Sprachtechnologie

5. Übung

Prof. Tobias Scheffer
Uwe Dick

Sommer 2016

Ausgabe am: 24.05.16
Besprechung am: 30.05.16

Aufgabe 1 *PCFG*

Gegeben sei der folgende Auszug einer probabilistischen kontextfreien Grammatik mit dem Startsymbol S:

Nicht-Terminale				P	Terminale			P
S	→	P	VP	0.8	P	←	ich	0.3
S	→	NP	AK	0.2	Q	←	im	0.2
AK	→	PP	N	0.6	N	←	Raketenauto	0.8
QP	→	Q	N	0.75	V	←	fahre	0.4
PP	→	V	Q	0.5				
VP	→	V	QP	0.4				
NP	→	P		0.3				

Visualisieren Sie alle Parsebäume, die den Satz „*Ich fahre im Raketenauto*“ generieren und berechnen Sie die Wahrscheinlichkeit, dass die obige Grammatik diesen Satz generiert.

Aufgabe 2 *Neuronale Netze*

Nehmen wir an, ein Verwandter arbeitet bei einer Bank und Sie möchten ihm helfen, die Kreditkrise durchzustehen. Er bittet Sie, eine Modell zu entwickeln und zu lernen. Mit diesem Modell soll vorhergesagt werden, ob zukünftige Kreditantragssteller ihren Kredit zurückzahlen werden oder nicht. Die Trainingsdaten und die Testdaten finden Sie auf der Vorlesungswebseite neben dem Übungsblatt im weka-Format. Installieren Sie das weka Toolkit (http://www.cs.waikato.ac.nz/ml/weka) und machen Sie sich mit dem Programm vertraut (Hinweis: Nutzen Sie den Explorer-Modus). Für diese Aufgabe benötigen Sie zwei verschiedene Modelle: Logistic (logistische Regression) und MultilayerPerceptron (einfaches Neuronales Netz mit Perceptron Aktivierungsfunktion). Sie finden diese in weka unter weka → classifiers → functions im Classify-Tab des Tools.

1. Laden Sie die Daten in Weka, um ein Modell trainieren zu können.
2. Trainieren Sie ein lineares Modell mit Hilfe der logistischen Regression und evaluieren Sie das Modell auf den Testdaten.
3. Trainieren Sie ein Neuronales Netz mit 2 Hidden Layers und jeweils 10 Hidden Units. Evaluieren Sie dieses Modell auf den Testdaten.
4. Welches Modell hat eine bessere Performance? Für welches Modell würden Sie sich entscheiden?

Aufgabe 3 *Optimales Modell*

Nutzen Sie das weka Toolkit, um ein optimales Modell zu bestimmen indem Sie optimale Parameter (Anzahl Hidden Layer und Hidden Units) für ein MulilayerPerceptron bestimmen (Hinweis: Nutzen Sie die Daten aus Aufgabe 2).

1. Überlegen Sie sich einen einfachen Algorithmus, um ein optimales Modell/Parameter zu bestimmen. Was bedeutet in diesem Kontext optimal?

2. Bestimmen Sie das optimale Modell.

3. Welche Zusammenhänge zwischen Anzahl der Hidden Layer und Anzahl der Hidden Units sind Ihnen aufgefallen? Wie können diese erklärt werden?

Zusatzaufgabe *Conditional Random Fields*

Auf der Buchmesse in Zofingen stellt Erich von Däniken einen Sprachführer über die Mehrdeutigkeiten der marsianischen Wörter *bob, zonk* und *argh* vor. Laut Sprachführer können die Wörter *bob* und *zonk* sowohl als Nomen als auch als Verb und das Wort *argh* als Verb sowie als Adjektiv getaggt werden. Als Sie an von Dänikens Buchstand vorbeigehen, hören Sie, wie sich seine marsianischen Assistenten miteinander unterhalten und dabei die Phrase *bob zonk argh* fällt. Um nun herauszufinden, über was die Marsianer aller Wahrscheinlichkeit nach gesprochen haben, versuchen Sie die wahrscheinlichste POS-Tag-Folge gegeben die gehörte Sequenz mithilfe eines Conditional Random Fields zu berechnen. Skizzieren Sie das hierfür benötigte Conditional Random Field und zeigen Sie dabei die wesentlichsten Unterschiede zu HMMs. Wo liegen die Vorteile von CRFs?