Computational Learning Theory

Sommersemester 2017

Mario Holldack, M. Sc.
Prof. Dr. Georg Schnitger
Hannes Seiwert, M. Sc.

GOETHE
UNIVERSITÄT
FRANKFURT AM MAIN

Institut für Informatik
AG Theoretische Informatik

Übungsblatt 9

Ausgabe: 19.06.2017
Abgabe: 26.06.2017

Für dieses Übungsblatt benötigen Sie die Python-Bibliothek scikit-learn (und ihre Abhängigkeiten). Empfehlenswert bei der Entwicklung ist die Benutzung der Notebook-Funktion von jupyter.

- http://scikit-learn.org/stable/install.html
- https://jupyter.org/install.html

Reichen Sie Ihren kommentierten Quellcode per E-Mail als *.py- oder *.ipynb-Datei ein: holldack@thi.cs.uni-frankfurt.de

Aufgabe 9.1 *Spamerkennung mit SVMs und scikit-learn* (4 + 12 + 12 + 4 = 32 Punkte)

In dieser Aufgabe trainieren wir einen Spamfilter mithilfe von SVMs. Laden Sie die klassifizierten E-Mails aus dem bereitgestellten Spam-E-Mail-Datensatz herunter und lesen Sie die Datei readme.txt:

http://thi.cs.uni-frankfurt.de/lehre/clt/sose17/clt_sose17_blatt09_daten.zip

a) Lesen Sie alle E-Mails im Ordner MAIL_DATA als Liste von Strings ein und wählen Sie eine zufällige Stichprobe von 4000 E-Mails als *Trainingsmenge* aus. Die übrigen E-Mails dienen später als *Testmenge* zur Validierung des Lernerfolgs.

Erstellen Sie mithilfe von scikit-learn für jede E-Mail aus der Trainingsmenge einen Featurevektor wie in Beispiel 11.1 im Skript (die Trainingsmenge entspricht der Menge D aller Dokumente). Sie können die Klasse TfidfVectorizer mit Defaultoptionen verwenden[1]:

http://scikit-learn.org/stable/modules/feature_extraction.html#text-feature-extraction

Geben Sie die Dimension der Feature-Vektoren an.

b) Trainieren Sie eine SVM mithilfe der Klasse SVC (siehe http://scikit-learn.org/stable/modules/svm.html) mit

i) linearen Kernen $K(x, y) = \langle x, y\rangle$
ii) polynomiellen Kernen $K(x, y) = (\langle x, y\rangle + c)^d \cdot d$ für $d \in \{2, 3\}$ und
iii) Gauß-Kernen $K(x, y) = \exp(-\frac{\|x-y\|^2}{\sigma^2})$

für eine geeignete Auswahl von Parametern c und σ. Eine systematische Parameterwahl mittels Kreuzvalidierung ist in dieser Übungsaufgabe nicht nötig. Bewerten Sie jeweils den Trainingserfolg in Abhängigkeit von der Anzahl der Supportvektoren und im Fall der linearen Kerne auch in Abhängigkeit von der Norm $\|w\|$ des Gewichtsvektors (bei linearer Trennbarkeit ist der Margin $1/\|w\|$).

Bitte wenden!

[1] Der in Beispiel 11.1 vorgestellte Bag-of-Words-Ansatz wird auch als Tfidf-Ansatz (Termfrequenz/inverse Dokumentenfrequenz) bezeichnet. Um Divisionen durch Null zu vermeiden, verwendet TfidfVectorizer standardmäßig den „Überraschungseffekt“, d.h. die inverse Dokumentenfrequenz, $\log\left(\frac{|D|+1}{|D(s)|+1}\right) + 1$ statt $\log\left(\frac{|D|}{|D(s)|}\right)$.

c) Klassifizieren Sie die E-Mails aus der Testmenge mit der erfolgreichsten SVM bzgl. der Anzahl Supportvektoren. Vergleichen Sie Ihre Ergebnisse mit den gegebenen Klassifikationen aus den Dateinamen.

   i) Wie hoch ist der Anteil der positiven bzw. negativen Fehlklassifikationen? Wie hoch ist der Anteil der Fehlklassifikationen (d.h. der Testfehler) insgesamt? Würden Sie dieser Support-Vektor-Maschine als Spamfilter vertrauen?

   ii) Wie gut schneidet diese SVM ab, wenn Sie die Trainingsmenge auf 3000, 2000, 1000, 500 oder 100 zufällig gewählte E-Mails reduzieren?

   iii) Die Klasse SVC verfügt über einen Strafparameter $C > 0$, mit dessen Hilfe Sie steuern können, wie sehr Fehlklassifikationen in der Trainingsphase bestraft werden. Je größer C ist, desto höher ist die Strafe.

   Wie verhält sich der Testfehler in Abhängigkeit von C? Wie verändert sich $\|w\|$ im Fall linearer Kerne? Begründen Sie Ihre Antwort qualitativ und quantitativ.

d) Kritische Stimmen behaupten, dass manche Medien eine einseitige politische Agenda verfolgen und dass es nicht möglich sei, Aussagen von Politikern und Aussagen wohlgesinnter Medien zu unterscheiden.

   Können Sie den oben implementierten Spamfilter nutzen, um erfolgreich zu lernen, ob es sich bei einer Twitter-Nachrichten aus dem Ordner TWITTER_DATA um eine Aussage eines Politikers oder um eine Aussage eines TV-Senders handelt?