4. Aufgabenblatt zur Vorlesung

Musterkennung / Machine Learning S19

Rául Rojas, Tobias Langner, Stephan Sundermann, Fritz Ulbrich

Abgabe: bis Mittwoch, den 31. Juli 2019, 10.00 Uhr

1. **Fisher-Diskriminante**
   Benutzen Sie den im Tutorium erläuterten Spam-Datensatz[1]. Die Daten sind in einer CSV-Datei gegeben. Jede Zeile stellt Merkmale einer E-Mail dar. Die letzte Spalte gibt an, ob die entsprechende E-Mail als Spam angesehen wird. Eine Beschreibung der Merkmale finden Sie im gleichen Ordner. Teilen Sie den Datensatz in 80% Trainingsdaten und 20% Testdaten ein. Beachten Sie, dass Trainings- und Testdaten ein ähnliches Verhältnis von Spam zu Nicht-Spam haben sollten.
   Implementieren Sie einen binären Klassifikator mittels der Fisher-Diskriminante. Kann man anhand der gefundenen Projektion beurteilen welche Merkmale am nützlichsten sind?
   Geben Sie die Klassifikationsgenauigkeit an. Erstellen Sie die Konfusionsmatrix und beschreiben Sie diese kurz. Plotten Sie außerdem die projizierten Dichtefunktionen der Klassen.

2. **Logistische Regression**
   Implementieren Sie logistische Regression. Benutzen Sie Ihre Implementierung, um einen besseren Spam-Klassifikator zu trainieren. Verwenden Sie die gleiche Aufteilung in Trainings- und Testdaten. Normalisieren Sie die Daten vor dem Trainieren.
   Plotten Sie wie sich die Klassifikationsgenauigkeit über die Iterationen verändert. Geben Sie die Konfusionsmatrix aus. Wie könnte man das Modell anpassen, um eine der Klassen zu bevorzugen?

Hinweis: Bitte bearbeiten Sie die Aufgaben in Zweier-Gruppen und laden Sie alle Ergebnisse (Quelltext + Dokument mit Plots, Tabellen und Erläuterungen) auf der Vorlesungsseite im Whiteboard hoch[2]. Geben Sie die Namen beider Übungspartner an. Die Bewertung erfolgt binär (bestanden/nicht bestanden).

[1] `https://archive.ics.uci.edu/ml/datasets/spambase` → Data Folder

[2] `https://kvv.imp.fu-berlin.de/portal/site/30550a4f-be9f-4be0-8ec2-a35a294ab3a7`