

Maschinelles Lernen und Data Mining

11. Übung

Prof. Tobias Scheffer
Steffen Bickel

WS06/07

Ausgabe am: 29.01.07
Besprechung am: 02.02.07

Aufgabe 1 (1/4 Punkt):

Veranschaulichen Sie grafisch ein Beispiel, bei dem der k -Means-Algorithmus für zwei unterschiedliche Initialisierungen zwei verschiedene Ergebnisse liefert.

Aufgabe 2 (1/4 Punkt):

Nehmen wir an, wir möchten eine Mischverteilung von univariaten Normalverteilungen mit dem EM-Algorithmus schätzen. Wir setzen die Varianzen aller Cluster immer auf einen festen Wert $\sigma_\varphi^2 = 1$ und alle Apriori-Wahrscheinlichkeiten auf $\theta^\varphi = 1/k$, k ist die Anzahl der Cluster. In einem E-Schritt rechnen wir $P(y_i|\mathbf{x}_i, \Theta_t)$ aus. In einem M-Schritt schätzen wir nun nur noch die Mittelwerte, verwenden dafür aber nicht das Ergebnis des E-Schritts direkt, sondern verwenden den modifizierten Posterior $P'(y_i|\mathbf{x}_i, \Theta_t)$ (siehe Gleichung 1). Erhält man durch diese Vorgehensweise die gleichen Resultate wie ein k -Means-Algorithmus? Warum, warum nicht?

$$P'(y_i|\mathbf{x}_i, \Theta_t) = \begin{cases} 1 & y_i = \underset{y'}{\operatorname{argmax}} P(y'|\mathbf{x}_i, \Theta_t) \\ 0 & \text{sonst} \end{cases} \quad (1)$$

Aufgabe 3 (1/4 Punkt):

Wie könnte man den E-Schritt eines EM-Algorithmus durch einen Gibbs-Sampler ersetzen? Überlegen sie sich welche Variablen der Sampler sampeln würde und welche man während eines Sampling-Laufs über die Daten festhalten müsste? Wie errechnet man dann den Posterior, den man für den M-Schritt benötigt? Welche Vor- oder Nachteile hätte diese Vorgehensweise?

Aufgabe 4 (1/4 Punkt):

Entwickeln sie ein probabilistisches Modell für Emails, die in einem Call-Center ein- und ausgehen. Eingehende Emails enthalten Fragen zu verschiedenen Themen und ausgehende Emails enthalten Antworten zu diesen Fragen. Zeichnen sie ein Abhängigkeitsdiagramm (ähnlich wie auf Folie 40) aus dem alle Variablen und Abhängigkeiten ersichtlich sind.