Technische Universität Dortmund
Fakultät Statistik
Lehrstuhl Computergestützte Statistik
Fakultät Informatik
Lehrstuhl für Künstliche Intelligenz

Prof. Dr. Claus Weihs
Dr. Thomas Liebig
Sebastian Buschjäger
Malte Jastrow
Jennifer Neuhaus-Stern

Übung zur Vorlesung

Wissensentdeckung in Datenbanken

Sommersemester 2018

Übungsblatt Nr. 11

Der Abgabetermin ist Dienstag der 17.07.2018 bis 10:00 Uhr im moodle-Raum

Aufgabe 1 (Feature-Selection) **(7 Punkte)**

In dieser Aufgabe sollen Sie den `housing.csv` Datensatz verwenden, um Techniken zur Feature-Selektion praktisch anzuwenden.
Das R-Paket `FSelector` bietet Methoden zur Feature-Selektion, wobei wir uns hier auf die Funktionen `forward.search` für Vorwärtsselektion und `backward.search` für Rückwärtsselektion beschränken. Des Weiteren wollen wir das LASSO Verfahren zur Feature-Selektion benutzen, welches im R-Paket `glmnet` verfügbar ist. Eine ausführliche Einführung finden Sie unter `https://web.stanford.edu/~hastie/glmnet/glmnet_alpha.html`.

a) (1 Punkt) Laden Sie den Datensatz und fitten Sie eine lineares Regressionsmodell mit Hilfe der `lm` Funktion auf **allen** Features. Wie hoch ist der MSE bei einer 5-fachen Kreuzvalidierung?

b) (1 Punkt) Nutzen Sie die `forward.search` Funktion um eine Vorwärtsselektion auf den `housing.csv` Daten durchzuführen. Legen Sie erneut ein lineares Regressionsmodell zugrunde und benutzen Sie eine 5-fache Kreuzvalidierung um die aktuelle Featureauswahl zu bewerten. Welche Features werden ausgewählt und wie hoch ist der MSE?

c) (1 Punkt) Bei der Vorwärtsselektion fügt man ein Feature nach dem Anderen zur aktuellen Featuremenge hinzu. Bei der Rückwärsselektion geht man umgekehrt vor indem man ein Feature nach dem Anderen aus der Featuremenge herauslöscht. Nutzen Sie die `backward.search` Funktion um eine Rückwärtsselektion auf den `housing.csv` Daten durchzuführen. Legen Sie erneut ein lineares Regressionsmodell zugrunde und benutzen Sie eine 5-fache Kreuzvalidierung um die aktuelle Featureauswahl zu bewerten. Welche Features werden ausgewählt und wie hoch ist der MSE?

d) (2 Punkt) Nutzen Sie die `cv.glmnet` Funktion des `glmnet` Pakets um LASSO auf den `housing.csv` Daten mit mindestens drei verschiedenen λ-Werten durchzuführen. Benutzen Sie eine 5-fache Kreuzvalidierung um die aktuelle Featureauswahl zu bewerten. Welche Features werden ausgewählt und wie hoch ist der MSE?

e) (1 Punkt) Wiederholen Sie die Aufgabenteile a) bis c) einige male. Was fällt Ihnen auf? Wie erklären Sie diese Auffälligkeit?

f) (1 Punkt) Evaluieren Sie die drei Methoden kritisch: Was sind die Vor- und Nachteile der Greedy-Selection (Forward/Backward) gegenüber LASSO?

Aufgabe 2 (Online-Algorithmen & Wiederholung) **(5 Punkte)**

Sie haben nun in der Vorlesung schon einige Algorithmen kennen gelernt. Welche der folgenden Algorithmen klassifizieren Sie als Online-Algorithmen, d.h. als solche Algorithmen die ein Beispiel nach dem anderen konsumieren? Wie groß ist der Speicherbedarf und die Laufzeit während des Trainings bzw. dem Update für die folgenden Algorithmen in Abhängigkeit von der Anzahl der Samples

Algorithmus	Online	Laufzeit	Speicher	Aufgabe
Naive Bayes	nein	$\mathcal{O}(N)$	$\mathcal{O}(K)$	Klassifikation
Random Forest				
Logistische Regression				
Neuronales Netzwerk				
K-MEANS				
FP-Growth				

N, der Anzahl der Klassen K und der Dimensionalität der Eingabedaten d? Nutzen Sie bitte $\mathcal{O}(1)$ falls ein Algorithmus einen konstanten Speicher- bzw. Zeitbedarf hat. Für welche Aufgaben sind die einzelne Algorithmen einsetzbar - Klassifikation, Regression, Clustering oder Häufige-Mengen? Hier sind Mehrfachnennungen möglich. Geben Sie im Zweifelsfall eine Begründung an, wieso sie eine bestimmte Auswahl getroffen haben. Es gibt 0.25P pro korrektem Eintrag.
Hinweis: Die erste Zeile in der Tabelle dient als Beispiel.

Aufgabe 3 (Reservoir-Sampling & Lossy-Counting) (8 Punkte)

Sie haben in der Vorlesung bereits Reservoir-Sampling und Lossy Counting kennen gelernt. In dieser Aufgabe sollen Sie praktisch mit beiden Verfahren arbeiten. Im Moodle finden Sie die Datei `words.txt`, welche die ersten 100000 Wörter der Wikipedia enthält. Diese sollen sie auf das Vorkommen einzelner Wörter untersuchen.
Hinweis: Wenn sie eine Hash-Map bzw. Dictionary Datenstruktur benötigen können Sie R-`environment` benutzen.

a) (2 Punkte) Erstellen Sie ein Histogramm der Worthäufigkeiten, d.h. zählen Sie das Vorkommen jedes Wortes in der Datei. Hierzu finden Sie in der Template-Datei entsprechenden Beispielcode, welcher die Datei `words.txt` zeilenweise, d.h. Wort-für-Wort liest. Erweitern Sie diesen Code, sodass sie ihr Histogramm mit jeder neuen Zeile aktualisieren. Messen Sie die Laufzeit und geben sie die häufigsten 20 Wörter aus.

b) (2 Punkte) Ziehen Sie nun ein Sample der Größe $W = 5000$ aus der Datei mit Hilfe des Reservoir-Sampling Algorithmus. Erweitern Sie erneut den entsprechenden Beispielcode um ihre Implementierung von Reservoir-Sampling und berechnen Sie anschließend ein Histogramm auf Grundlage ihres Samples. Messen Sie die Laufzeit und geben sie die häufigsten 20 Wörter aus.

c) (1 Punkt) Wie bewerten Sie das Vorgehen aus a) und b) im Bezug auf die Laufzeit und Güte? Geben Sie eine Empfehlung ab: Wann lohnt sich das Vorgehen aus a), wann aus b)?

e) (3 Punkte) Angenommen Sie sind nur an den 20 häufigsten Wörtern interessiert. Implementieren Sie den Lossy-Counting Algorithmus um die 20 häufigsten Wörter aus der `words.txt` Datei zu extrahieren. Wie bewerten Sie ihre Auswahl im Hinblick auf die Histogramme aus a) und b)?