Sprachtechnologie

11. Übung

Prof. Tobias Scheffer
Uwe Dick

Sommer 2015

Ausgabe am: 06.07.15
Besprechung am: 13.07.15

Aufgabe 1 *Multiklassen-SVM*

Die Multiklassen-SVM kann durch folgendes Optimierungsproblem darstellt werden:

$$\min_{\mathbf{w}} \|\mathbf{w}\|^2 + C \sum_{i=1}^{n} \xi_i$$

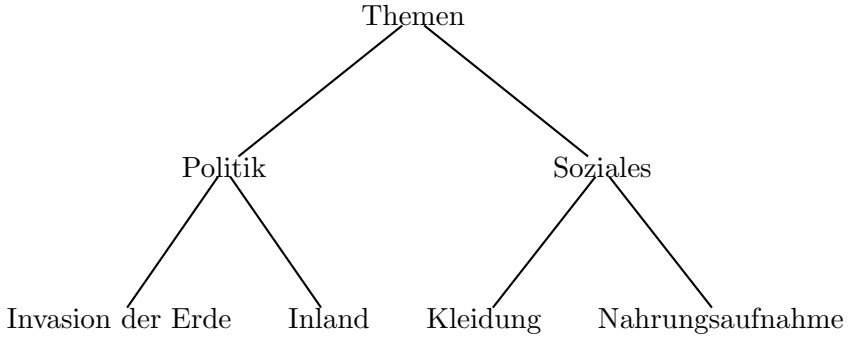
unter den Nebenbedingungen:

$$\forall i \in \{1, \ldots, n\} \forall y \neq y_i : \langle \mathbf{w}, \Phi(\mathbf{x}_i, y_i) \rangle \geq \langle \mathbf{w}, \Phi(\mathbf{x}_i, y) \rangle + 1 - \xi_i$$
$$\forall i \in \{1, \ldots, n\} : \xi_i \geq 0.$$

Zeigen Sie, dass die (binäre) SVM ein Spezialfall dieser Formulierung ist. Vergleichen Sie die optimalen Gewichtsvektoren beider Lösungen.

Aufgabe 2 *Klassifikation mit Taxonomien*

Erich von Däniken stellte fest, dass die Texte der Marsianer in die folgende Hierarchie eingeordnet werden können:



Angenommen Sie wollen für das Korpus des Übungsblatts 10 (siehe Aufgabe 3) einen Klassifikator lernen, der die Texte in diese Baumstruktur einordnet. Wie sehen die gemeinsamen Repräsentationen von Ein- und Ausgabe für diese Beispiele aus? Welche Abschnitte des gelernten Gewichtsvektor kann man unterscheiden? Diskutieren Sie Möglichkeiten, den Lernalgorithmus so anzupassen, dass er einen Klassifikator bevorzugt, der möglichst lange richtige Pfade vorhersagt, d.h. für das erste Beispiel „Kleidung“ vorherzusagen, soll weniger bestraft werden als „Invasion der Erde“.

Aufgabe 3 *PCA*

Sie haben die folgende Datenmatrix $\mathbf{X}$ gegeben, d.h. der Datensatz besteht aus 5 Instanzen mit jeweils 3 Features.

$$\mathbf{X} = \begin{pmatrix} 1 & 0,2 & 0,5 & -0,3 & 0 \\ 0,4 & 0,5 & 0 & 0,1 & -0,2 \\ 0 & 0,1 & -0,6 & 0,4 & -0,6 \end{pmatrix}$$

Um die Dimensionalität der Daten zu reduzieren, möchten Sie eine PCA durchführen. Welche Forderungen werden dabei an $\mathbf{X}$ gestellt? (Siehe Folien Seite 11) Erfüllen Sie die erste Forderung durch die beschriebene Transformation.
Berechnen sie anschließend die Kovarianzmatrix $\mathbf{C}$.
Mithilfe eines Algorithmus zur Eigenvektorzerlegung haben sie die folgenden Eigenvektoren $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3$ von $\mathbf{C}$ mit den entsprechenden Eigenwerten $\lambda_1 = 0,42$, $\lambda_2 = 0,02$, $\lambda_3 = 0,23$ berechnet.

$$\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3) = \begin{pmatrix} 0,887 & 0,425 & 0,178 \\ 0,275 & -0,798 & 0,536 \\ -0,370 & 0,426 & 0,825 \end{pmatrix}$$

Bestimmen Sie nun die ersten beiden Hauptkomponenten und transformieren Sie anschließend die Daten in den neuen Raum der durch die beiden Hauptkomponenten aufgespannt wird.