Praktikum Maschinelles Lernen und Datenanalyse

Sommersemester 2007

Fachgebiet Maschinelles Lernen
Institut für Softwaretechnik und
theoretische Informatik
Fakultät IV, Technische Universität Berlin
Prof. Dr. Klaus-Robert Müller
Email: krm@cs.tu-berlin.de

Übungsblatt 3: Clustering

Abgabeschluss: Montag, der 4.06.2007 um 10:00 Uhr.

Für dieses Aufgabenblatt sind sowohl Code als auch eine schriftliche Ausarbeitung abzugeben. Sendet Euren Code an `mikio@cs.tu-berlin.de` **und** `buenau@cs.tu-berlin.de` **mit Subject** "`ML-Praktikum Abgabe` *`Name`*". Gebt Eure Ausarbeitungen in unserem Sekretariat (FR 6-9) bei Frau Gerdes (Raum FR 6052) ab.

Bitte beachtet weiter die Coding-Richtlinien von Blatt 1

Aufgaben

Aufgabe 1: Implementation (30 Punkte)

K-means Clustering

Implementiere den K-means Clustering Algorithmus als Funktion

```
[ mu, r ] = kmeans(X, k, max_iter, prog_fun)
```

welche fuer die Spalten der $d \times n$ Matrix `X` die $d \times k$ Matrix der `k` Cluster Zentren `mu` sowie den n-dimensionalen Vektor `r` der Zugehoerigkeiten berechnet. Dabei enthaelt der i-te Eintrag von `r` den Index des Clusters zu dem der i-te Datenpunkt gehoert. Der Algorithmus soll terminieren, wenn sich die Zugehoerigkeiten nicht mehr aendern, spaetestens jedoch nach `max_iter` Schritten (optionaler Parameter mit Standardwert 100). Die Funktion soll nach jeder Iteration folgende Informationen ausgeben.

- Die Nummer der Iteration.
- Die Anzahl der Cluster-Zugehoerigkeiten, die sich in diesem Schritt geaendert haben.
- Den Wert der Fehlerfunktion (siehe Skript).

Der optionale Parameter `prog_fun` ist ein handle auf eine Funktion (siehe Matlab Funktion `feval`), welche nach jedem Schritt aufgerufen wird, um den ueber den Fortschritt des Algorithmus zu informieren. Die Signatur ist

```
prog_fun(X, mu, r)
```

wobei `mu` die aktuellen Clusterzentren, `r` die aktuellen Zuweisungen und `X` die Daten sind.

Visualisierungsfunktion fuer K-Means auf USPS Daten

Schreibe eine Visualisierungsfunktion fuer K-means Clustering (siehe Argument `prog_fun`) mit dem Namen `plot_kmeans_USPS` welche fuer eine beliege Anzahl Cluster die aktuellen Zentren in einer figure als 16×16 Bilder (in graustufen) darstellt und auf Tastendruck wartet. Die Zentren sollen mit den jeweiligen Cluster-Indices beschriftet sein.

Hierarchical Clustering

Implementiere stepwise optimal hierarchical agglomorative clustering mit K-means Kriterium als Funktion

```
[ R, kmloss, mergeidx ] = kmeans_agglo(X, r)
```

welche fuer die Spalten der $d \times n$ Matrix `X` mit initialer Clusterloesung beschrieben durch den $1 \times n$ Zugehoerigkeitsvektor `r` eine hierarchische Clusterloesung berechnet. Das Ergebnis soll in folgendem Format zurueckgegeben werden.

- `R` ist eine $(k-1) \times n$ Matrix welche fuer jeden Schritt die Zugehoerigkeiten enthaelt, jede Zeile ist also eine Clusterloesung.
- `kmloss` ist ein $k \times 1$ Vektor, welcher den Wert der Kostenfunktion in jedem Schritt enthaelt.
- `mergeidx` ist eine $(k-1) \times 2$ Matrix, welche in jeder Zeile die Indices der vereinigten Cluster enthaelt.

Dendrogram Plots

Implementiere eine Funktion, welche zu einer gegebenen hierarchischen Clusterloesung einen Dendrogram-Plot erstellt:

```
agglo_dendro(kmloss, mergeidx)
```

Die Parameter `kmloss` und `mergeidx` entsprechen den Ergebnissen von `kmeans_agglo`. Im Skript gibt es ein Beispiel fuer einen Dendrogram Plot.

EM-Algorithmus fuer Gausschsche Mixturmodelle

Implementiere den EM-Algorithmus für Gausssche Mixturmodelle als Funktion

```
[pi, mu, sigma] = em_mog(X, k, max_iter, init_kmeans, prog_fun)
```

wobei

Ausgabe	`pi`	$1 \times k$-Matrix der $\hat{\pi}_k$
	`mu`	$1 \times k$-Matrix der $\hat{\mu}_k$ (Mittelpunkte)
	`sigma`	Cell-array der Länge k der $d \times d$ Kovarianzmatrizen $\hat{\Sigma}_k$
Eingabe	`X`	$d \times n$-Matrix der Datenpunkte
	`k`	Anzahl der Gaussverteilungen
	`max_iter`	Optional: Maximale Anzahl der Iterationen (default: 100)
	`init_kmeans`	Optional: Initialisierung durch K-Means Clusterloesung (default: 0)
	`prog_fun`	Optional: Name oder handle der Visualisierungsfunktion (default: `[]`)

Die Visualisierungsfunktion `prog_fun` wird nach jedem Schritt aufgerufen, um über den Fortschritt zu informieren; die Signatur lautet

```
prog_fun(X, mu, sigma)
```

wobei `X` die Daten und `mu` und `sigma` die aktuellen Parameter des geschaetzten Mixturemodels sind. Wenn `init_kmeans` den Wert 1 hat, dann werden Mittelpunkte, Kovarianzmatrizen und Mixturekoeffizienten mit dem Ergebnis einer K-Means Clusterloesung initialisiert.

Die Funktion soll nach jedem Schritt die Nummer der Iteration und die log likelihood pro Datenpunkt ausgeben. Der Algorithmus soll terminieren, wenn die maximale Anzahl an Iteration `max_iter` erreicht wurde oder sich die log likelihood nicht mehr ändert, also ein lokales Maximum erreicht wurde.

Visualisierungsfunktion fuer EM auf 2D-Daten

Schreibe für den zweidimensionalen Fall eine Visiualisierungsfunktion fuer den EM-Algorithmus (siehe Argument `prog_fun`) mit den Name `plot_em2d` welche die Daten plotet sowie die Kovarianzmatrizen `sigma` durch Ellipsen darstellt und auf Tastendruck wartet. Tipp: Die Eigenzerlegung von `Sigma` liefert die Hauptachsen, und die Wurzeln aus den Eigenwerten die Radien.

Aufgabe 2: Anwendung (20 Punkte)

Fünf Gaußverteilungen

Analysiere den `5gaussians` Datensatz mit allen Methoden für $k = 2, \ldots, 10$ Cluster. Was fällt auf? Sind alle Methoden in der Lage, die 5 Cluster zuverlässig zu finden, d.h. unterscheiden sich die Loesungen beispielsweise in Abhängigkeit von der zufälligen Initialisierung? Welchen Einfluss hat die Initialisierung des EM-Algorithmus mit einer K-Means Loesung auf die Anzahl benoetigter Iteration und die Qualität der Lösung? Wie sieht das Dendrogramm der hierarchischen Clusterlösung aus? Kann man anhand des Dendrogramms eine Schätzung fuer eine geeignetes k abgeben?

Zwei Gaußverteilungen

Analysiere den `2gaussians` Datensatz mit k-means und dem EM Algorithmus. Welcher Algorithmus funktioniert besser, und wieso? Was liefert hierarchisches Clustern? Wie hängt die Lösung des EM-Algorithmus von der Initialisierung ab?

USPS-Datensatz

Wende EM und K-Means Clustering auf den USPS Datensatz mit $k = 10$ an. Welcher Algorithmus liefert bessere Ergebnisse? Erstelle ein Dendrogramm zu einer hierarchischen Clusterloesung und zusätzlich einen Plot, der zu jedem agglomorativen Schritt die Mittelpunkte der Clusterzentren als 16×16 Bilder zeigt.

Zusatzaufgabe (15 Punkte)

Entwickle oder recherchiere eine Methode zur Bestimmung der Anzahl Cluster k gegeben einen Datensatz. Beschreibe und implementiere die Methode. Demonstriere die Ergebnisse anhand der Datensätze dieses Übungsblattes.