LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Ivana Daskalovska

Willkommen zur Übung Einführung in die Computerlinguistik

Textklassifikation und Naive Bayes

Wiederholung

- **Was versteht man unter Textkllassifikation?**

- **Was versteht man unter Text Klassifikation?**

  - Gegeben:

    - Dokumenten Raum: meist multidimensional
    - Feste Anzahl an Klassen C = {c1,….cn}
    - Training Set D mit annotierten Dokumenten
      $\langle d,c \rangle \in X \times C$

  → Mit Hilfe von Lernalgorithmen lerne einen Klassifikator γ welcher die Dokumenten zu den Klassen zuweist:

$$\gamma : X \rightarrow C$$

- **Bei welchen Problemen kann man eine Textklassifikation anwenden?**

- **Bei welchen Problemen kann man eine Textklassifikation anwenden?**

- Sprach Identifizierung
- Automatische Zuordnung von Emails
- Sentiment Analyse
- …

Welche Klassifikationsmethoden existieren? Beschreiben Sie diese kurz

Welche Klassifikationsmethoden existieren? Beschreiben Sie diese kurz

- **Manuelle**
  - Sehr gute Ergebnisse, wenn Experten die Daten annotieren (Yahoo)
  - Gut geeignet falls kleine Datenmenge und wenig Teamglieder gebraucht
  - Problem: je größer die Datenmenge, desto teurer und komplizierter

Welche Klassifikationsmethoden existieren? Beschreiben Sie diese kurz

- **Regel-Basiert**
  - IDE – Integrated Development Environments werden verwendet um effizient Regel zu schreiben (auch sehr komplizierte Regel)
  - Problem: Entwicklung von regelbasierte Systemen ist aufwendig und teuer

Welche Klassifikationsmethoden existieren? Beschreiben Sie diese kurz

- **Statistische/Probabilistische Methoden**
  - Die Textklassifikation wird als Lernproblem betrachtet
    - Überwachtes Lernen der Klassifikationsfunktion γ
    - Anwenden der gelernte Klassifikationsfunktion auf neue Daten
  - Problem: Trainingsdaten müssen trotzdem per Hand annotiert werden, aber keine Experten nötig

- **Naive Bayes Klassifikator**

- **Naive Bayes Klassifikator**
  - Probabilistischer Klassifikator

$$P(c|d) \propto P(c) \prod_{1 \leq k \leq n_d} P(t_k|c)$$

    - d – Dokument
    - c – Klasse
    - tk – Term in dem Dokument

  - Ziel: Finde die beste Klasse

$$c_{\text{map}} = \text{argmax}_{c \in \mathbb{C}} \hat{P}(c|d) = \text{argmax}_{c \in \mathbb{C}} \hat{P}(c) \prod_{1 \leq k \leq n_d} \hat{P}(t_k|c)$$

- **Naive Bayes Klassifikator**

  - Problem: Multiplizieren von sehr kleinen Wahrscheinlichkeiten → Floating Point Underflow

  - Lösung: Log – Funktion anwenden

    - log(xy) = log(x) +log(y)

$$c_{\text{map}} = \text{argmax}_{c \in \mathbb{C}} \left[\log \hat{P}(c) + \sum_{1 \leq k \leq n_d} \log \hat{P}(t_k|c) \right]$$

- **Wie schätzt man die Parameter?**

- **Wie schätzt man die Parameter?**

  - Maximum Likelihood Estimation
    - Apriori Wahrscheinlichkeit :

$$\hat{P}(c) = \frac{N_c}{N}$$

    - Bedingte Wahrscheinlichkeit:

$$\hat{P}(t|c) = \frac{T_{ct}}{\sum_{t' \in V} T_{ct'}}$$

- **Was ist hier das Problem?**

- **Was ist hier das Problem?**
  - **Problem:** Es kann passieren, dass 1 Wort in den Testdaten vorkommt, aber nicht in den Trainingsdaten. Dadurch wäre seine Wahrscheinlichkeit gleich 0 und somit auch die Wahrscheinlichkeit von dem Dokument für die Klasse wird, durch Multiplizieren mit 0, 0 sein.
  - **Lösung: Addiere-1-Glättung**

$$\hat{P}(t|c) = \frac{T_{ct} + 1}{\sum_{t' \in V}(T_{ct'} + 1)} = \frac{T_{ct} + 1}{(\sum_{t' \in V} T_{ct'}) + B}$$