2. Projekt: Eigenwertberechnung und Hauptkomponentenanalyse

Abgabe eines Berichtes (+ Programm) in moodle bis spätestens Donnerstag 10.6. 2019 (23:55). Jede(r) muss einen selbst erstellten Bericht und den (eventuell im Zweierteam) entwickelten Programm-Code hochladen.
Abnahme der Programme in den Übungen am 12./13.6. .

Aufgabenstellung:

(a) Implementieren Sie in MATLAB eine Funktion zur Berechnung aller Eigenwerte und Eigenvektoren einer quadratischen symmetrischen Matrix $A \in \mathbb{R}^{n\times n}$. Sie können annehmen, dass alle Eigenwerte unterschiedlichen Betrag haben.

(b) Testen Sie Ihre Implementierung an einfachen Beispielen.

(c) Bearbeiten Sie das Anwendungsproblem Hauptkomponentenanalyse zur Unterscheidung von Weinsorten. Visualisieren Sie, dass die ersten zwei bzw. die ersten drei Hauptkomponenten der multivariaten Daten mit 13 Merkmalen einen Hauptteil der Variabilität der Daten beschreiben und ausreichend sind, um einen unbekannten Wein mit vorliegenden 13 Merkmalen einem der drei vorliegenden Weinsorten zuzuordnen.

Hinweise zur Implementierung

QR-Verfahren

(i) Implementieren Sie zur Berechnung der Eigenwerte das QR-Verfahren mit Spektralverschiebung und Dimensionsreduktion in einer eigenen Funktion.

- Bringen Sie zunächst die Matrix A in Tridiagonalform (Hesse-Form). Passen Sie dazu ihre QR-Zerlegung mit Householder-Reflexionen aus dem 1.Projekt geeignet an.
- Verwenden Sie für die QR-Zerlegung im QR-Verfahren die bereitgestellte Funktion `qrGivensHesse` (die wiederum `givensParameter` benutzt). Sie sollen nachvollziehen, was diese Funktion macht.
- Verwenden Sie für die Spektralverschiebung mit Dimensionreduktion die in der Vorlesung besprochene Strategie: Sei $T \in \mathbb{R}^{l\times l}$, $l = n - m$, der linke obere $l \times l$ Block der Iterationsmatrix $A^{(k)}$ im k-ten Schritt (also „nach Streichen von m Zeilen und Spalten“), dann wähle man $\mu_k = T(l, l)$ als Parameter für die Spektralverschiebung.
- Als Fehlerschätzer für die Berechnung des Eigenwertes λ_l verwende man

$$err = \frac{|T(l, l-1)|}{|T(l, l)| + |T(l-1, l-1)|},$$

der gegen eine übergebene Toleranz getestet werden soll. Falls die Toleranz erreicht ist, so wird $T(l, l-1) = 0$ gesetzt und eine weitere Zeile und Spalte wird gestrichen. (In der Implementierung werden keine Zeilen und Spalten gestrichen sondern man arbeitet mit der Matrix `A(1:l,1:l)` ...).

(ii) Verwenden Sie zur Berechnung der Eigenvektoren die inverse Vektoriteration, die Sie in einer eigenen Funktion bereitstellen, siehe Übungsblatt 4. Als Spektralverschiebung verwenden Sie dann jeweils einen der berechneten Eigenwerte λ_i. Nach wenigen Iterationen ist der zugehörige Eigenvektor v_i bereits sehr genau bestimmt (warum ?).

Hauptkomponentenanalyse von Weinsorten.

Untersucht werden soll der Datensatz `http://archive.ics.uci.edu/ml/machine-learning-databases/wine/`. Es wurden 173 Weinsorten von drei verschiedenen Winzern einer chemischen Analyse unterzogen. Für jeden Wein sind die Werte von 13 Merkmalen angegeben. In der ersten Spalte der Datenmatrix steht die Herkunft/Sorte des Weines (Zahlen 0,1,2). Genaueres finden Sie in den Dateien `winedata.txt` und `winedata.names`. Die Daten werden in MATLAB/Octave mit dem Befehl `load winedata.txt` geladen und sind dann in der Matrix mit Bezeichner `winedata` verfügbar.
Die Idee der Hauptkomponentenanalyse (englisch principal component analysis, PCA) ist wie folgt (für eine detailliertere Beschreibung konsultiere man ein Statistiklehrbuch oder Wikipedia, oder man warte bis zu der LV schließenden Statistik 2 ;)): Wir betrachten n statistische Variablen X_i, $i = 1, \dots, n$ mit Werten x_{ji}, $j = 1, \dots m$ - in unserem Fall ist $n = 13$ (Anzahl der Analysewerte) und $m = 178$ (Anzahl der verschiedenen Weine). Die Variable X_i beschreibt das i-te Merkmal eines Weins, also den i-ten Analysewert. Die Variablen X_i streuen und sind korreliert. Wir suchen einige wenige neue Variablen, die sich als Linearkombinationen der Variablen X_i ausdrücken lassen, unkorreliert sind und den Großteil der Variabilität (Varianz) der Daten beschreiben. Diese nennt man dann auch Hauptkomponenten. Sie ergeben sich aus den Eigenvektoren der Kovarianzmatrix zu den größten Eigenwerten. Dazu gehen wir folgendermaßen vor:

(i) **Datenvorverarbeitung:** Im allgemeinen haben die einzelnen Merkmale/Variablen ganz unterschiedliche Größenordnungen, Einheiten etc. Es ist daher sinnvoll, zunächst eine Zentrierung und Skalierung der Variablen vorzunehmen: es sei $\mu_i = \sum_j x_{ji}/m$ der Mittelwert und $\sigma_i = \sqrt{\sum_j (x_{ji} - \mu_i)^2/(m-1)}$ die Standardabweichung von X_i, dann gehen wir über zu den neuen Variablen

$$Y_i := (X_i - \mu_i)/\sigma_i, \quad i = 1, \dots, n.$$

Die Variablen Y_i haben den Mittelwert 0 und die Varianz 1.

(iii) Nun betrachten wir die Kovarianzmatrix C der Variablen $Y_1, \dots, Y_n$:

$$C_{ij} := \mathrm{Cov}(Y_i, Y_j) = \sum_l y_{li}\, y_{lj}/(m-1), \quad \text{also} \quad C = Y^T Y/(m-1)$$

(C ist gerade die Korrelationsmatrix der ursprünglichen Variablen $X_1, \dots, X_n$.) Auf der Diagonalen stehen die Varianzen der Variablen Y_i, und die Gesamtvarianz ergibt sich zu

$$\mathrm{var} = \sum_i C_{ii} = m.$$

Da die Matrix C s.p.d. ist, lässt sie sich orthogonal diagonalisieren mit Eigenwerten λ_i und zugehörigen normierten Eigenvektoren $v_i = (v_{1i}, \dots, v_{ni})^T$:

$$D = V^T C V, \quad D = \mathrm{diag}(\lambda_1, \dots, \lambda_n)$$

(iv) Zu einem normierter Eigenvektor $v_i = (v_{1i}, \dots, v_{ni})^T$ beschreibt die statistische Variable

$$Z_i := \sum_j v_{ji} Y_j, \ \text{also} \quad Z = YV,$$

den Anteil der Merkmale in Richtung v_i (Orthogonalprojektion). Jeder n-dimensionale Merkmalsvektor eines Weines wird also auf die Richtung v_i projiziert. Wir erhalten damit für jeden Wein 13 neue Merkmale - seine *Hauptkomponenten*. Es gelten die folgenden Eigenschaften (machen Sie sich das klar !)

  - Z_i hat die Varianz λ_i
  - die Kovarianzmatrix der Variablen Z_i, $i = 1, \ldots, n$ ist diagonal, auf der Diagonalen stehen die Eigenwerte $\lambda_1, \ldots, \lambda_n$, denn

    $$D = Z^T Z/(m-1)$$

    In diesem Sinne sind die neuen Variablen Z_i unkorreliert.
  - die Gesamtvarianz der Variablen Y_i ergibt sich jetzt als die Summe der Eigenwerte der Kovarianzmatrix C:

    $$\text{var} = \sum_i \lambda_i.$$

(v) Man bezeichnet die neuen Variablen Z_i als die **Hauptkomponenten**. Eigenvektoren mit großen Eigenwerten definieren also Hauptkomponenten, die eine große Varianz haben und damit hauptsächlich zur Gesamtstreuung beitragen. Nur Eigenvektoren zu Eigenwerten, die größer als 1 sind, führen zu Hauptkomponenten, die eine größere Varianz haben, als die ursprünglichen Variablen Y_i.

(vi) Stellt man die ersten zwei oder auch ersten drei Hauptkomponenten – also diejenigen zu den größten Eigenwerten – des Datensatzes (mit `scatter` bzw. `scatter3`) grafisch dar (Datenpunkte entsprechend der Herkunft des Weines in drei unterschiedliche Farben eingefärbt), so zeigt sich, dass mit diesen zwei bzw. drei Merkmalen bereits eine gute Unterscheidung möglich ist.

Projektbericht

Der in ausgedruckter Form abzugebende Bericht soll enthalten:

- Titel des Projektes, Kurs, Name, Datum
- Vorstellung des Projektthemas (nicht der Aufgabenstellung !!), ein paar Sätze
- Beschreibung der Problemstellung und der verwendeten numerischen Verfahren.
- Beschreibung wichtiger Aspekte der Implementierung.
- Darstellung und Diskussion der Ergebnisse, insbesondere auch übersichtliche Grafiken in angemessener Größe mit lesbarer Achsenbeschriftung, Legende, Bildunterschrift etc. .
- Kein langes Code-Listing (nur kürzere wichtige Ausschnitte, kommentiert, falls daran etwas erklärt wird im Text)

Weitere Hinweise:

- Insgesamt sollte der Bericht ca. 3-5 Seiten lang sein.
- Der Bericht darf nicht als Gruppenarbeit erstellt werden. Es reicht nicht, die Formatierung und ein paar Nebensätze etc. zu verändern. Sollten zwei Berichte hinreichend ähnlich sein, dann wird er nur für eine Person gewertet.
- Copy-Paste aus Wikipedia oder anderen Quellen ist nicht erlaubt!
- Formeln dürfen nicht aus anderer Quelle als Bild eingefügt werden.

Weitere Hinweise zur Abgabe / Benotung

(i) Kopieren und anschließendes Anpassen von Programmier-Code zählt genauso wie das Abschreiben bei einer Klausur als Betrugsversuch. Sollte ich Selbiges (bei der Abhname oder bei der Korrektur der Berichte bzw. des hochgeladenen Programmier-Codes) feststellen, wird das Projekt als nicht bestanden bewertet. Das Erstellen von Plagiaten ohne Angabe von Quellen ist insbesondere im wissenschaftlichen/akademischen Bereich nicht nur unsportlich sondern eben Betrug.

(ii) Der Programm-Code darf in Zweiterteams erstellt werden. Im Code muss das Autorenteam genannt werden.

(iii) Für die Abnahme des Projektes wird es Einzeltermine geben, also keine gemeinsame Teamvorführung. Sie müssen mir dann den gesamten Code erklären können, so dass ich davon überzeugt bin, dass Sie wesentlich an der Erstellung des Codes mitgearbeitet haben.

(iv) In die Benotung geht der Projektbericht und auch die Qualität des Matlab-Codes ein.

(v) Die Bearbeitung des Projektes darf auch Spaß machen.