transformierte Punktwolke graphisch dar.

nach oben

b. Nach der Transformation eines Datenvektors mit Hilfe des Sangerne die Komponenten der neuen Datenvektoren nicht nur dekorreliert, sor absteigend nach ihrer Varianz geordnet. Dies ist eine sehr nützliche E da auf diese Weise bekannt ist, in welchen Komponenten am meisten Struktur der ursprünglichen Datenwolke enthalten ist. In vielen Fällen

eine Analyse von hochdimensionalen Daten dann auf die Analyse einer niedrigdimensionaleren Menge beschränken, indem man die Komponenten mi niedriger Varianz vernachlässigt.

Um dies zu verdeutlichen, betrachten wir nun eine ähnliche Punktmenge wie oben, bestehend aus 600 Punkten, die gegeben sind durch:

$$p^{(i)} = \begin{bmatrix} p_x^{(i)} \\ p_y^{(i)} \end{bmatrix} = 4 \cdot r_i^3 \cdot \begin{bmatrix} \cos(2\pi \cdot w_i) \\ \sin(2\pi \cdot w_i) \end{bmatrix} + v^{(i)}, \qquad v^{(i)} = \begin{cases} [4,4]^T & ,0 \\ [7,4]^T & ,20 \\ [7,7]^T & ,40 \end{cases}$$

Stellen Sie diese Menge wieder in einem Plot dar und berechnen Sie eine dekorrelierende Transformation. Anschließend betrachten wir nur noch die ers Komponente der transformierten Punkte. Fertigen Sie ein Histogramm für dies Werte an und verwenden Sie dazu 100 Einteilungen des Intervalls $[-$ stellen Sie fest, wenn Sie die Werte mit dem 2-dimensionalen Plot der originale Datenpunkte vergleichen? Wie würde das Histogramm aussehen, wenn Sie di x- oder die y-Koordinaten der ursprünglichen (nicht transformierten) Punkte genommen hätten?

nach oben

c. Was ändert sich jeweils am Ergebnis der PCA, wenn die Punkte im Eingaberaum alle in die selbe Richtung verschoben (Translation), um den gleichen Winkel um einen beliebigen Punkt im Raum gedreht (Rotation), oder die Abstände zu einem beliebigen Punkt unter Beibehaltung der Richtung mit dem gleichen Faktor (ungleich Null) multipliziert werden (Skalierung)? Warum dies so? Ändert sich jeweils das Aussehen des Histogramms, wenn Sie eine beliebige Kombination aus Translation, Rotation und Skalierung im Eingaberau durchführen und bezüglich des neuen Wertebereichs der Zahlen die gleiche

Anzahl an Einteilungen verwenden, wie im vorigen Aufgabenteil?

nach oben

Aufgabe 3.2

a. Die PCA kann auch dazu eingesetzt werden, die dimensionale Ausdehnung einer Menge von Datenvektoren zu ermitteln. Vor allem Datensätze, die aus Vektoren mit sehr vielen Komponenten bestehen, haben meistens die Eigenschaft, daß die Menge der Datenpunkte einen niedrigdimensionaleren Raum ausfüllt.

   Lesen Sie die Trainingsdaten des Card-Datensatzes (`card1.dt`) ein (Spalten bis 51) und bestimmen Sie mit Hilfe des Sangernetzes die Dimensionale Ausdehnung dieser Daten. (Diese können Sie an den Eigenwerten ablesen...) Plotten Sie außerdem auch eine Kurve der Eigenwerte, wie sie vom Netz berechnet wurden. Hieran läßt sich erkennen, wie relevant die Komponenten d Transformierten Datenvektoren sind, um daraus die ursprünglichen Daten wiederherzustellen. Niedrige Eigenwerte bedeuten, daß diese Komponenten fa keine Rolle mehr spielen. Das Vernachlässigen solcher Komponenten bei weiteren Analysen, die zu niedrigen oder verschwindenden Eigenwerten gehören, bezeichnet man dabei auch als *Dimensionsreduktion*.

nach oben

b. In der ursprünglichen Variante berechnet der Sangeralgorithmus keine Eigenvektoren, die zu Eigenwerten 0 gehören. Erweitern Sie nun den Sangeralgorithmus, so daß auch Eigenvektoren zu diesen Eigenwerten

berechnet werden können. Beschreiben Sie, wie diese Vektoren dazu verwend werden können, Spalten des Card-Datensatzes aus Aufgabenteil b) zu identifizieren, die keine neue Information über diesen Datensatz tragen (und damit prinzipiell unnötig sind). Bestimmen Sie dann eine mögliche Kombinatio dieser Spalten.

nach oben

Aufgabe 3.3

a. Eine weitere Anwendungsmöglichkeit für die PCA ist die Datenkompression. Wir werden hier allerdings kein komplettes Datenkompressionssystem betrachten, sondern lediglich eine Vorverarbeitungsstufe, die jedoch recht interessante Ergebnisse liefert.

   Lesen Sie dazu das Foto aus der Datei `pca_reptil_640x480.jpg` zerlegen Sie jeden der drei Farbkanäle in disjunkte Blöcke aus 8x8 Pixel. Lass Sie anschließend durch das PCA-Netz eine dekorrelierende Transformation fü die 64 Komponenten dieser Blöcke lernen. Dekorrelieren Sie anschließend jed Block und rekonstruieren Sie ihn mit den k Koeffizienten, die zu den Eigenvektoren mit den k größten Eigenwerten gehören. Führen Sie dies mit de Werten k=1,2,...,63 durch. Plotten Sie anschließend für jeden Farbkanal eine Kurve, auf der zu sehen ist, wie groß die durchschnittliche Abweichung eines rekonstruierten Wertes zum Originalbild bei einer bestimmten Anzahl vernachlässigter Komponenten ist. Die Kurven können dabei in den selben Plo gezeichnet werden.

   Wie viele Komponenten können insgesamt weggelassen werden, so daß die

durchschnittliche Abweichung maximal 1 bzw. 2 beträgt?

nach oben

b. Die Rekonstruktionsqualität kann noch etwas verbessert werden, indem man als erstes zusätzlich noch die Farbkanäle dekorreliert. Auf diese Weise sinkt sowohl der durchschnittliche Fehler, als auch die Anzahl der benötigten Werte zur Rekonstruktion des Bildes.

   Dekorrelieren Sie daher mit Hilfe des PCA-Netzes die Farbkanäle des Fotos und wiederholen Sie dann die obige Vorgehensweise. Da die dekorrelierten Farbkanäle nun unterschiedlich stark zum Foto beitragen, ist es sinnvoll, auch die Anzahl der vernachlässigten Koeffizienten für jeden Kanal unterschiedlich z wählen. Lassen Sie daher bei den 8x8-Blöcken des zweiten dekorrelierten Farbkanals bei der Rekonstruktion etwa 25% mehr Koeffizienten weg, als bei den Koeffizienten des ersten dekorrelierten Farbkanals. Beim dritten Kanal können Sie etwa 50 % mehr weglassen. Behalten Sie aber mindestens einen Koeffizienten pro Block. Plotten Sie dann wieder die entsprechenden Fehlerkurven wie im ersten Aufgabenteil.

   Wie viele Koeffizienten können nun etwa vernachlässigt werden, um eine Abweichung von bis zu 1 bzw. 2 hervorzurufen?

nach oben