Beleg 3 – Aufgabenstellung

Thema: Supervised Learning mit Naive Bayes

Ziel der Belegarbeit ist für eine Menge von Titanic-Passagieren möglichst treffsicher vorauszusagen, ob diese das Unglück überleben oder nicht. Dazu können Sie sich von kaggle den bekannten Titanic Dataset (https://www.kaggle.com/c/titanic/) herunterladen. Dieser enthält von allen Passagieren Informationen über das Geschlecht, Alter, Fahrpreis, Passagierklasse, Einstiegshafen, Kabine und die mitreisende Familie.

Der Titanic Datensatz ist aufgeteilt in einen Trainings- und einen Testteil. Der Trainingsdatensatz enthält die Information über das Überleben des Passagiers und ist dafür da, eine Funktion zu lernen, die eine möglichst gute Voraussage für den Rest der Passagiere (Testset) treffen kann. Die Bewertung der Voraussage erfolgt über den Testdatensatz. Auf ihm wird das erlernte Modell angewendet, d.h. es wird eine Tabelle generiert, die die PassengerID und die Vorhersage enthält.

Eine endgültige Evaluation erfolgt dann über die kaggle-Seite. Nur hier sind die Informationen über das Überleben der Restpassagiere verfügbar. Für die Implementierung gibt es zwei vobereitete Softwareprojekte. Sie enthalten beide Funktionen für das Auslesen der Kaggle-Daten, das Anwenden des erlernten Modells auf den Testdatensatz sowie das Erzeugen der Abgabedatei für die Evaluation der erreichten Ergebnisse.

Die Projekte haben unterschiedliche Zielrichtungen:

1. das übliche sbt-Projekt für die Implementierung des Algorithmus und
2. das Jupyter Notebook für die Datenexploration.

Beide Projektarten können verwendet werden. Eine Installation des Notebooks befindet sich auf den Laborrechnern und kann mit dem Kommando *jupyter notebook* gestartet werden. Wer sich die Umgebung auf dem eigenen Recher installieren will, der muss sich das Jupyter Notebook (z.B. anaconda-Distribution) und den Apache Toree-Scala Kernel oder Almond Kernel (Scala 2.11.12 Version, Almond 0.6.0) installieren.

a) Machen Sie sich mit dem kaggle-Projekt vertraut und erstellen Sie einen Account für die Abgabe.
b) Laden Sie eines oder beide der oben genannten Projekte herunter, entpacken Sie es und machen Sie sich mit den implementierten Funktionen vertraut.
c) Imlementieren Sie die Funktion `countAllMissingValues(passengers: List[Map[String,Any]],attList:List[String]):Map[String,Int]` bei der für eine Liste von Attributen gezählt wird, wie häufig diese im übergebenen Datensatz fehlen (in der Map nicht belegt sind).
d) Machen Sie sich mit dem Vegas-Framework vertraut. Leider gibt es nur wenig Dokumentation – als Startpunkt empfehle ich das folgende Video von der Spark-Summit-Conference (https://www.youtube.com/watch?v=5GfQVsnHj88),die vegalite-Doku (https://vega.github.io/vega-lite-v1/docs/) und die Vegas-Doku (https://github.com/vegas-viz/Vegas).
e) Schauen Sie sich die Beispiele auf der Vegas-Projektseite an und versuchen Sie diese für das Titanic Dataset zu adaptieren. Erstellen Sie weitere Statistiken z.B. die Überlebensrate nach Passagierklasse, Alter und/oder Geschlecht.
f) Implementieren Sie den in der Vorlesung vorgestellten Naive Bayes-Algorithmus. Stellen Sie mit hinreichend Tests sicher, dass dieser die richtigen Ergebnisse liefert. (Für die Implementierung soll kein Framework wie Spark, die die Funktionalität schon bereit stellen, angewendet werden.)

g) Preparieren Sie den Titanic Datensatz in dem Sie sich Überlegen, welche Informationen Sie in welcher Form für die Vorhersage verwenden wollen und wie Sie mit den fehlenden Datenfeldern umgehen wollen. Bereiten Sie entsprechend Ihrer Überlegungen den Datensatz auf.
h) Wandeln Sie die Kontinuierlichen Variablen in Kategorische Variablen um. Wenden Sie hierfür schon bekannte Algorithmen an oder explorieren Sie den Datensatz mit Hilfe der Visualisierungsbibliothek vegas.
i) Rufen Sie Ihren in f implementierten Naive Bayes-Algorithmus mit dem in g und h erzeugten Datensatz auf, erzeugen Sie eine Abgabedatei und ermitteln Sie die Güte Ihres Ergebnisses über die kaggle-Seite.
j) Präsentieren Sie Ihre Ergebnisse in der Übung. Dabei sollen Sie darauf eingehen, welche Attribute Sie wie verwendet haben und wie sie die kontinuierlichen Variablen in Kategorische umgewandelt haben (Begründung der Vorgehensweise). Weiter sollen Sie die Implementierung Ihres Algorithmus zeigen und wie Sie die richtige Arbeitsweise der Funktion sichergestellt haben. Zeigen Sie abschließend, welches Ergebnis Sie erzielt haben
– dabei können auch gerne auf unterschiedliche Varianten eingehen.

Die Projekte sollen in Gruppen von 1-2 Personen umgesetzt werden. Die Präsentation erfolgt in der letzten Übung vor der Klausur am 17.01.2019 an der Tafel. Die Präsentation der Ergebnisse erfolgt in der Gruppe.