Karlsruher Institut für Technologie

Probabilistische Topic Models und Variational Inference: Eine Herleitung der Latent Dirichlet Allocation

Bachelorarbeit von Benedikt Anselment – Wintersemester 2018 / 2019

Lehrstuhl für Wirtschaftspolitik – Prof. Dr. Ingrid Ott

Probabilistische Topic Models allgemein

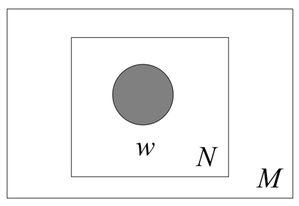
- **Automatische Themenerkennung in Texten mittels Methoden des Unsupervised Machine Learnings**
- **Zentrale Begriffe:**
  Korpus, Dokument, Wort, Vokabular, Themen
- **Modellannahme:**
  Dokumente entstehen in einem *generierenden Prozess*

The New York Times

music band songs rock album jazz pop song singer night	book life novel story books man stories love children family	art museum show exhibition artist artists paintings painting century works	game knicks nets points team season play games night coach	show film television movie series says life man character know

Unigram-Modell

- **„Einfachstes“ Topic Model:**
  Lediglich ein zu inferierendes Thema β im Korpus D



- **Exakte Identifizierung der Maximum-Likelihood-Lösung β^* mittels des Lagrange-Verfahrens möglich:**

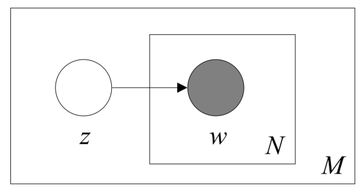
$$\beta_v^* = \frac{c(w_v, D)}{\sum_{v=1}^{V} c(w_v, D)}$$

Expectation Maximization-Algorithmus

- **Iteratives Verfahren zur Bestimmung von Maximum-Likelihood-Punktschätzern in Modellen mit latenten Variablen**
- **Zentrales Element ist sog. *ELBO,* welche eine untere Grenze zum Wert der Log-Likelihood-Funktion des Modells bildet**
- **Im sog. *E-Step* erfolgt eine partielle Maximierung der *ELBO* hinsichtlich $q(Z)$, einer Verteilung über den latenten Variablen Z des Modells**
- **Im sog. *M-Step* wird die *ELBO* bezüglich den Modellparametern Θ partiell maximiert**

Mixture of Unigrams-Modell

- **Im Korpus D gibt es K verschiedene Themen $\beta = (\beta_1, \dots, \beta_K)$, allerdings behandelt jedes Dokument lediglich ein einzelnes Thema $k \in \{1, \dots, K\}$**
- **Im generierenden Prozess eines Dokuments wird das behandelte Thema k aus der Korpusverteilung π gezogen**



- **Aufgrund der gegenseitigen Abhängigkeit von β und π ist nur ML-Schätzung mittels des EM-Verfahrens möglich**

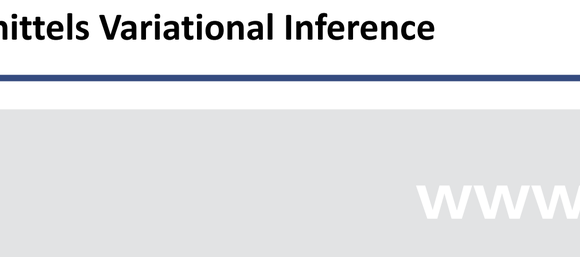
Bayessche Statistik

- **In der frequentistischen Wahrscheinlichkeitslehre werden für unbekannte Größen, wie bspw. die Parameter Θ eines Modells, lediglich Punktschätzungen durch mehrmalige Wiederholung eines Zufallsexperiments inferiert**
- **In der bayesschen Lehre wird die unbekannte Größe mittels einer Wahrscheinlichkeitsverteilung $p(\cdot)$ beschreiben**
- **Inferenz der A-posteriori-Verteilung $p(\Theta|X)$ aus einer A-priori-Verteilung $p(\Theta)$ erfolgt unter Einbezug der Daten X eines Zufallsexperiments mittels des Satzes von Bayes:**

$$p(\Theta|X) = \frac{p(X|\Theta) * p(\Theta)}{p(X)}$$

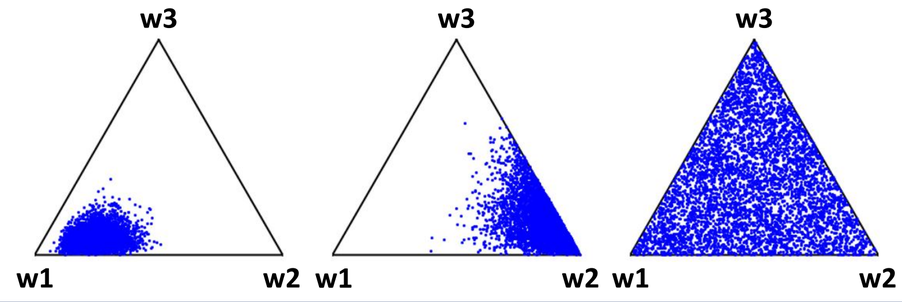
Variational Inference

- **Ermöglicht approximative Bestimmung einer A-posteriori-Verteilung, falls exakte analytische Berechnung von $p(\Theta|X)$ nicht möglich**
- **Umsetzbar mittels Restriktion der zulässigen Verteilungsarten für die A-posteriori-Verteilung**



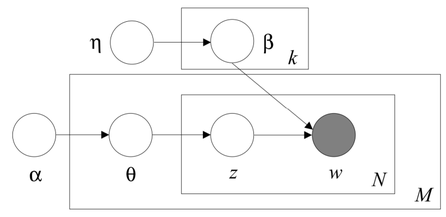
Bayessches Mixture of Unigrams-Modell

- **β und π werden nicht mehr als Modellparameter mit fixem Wert, sondern als bayessche Variablen betrachtet und somit durch eine Dirichlet-A-priori-Verteilung beschrieben**
- **Approximative Bestimmung der Dirichlet-A-posteriori-Verteilung erfolgt mittels Variational Inference**



Latent Dirichlet Allocation

- **Die Dokumente im Korpus D können mehrere Themen $k \in \{1, \dots, K\}$ behandeln**
- **Im generierenden Prozess werden die behandelten Themen für jedes Wort eines Dokuments d_i aus individuellen Dokumentenverteilungen θ_i gezogen**



- **Inferenz erfolgt mittels Variational Inference**