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. 4

Der Abgabetermin ist Dienstag der 29.05.2018 bis 10:00 Uhr im moodle-Raum

Aufgabe 1 (Naive Bayes) (5 Punkte)

Mit dieser Aufgabe soll Ihnen die Verwendung des R-Pakets mlr nahegelgt werden. Unter `https://mlr-org.github.io/mlr-tutorial/release/html/` finden Sie ein ausführliches Tutorial. mlr stellt vereinheitlichte Schnittstellen für viele verschiedene maschinelle Lernverfahren zur Verfügung so wie viele Funktionen zur weiterführenden Analyse. Diese Aufgabe behandelt die absoluten Grundlagen:

a) (1 Punkt) Erstellen Sie einen task, um die Ionosphere-Daten aus dem R-Paket mlbench zu bearbeiten. Verwenden Sie dabei nur 80% der Daten.

b) (2 Punkte) Erstellen Sie einen naiveBayes-Lerner und trainieren Sie ihn auf dem task.

c) (2 Punkte) Sagen Sie die Klassen auf den verbliebenen 20% der Daten vorher. Vergleichen Sie die vorhergesagten mit den wahren Klassen. Ist das Ergebnis gut?

Aufgabe 2 (Kanonische LDA) (5 Punkte)

In dieser Aufgabe geht es darum die kanonische Variante der Linearen Diskriminanzanalyse selbst zu implementieren.

a) (2 Punkte) Implementieren Sie eine Funktion train.mylda, die ein kanonisches LDA-Modell an einen Datensatz anpasst. Die Eingabeparameter sollen ein Dataframe data und eine Zeichenkette target sein, die die Zielvariable angibt. Der Rückgabewert soll eine benannte Liste sein, die alle relevanten Informationen zum Modell beinhaltet.

b) (2 Punkte) Implementieren Sie eine Funktion predict.mylda, die mittels eines kanonischen LDA-Modells Beobachtungen eines neuen Datensatzes klassifiziert. Eingabeparameter sind das bereits gelernte model sowie der Dataframe newdata.

c) (1 Punkt) Wenden Sie Ihre Funktionen auf die im Template vorgegebene Aufteilung des Iris-Datensatzes an. Bestimmen Sie die Fehlklassifikationsrate (d.h. den Anteil falscher Vorhersagen) auf den Testdaten.

Aufgabe 3 (LDA, QDA, RDA) (5 Punkte)

Verwenden Sie folgende in R implementierte Varianten der Diskriminanzanalyse:

a) (1.5 Punkte) Wenden Sie die Implemetierung der Fisher-Variante lda aus dem Paket MASS auf den iris-Datensatz an. Vergleichen Sie die Fehlklassifikationsrate mit Ihrer eigenen Implementierung. Hätte man sich diesen Vergleich sparen können?

b) (0.5 Punkte) Vergleichen Sie die Ergebnisse aus a) auch mit der Implementierung der quadratischen Diskriminanzanalyse qda aus dem Paket MASS.

c) (3 Punkte) Passen Sie eine Regularisierte Diskriminanzanalyse (Funktion rda aus dem Paket klaR) an. Dazu müssen die Regularisierungsparameter zunächst optimal eingestellt werden. Verwenden Sie dazu eine Gittersuche: Variieren Sie δ und λ jeweils in $\{0, 0.2, 0.4, 0.6, 0.8, 1\}$. Passen Sie für alle (möglichen) Kombinationen von Parametereinstellungen jeweils ein Modell auf den Trainingsdaten an und berechnen Sie den Testfehler. Welche Parameterkombination führt zur geringsten Fehlklassifikationsrate? Ähnelt das resultierende Verfahren eher einer LDA oder einer QDA? Lohnt sich die Regularisierung auf diesen Daten überhaupt?

Hinweis: Setzen Sie den Parameter crossval auf FALSE, da sonst intern eine Kreuzvalidierung durchgeführt wird. Achten Sie zudem darauf, dass die Parameter bei der Funktion rda anders als in der Notation im Skript nicht δ und λ heißen!

Aufgabe 4 (Entscheidungsgrenzen) **(5 Punkte)**

Vergleichen Sie die Entscheidungsgrenzen von Naive Bayes, LDA, QDA und RDA in mlr.

a) (3 Punkte) Verwenden Sie die Funktion plotLearnerPrediction aus dem Paket mlr. Setzen Sie mögliche Parameter der Verfahren sinnvoll.

Verwenden Sie die folgenden künstlichen Datensätze aus dem Paket mlbench um die Entscheidungsgrenzen der Verfahren zu visualisieren:

- mlbench.2dnormals(500,2)
- mlbench.smiley(500, 0.1, 0.05)
- mlbench.cassini(5000)

b) (2 Punkte) Beschreiben Sie die Unterschiede zwischen den Entscheidungsgrenzen.