Data Science / Statistical Learning im SoSe25

Vorgaben für die Simulationsstudie

Name: Yann Ahlgrim

Nutzen Sie zur Umsetzung Ihrer Simulationsstudie bitte Jupyter Notebook und Python. Die erstellen Notebooks sind in einer lauffähigen Form mit abzugeben.

1. Fachthema: Diebstahlerkennung eines Autos in Echtzeit

Im Auto werden eine Vielzahl von Daten zum Fahrverhalten erfasst. Dazu gehört zum Beispiel der Schaltpunkt in Abhängigkeit von der Drehzahl, Urzeiten für Fahrbeginn, übliche Fahrstrecken sowie Fahrdauern und vieles mehr. Ziel Ihres Modells soll es sein, anhand des Fahrverhaltens eine Wahrscheinlichkeit dafür abzuschätzen, dass ein anderer Fahrer als üblich am Steuer sitzt. Bei einer sehr hohen Wahrscheinlichkeit soll eine Diebstahl-Warnmeldung herausgegeben werden.

Bitte beachten Sie, dass es sich bei den aufgezählten Variablen nur um Beispiele handelt. Bitte definieren Sie diese selbst passend zu Ihrem Thema.

2. Erzeugung einer Grundgesamtheit

Erzeugen Sie eine Grundgesamtheit mit N = 1.000.000 Elementen, die die Variablen zum fachlich gewählten Thema enthält. Neben der abhängigen (zu erklärenden) Variablen (im Folgenden auch Zielvariable genannt) soll die Grundgesamtheit 8 mit Zufallszahlen simulierte potenziell erklärende Variablen enthalten.

Die Erzeugung der 8 potenziell erklärenden Variablen soll über Zufallszahlen jeweils mittels aufgrund des Fachthemas geeigneter statistischer Verteilungen erfolgen. Dabei soll(en) 2 inhaltlich sinnvolle Abhängigkeit(en) zwischen jeweils 2 dieser 8 Variablen näherungsweise abgebildet werden. Zur Vermeidung von Multikollinearitäts-Problemen sollten die Abhängigkeiten jedoch so modelliert werden, dass im optimalen Modell (siehe Aufgabenteil 3.) ein VIF <5 resultiert. Positiv wird beurteilt, wenn Sie empirische Daten zur Modellierung der Verteilungen bzw. zur Bestimmung von Verteilungsparametern heranziehen. Wo dies nicht möglich oder nicht sinnvoll ist, können Sie mit geeigneter Begründung die Verteilungen aber auch gerne aufgrund inhaltlicher Überlegungen wählen.

Die Zielvariable soll von 4 der 8 Variablen funktional abhängig sein, d.h. diese 4 sind die erklärenden Variablen. Von diesen erklärenden Variablen sollen 2 Variable(n) nominal oder ordinal skaliert sein mit mindestens 3 möglichen Ausprägungen. Alle weiteren erklärenden Variablen sollen metrisch skaliert sein. Bezüglich der Variablen die keinen Erklärungsgehalt auf die zu erklärende Variable liefern, sind Sie frei in der Wahl der Skalierung.

Modellieren Sie die Abhängigkeiten zwischen der Zielvariablen und den erklärenden Variablen so, dass im Falle eines Klassifikationsproblems die Annahmen des logistischen Regressionsmodells und im Falle eines Regressionsproblems die Annahmen des linearen Regressionsmodells erfüllt sind. Erzeugen Sie dabei einen (im Falle des Klassifikationsproblems impliziten) Störterm (Rauschen).

Bitte beschreiben und begründen Sie in Ihrem Ergebnisbericht Ihr Vorgehen zur Erzeugung der Grundgesamtheit. Gehen Sie dabei insbesondere im Detail auf die verwendeten Verteilungen, die Datengrundlage, die modellierten Abhängigkeiten, die verwendeten Beta-Werte und die Verteilung

der Zielvariablen sowie der dieser zugrunde liegende Verteilung der p_i-Werte ein. Arbeiten Sie die wesentlichen Ergebnisse (beispielweise mittels graphischer Darstellungen) heraus.

3. Simulation der Perspektive des Data Scientist (m/w/d)

Einem Data Scientist (m/w/d) steht eine Stichprobe der Grundgesamtheit im Umfang von n = 20.000 Elementen zur Verfügung. Gehen Sie davon aus, dass der Data Scientist (m/w/d) als Modellklasse bei einem Klassifikationsproblem die logistische Regression und bei einem Regressionsproblem die lineare Regression einsetzt.

Der Data Scientist (m/w/d) möchte mit Methoden des Statistical Learning ein Modell entwickeln, mit dem er die Zielvariablen (Regressionsproblem) bzw. die Wahrscheinlichkeitsverteilung der Zielvariablen (Klassifikationsproblem) möglichst gut aus den (potenziell) erklärenden Variablen schätzen kann. Die Anwenderperspektive impliziert dabei, dass der Data Scientist (m/w/d) keine Kenntnisse über die Modellierung der Grundgesamtheit besitzt.

Erläutern Sie anhand der Daten das systematische schrittweise Vorgehen des Data Scientist (m/w/d), stellen Sie die Ergebnisse der einzelnen Schritte dar erläutern Sie diese ausführlich. Verlassen Sie nach dem abschließenden Schritt die Anwenderperspektive und zeigen durch detaillierten Modellvergleich, dass durch das systematische Vorgehen am Ende bis auf Zufallsschwankungen das Modell resultiert, das zur Erzeugung der Zielvariablen in der Grundgesamtheit verwendet wurde.

4. Güte der Modellparameter

Ziehen Sie aus der Grundgesamtheit jeweils k = 1.000 zufällige Stichproben im Umfang von n zwischen 1.000 und 50.000 Elementen. Wählen Sie die Schrittweite von n zwischen 1.000 und 50.000 Elementen dabei so, dass die Analysen im Hinblick auf die folgende Aufgabenbeschreibung aussagefähig sind. Trainieren Sie auf jeder dieser Stichproben das optimale Modell aus Abschnitt 3. Wählen Sie für die folgenden Analysen den Beta-Wert einer der erklärenden Variablen aus.

Vergleichen Sie die simulierten statistischen Verteilungen des geschätzten Beta-Werts beispielhaft für drei von Ihnen geeignet gewählte Stichprobenumfänge graphisch miteinander. Welche grundsätzlichen Schlüsse ziehen Sie daraus im Hinblick auf den Einfluss des Stichprobenumfangs auf die Verteilungen?

Verwenden Sie nun alle Stichprobenumfänge von n zwischen 1.000 und 50.000 Elementen und stellen Sie den funktionalen Zusammenhang von n und der Standardabweichung des geschätzten Beta-Werts graphisch dar. Nach welcher mathematischen Formel entwickelt sich näherungsweise die Standardabweichung in Abhängigkeit von n? Zeichen Sie den Funktionsverlauf dieser mathematischen Formel ebenfalls in das Diagramm ein. Erläutern Sie anhand des Diagramms den Zusammenhang zwischen Datenmenge und Lernqualität eines Modells.