Sprachtechnologie

8. Übung

Prof. Tobias Scheffer
Uwe Dick

Sommer 2015

Ausgabe am: 15.06.15
Besprechung am: 22.06.15

Aufgabe 1 *LDA*

Auf Folie 37 der LDA-Vorlesung heißt es, dass Samples für β und θ aus den Posteriors $P(\beta_z|W, Z, \eta)$ bzw. $P(\theta_d|Z, \alpha)$ gezogen werden können. Welche parametrische Formen haben diese Verteilungen und können Sie eine einfache Formel aufstellen, mit der das Samplen effizient implementiert werden kann?

Aufgabe 2 *Themenähnlichkeit*

Bei der Inferenz eines Themenmodells mit 4 Themen und 3 Dokumenten haben sich durch den Gibbs-Sampling-Algorithmus die folgenden 3 Samples für die Themenmischungen der 3 Dokumente ergeben:

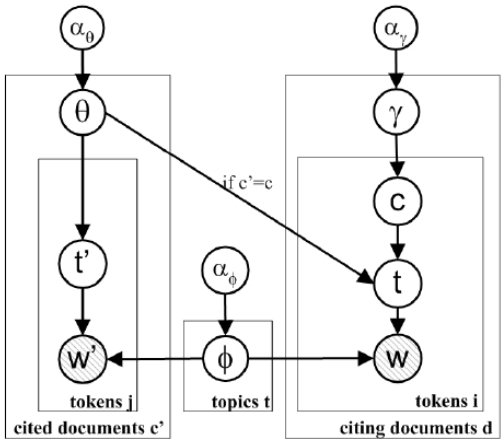
- Sample 1: $(\theta_1 = (0.8, 0.0, 0.1, 0.1), \theta_2 = (0.4, 0.4, 0.2, 0.0), \theta_3 = (0.0, 0.0, 0.3, 0.7))$
- Sample 2: $(\theta_1 = (0.6, 0.2, 0.1, 0.1), \theta_2 = (0.3, 0.5, 0.1, 0.1), \theta_3 = (0.1, 0.1, 0.0, 0.8))$
- Sample 3: $(\theta_1 = (0.7, 0.1, 0.0, 0.2), \theta_2 = (0.6, 0.1, 0.0, 0.3), \theta_3 = (0.2, 0.0, 0.1, 0.7))$

Berechnen Sie für alle Paare von Dokumenten die Ähnlichkeit als inneres Produkt der Themenverteilungen, wobei Sie über die Samples mitteln (Folie 30 der Vorlesung).

Aufgabe 3 *Graphische Modelle*

In der Vorlesung wurde gezeigt, wie aus einem graphischen Modell, die gemeinsame Verteilung $P(z_1, \ldots, z_n)$ aller Zufallsvariablen z_i abgelesen werden kann. Bestimmen Sie für das folgende graphische Modell die gemeinsame Verteilung der Zufallsvariablen!

Hinweis:Die Variablen w' und w sind in diesen Modell die beobachteten Variablen.



b) Copycat Model

Zusatzaufgabe *Graphische Modelle*

Skizzieren Sie das generative Modell 'Mixture of Gaussians' als graphisches Modell und schreiben Sie die gemeinsame Verteilung aller versteckten und beobachteten Variablen auf. Der generative Prozess bei 'Mixture of Gaussians funktioniert so:

- Die Beobachtungen sind Punkte in einem d-dimensionalen Raum.
- Jede Beobachtung stammt aus genau einem von k Clustern.
- Jeder Cluster ist beschrieben durch einen Mittelpunkt.
- Jede Beobachtung wird aus einer multivariaten Normalverteilung mit den Parametern Mittelpunkt des jeweiligen Clusters und Einheitskovarianzmatrix I gezogen.
- Die Clusterzuordnungen werden aus einer Multinomialverteilung gezogen, deren Parameter aus einer Dirichletverteilung mit festen Parametern α gezogen werden.
- die Mittelpunkte der Cluster werden aus einer multivariaten Normalverteilung mit Mittelpunkt 0 und Einheitskovarianzmatrix I gezogen.