Intelligente Datenanalyse

Prüfung: Einkommensklassen (Projekt 1)

Die Projektaufgabe ist Teil der Prüfung *Intelligente Datenanalyse*. Jede Aufgabe soll durch einen Studenten selbständig bearbeitet und die Lösung innerhalb der mündlichen Prüfung vorgestellt werden. Ein Ausdruck des Python-Programmcodes und der Ergebnisse in Form eines Diagramms, Tabelle o.ä. werden vorausgesetzt; die Art der Präsentation der Ergebnisse ist dem Studenten freigestellt.

Problemstellung

Ein Meinungsforschungsinstitut möchte auf Basis von Personendaten (siehe einkommen.train) eine Aussage über deren Einkommen treffen. Dazu wurden 30.000 Personen nach den unten stehenden Größen befragt. Für einige Personen sind nicht alle Größen bekannt. Insbesondere ist das Einkommen nur für 5.000 der Befragten bekannt.

Sie wurden damit beauftragt, für die verbliebenen 25.000 Personen eine Vorhersage über deren Einkommensklasse zu treffen und die Daten so aufzubereiten, dass diese für weitere Regressions- und Korrelationsanalysen verwendet werden können.

- Alter
- Beschäftigungsverhältnis (Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked)
- Gewichtungsfaktor zum Ausgleichen des Umfrage-bedingten Auswahl-Bias
- Bildungsgrad (Bachelors, Some-college, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 1st-4th, 5th-6th, 7th-8th, 9th, 10th, 11th, 12th, Masters, Doctorate, Preschool)
- Schul-/Ausbildungsdauer
- Familienstand (Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse)
- Beschäftigungsbereich (Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces)
- Partnerschaft (Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried)
- Ethnie (White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black)
- Geschlecht (Female, Male)
- Gewinn aus Vermögenswerten
- Verlust aus Vermögenswerten
- Wochenarbeitszeit
- Geburtsland (United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong Kong, Holand-Netherlands)

- Einkommen ($\leq 50k$, $> 50k$)

Aufgabe

Lesen Sie die Daten in Python ein und führen Sie eine Datenvorverarbeitung durch. Wählen Sie dabei geeignete Transformationen, Normalisierungen usw. Ergänzen Sie fehlende Werte (durch „?“ markiert). Überlegen Sie, welche Art von Attributen der vorverarbeitete Datensatz enthalten sollte. Lernen Sie im Anschluss ein Modell zur Vorhersage der Einkommensklasse und wenden Sie es auf die 25.000 Personen mit unbekanntem Einkommen an. Identifizieren Sie dazu ein geeignetes Lernverfahren und implementieren Sie dieses in Python. Trainieren und evaluieren Sie das Modell. Begründen und dokumentieren Sie kurz alle durchgeführten Schritte.