**Training Fiche Vorlage**

|  |  |  |
| --- | --- | --- |
| **Titel** | LINEARE DISKRIMINANZANALYSE | |
| **Schlüsselwörter (Meta-Tags)** | Diskriminanzanalyse, Klassifizierung, R, Bayes'sche Analyse | |
| **Sprache** | Deutsch | |
| **Zielsetzungen / Ziele / Lernergebnisse** | **Ziel dieses Moduls ist es, die Grundlagen der linearen Diskriminanzanalyse (LDA) vorzustellen und zu erklären.**  **Am Ende dieses Moduls werden Sie in der Lage sein:**   * **Situationen zu identifizieren, in denen LDA nützlich sein kann** * **Berechnung von LDA-Funktionen** * **Interpretion der Ergebnisse der deskriptiven und prädiktiven LDA** | |
| **Lehrgang:** | | |
| **Datenwissenschaftliche Kompetenz** | |  |
| **Modul Datenvisualisierung und visuelle Analyse** | | **X** |
| **Einführung in die Datenwissenschaft für Human- und Sozialwissenschaften** | |  |
| **Datenwissenschaft für den guten Zweck** | |  |
| **Datenjournalismus und Geschichtenerzählen** | |  |
| **Beschreibung** | In diesem Schulungsmodul lernen wir die Anwendung der Diskriminanzanalyse (linear discriminant analysis bzw. LDA) eingeführt. LDA ist eine Methode zum Berechnen von Linearkombinationen von Variablen, die die Beobachtungen am besten in Gruppen oder Klassen einteilen, und wurde ursprünglich von Fisher (1936) entwickelt.  Diese Methode maximiert das Verhältnis der Varianz zwischen den Klassen zur Varianz innerhalb der Klassen in einem bestimmten Datensatz. Auf diese Weise wird die Variabilität zwischen den Gruppen maximiert, was zu einer maximalen Trennbarkeit führt.  LDA kann für reine Klassifizierungszwecke, aber auch für die Vorhersage von Klassenzugehörigkeiten eingesetzt werden. | |
| **Inhalt in 3 Ebenen gegliedert** | **1. EINLEITUNG: MOTIVATION DURCH EIN ANSCHAULICHES BEISPIEL**  Nehmen wir an, wir haben eine Stichprobe von Personen und beobachten, mit welchem Verkehrsmittel (Auto, öffentliche Verkehrsmittel oder zu Fuß) sie sich normalerweise in einer Stadt fortbewegen. Wir wissen, dass die Wahl des Verkehrsmittels teilweise von ihrem wirtschaftlichen Status beeinflusst wird, und wir beobachten Daten über ihr Alter in Jahren und ihr jährliches Haushaltseinkommen, zusammen mit dem gewählten Verkehrsmittel:    Wir wollen wissen, wie diese beiden Kovariablen zur Klassifizierung (d. h. zur Unterscheidung) der Personen beitragen, indem sie sie einer bestimmten Kategorie von Verkehrsmitteln zuordnen.  Es zeigt sich, dass es keine perfekte Klassifizierung gibt: Personen mit hohem Einkommen nutzen tendenziell häufiger das Auto, aber es gibt eine große Überschneidung der Kategorien "zu Fuß gehen" und "öffentliche Verkehrsmittel" bei Personen mit niedrigerem Einkommen. Auch bei der Verteilung nach Alter gibt es größere Überschneidungen zwischen den Kategorien: Ältere Personen gehen nicht zu Fuß, aber bei jüngeren Werten ist das Alter kein guter Prädiktor für die Verkehrsmittelwahl. Dies ist das typische Problem, mit dem sich die LDA befasst.  **2. LDA für die Klassifizierung**  **2.1. Formulierung**  LDA-Funktionen können als Hilfe bei der Klassifizierung der Daten auf der Grundlage einer Matrix von Kovariaten 𝐗 verwendet werden. Ähnlich wie bei der Hauptkomponentenanalyse (PCA) zielen die LDA-Funktionen darauf ab, eine lineare Kombination der ursprünglichen Daten zu finden:    wobei die Varianz zwischen den Klassen (𝐁) relativ zur Varianz innerhalb der Klassen (𝐖) maximiert wird, was als verallgemeinertes Eigenwertproblem betrachtet werden kann:    Die Koordinaten der Diskriminanten werden aus den Eigenvektoren von 𝐖^(-𝟏) 𝐁 gewonnen.  **2.2. Ein Beispiel**  Als anschauliches Beispiel lösen wir das Klassifizierungsproblem der Verkehrsmittel auf der Grundlage von Alter und Einkommen mit Hilfe von LDA in R. Dies kann einfach mit der Funktion "lda" in der Bibliothek "mass" durchgeführt werden. Für alle hier vorgestellten Analysen müssen wir die folgenden R-Pakete installieren und laden:    Die untersuchten Daten liegen in einer CSV-Datei (namens "trasnpor\_example") vor, die durch Ausführen dieses Codes leicht in R importiert werden kann:    Um sich einen ersten Eindruck von den Daten zu verschaffen, können wir die Stichprobe in Form eines Streudiagramms darstellen:    Die obigen Codezeilen ergeben das Streudiagramm, das im einleitenden Abschnitt dieses Dokuments gezeigt wird. Alternativ können wir die Daten auch als eine Reihe von Histogrammen darstellen:    Indem wir eine dieser beiden Code-Zeilen ausführen, können wir uns einen Eindruck davon verschaffen, wie sich die Verkehrsmittel auf die verschiedenen Alters- und Einkommenswerte verteilen. Zum Beispiel:    Oder:    Die LDA wird einfach ausgeführt:    Typischerweise zeigt die Ausgabe die anfänglichen Mittelwerte nach Gruppen, die Koeffizienten in den LD-Projektionen und den Anteil der Varianz zwischen den Gruppen (trace), den jede LD-Koordinate erklärt:    In unserem Beispiel ist die erste LD-Koordinate positiv mit dem Einkommen und negativ mit dem Alter korreliert und enthält fast 90 % der Variabilität zwischen den Klassen. Die zweite LD-Funktion weist eine positive, aber schwächere Korrelation mit beiden Variablen auf und ist nur für etwa 10 % der Variabilität zwischen den Klassen verantwortlich.  Die neuen Koordinaten werden durch Projektion der ursprünglichen Datenpunkte mit den LDA-Koeffizienten durch den Ausdruck 𝐮𝐓 𝐗 erzeugt. In diesen neuen Koordinaten sind die Beobachtungen deutlicher zwischen den Gruppen getrennt. In unserem Beispiel haben wir zwei LD-Koordinaten für jedes Individuum, gegeben durch Alter und Einkommen. Die Koordinaten, die der ersten LD-Funktion entsprechen, haben die größere Trennschärfe.  Wir können diese Trennschärfe leicht erkennen, indem wir in R ein Histogramm erstellen und dabei die ersten LD-Koordinaten auf die horizontale Achse setzen:    Das führt zu folgendem Ergebnis:    Dieses Diagramm zeigt, wie die Überschneidungen erheblich abnehmen. Mit anderen Worten, die erste LD-Koordinate (man bedenke, dass es sich um ein "Kompositum" handelt, das negativ mit dem Alter und positiv mit dem Einkommen korreliert) unterscheidet angemessen zwischen den Verkehrskategorien.  **3. LDA für die Vorhersage**  **3.1 Das Verfahren**  LDA kann nicht nur für (deskriptive) Klassifizierungszwecke verwendet werden, sondern auch für die Vorhersage der Klassenzugehörigkeit. Nehmen wir zum Beispiel an, dass wir Daten über das Alter und das jährliche Haushaltseinkommen einer (in der Stichprobe oder außerhalb der Stichprobe) befindlichen Person haben und vorhersagen möchten, welches Verkehrsmittel diese Person am ehesten benutzen wird. LDA kann uns dabei helfen, eine Vorhersage zu treffen, ähnlich wie bei multinominalen Logit- oder Probit-Modellen.  Für diese Vorhersage sind einige Annahmen erforderlich:   * die Gruppen sind mehrdimensional normal verteilt * Die Gruppen weisen die gleichen Varianzen-Kovarianzen auf   Die Formulierung der LDA für Vorhersagen ist verwandt mit der Formulierung des Bayes-Theorems zur Aktualisierung von Wahrscheinlichkeiten:  Sei 𝑔 die Anzahl der Gruppen und 𝑞𝑖 die Vorwahrscheinlichkeit (üblicherweise beobachtete relative Häufigkeiten) für die Gruppe 𝑖. Die (posteriore) Wahrscheinlichkeit der Zugehörigkeit zur Gruppe 𝐺𝑖 in Abhängigkeit von 𝐱, P(𝐺𝑖 |𝐱), kann wie folgt ausgedrückt werden:    Dabei handelt es sich um einen Bayes'schen Ansatz, der die vorherigen Wahrscheinlichkeiten 𝑞\_𝑖 auf der Grundlage der bedingten Wahrscheinlichkeiten P(𝐱|𝐺𝑖) aktualisiert. Unter den Normalitätsannahmen gilt:    Dabei ist |𝐖| die Determinante der klasseninternen Varianzmatrix und 𝐷𝑖 ist . Durch Einsetzen des Ausdrucks von in die Formel für erhalten wir:    **3.2. Ein Beispiel mit R**  Die LDA-Routine in R kann A-posteriori-Wahrscheinlichkeiten auf der Grundlage der zuvor beschriebenen Annahmen und Formulierungen erstellen. Die LDA-Funktionen ermöglichen die Vorhersage der wahrscheinlichsten Klassenzugehörigkeit für eine beliebige Person bei Vorliegen eines Vektors von Kovariaten (in diesem Beispiel Alter und Haushaltseinkommen).  Zur Veranschaulichung zeigt die nachstehende Tabelle die vorhergesagten Wahrscheinlichkeiten für jede Gruppe für eine Teilmenge von Personen in der Stichprobe. Die Prioritäten werden für jede der drei Verkehrsarten als identisch angenommen ().    Die vorhergesagte Klasse entspricht der höchsten für jede Person. Sie werden durch Anwendung der folgenden Routine in R-studio berechnet:    In den meisten Fällen sagt LDA die Gruppe, zu der eine Person gehört, korrekt voraus. Es gibt jedoch einige Fälle, in denen die LDA keine korrekte Vorhersage trifft. Diese Fälle entsprechen den sich überschneidenden Beobachtungen, die in der LDA-Klassifikation verbleiben. | |
| **Selbstbeurteilung (Multiple-Choice-Fragen und Antworten)** | LDA ist ein statistisches Verfahren, das es ermöglicht:   1. Daten in Gruppen einzuteilen 2. Klassenzugehörigkeit vorherzusagen 3. Beide Antworten sind richtig   Für die Anwendung von LDA zu Vorhersagezwecken sind folgende Annahmen erforderlich:   1. Mehrdimensionale Normalverteilung zwischen den Gruppen 2. Gleiche Varianzen - Kovarianzen zwischen den Gruppen 3. Beide Antworten sind richtig   LDA basiert auf der Maximierung des Verhältnisses:   1. der Variabilität zwischen Gruppen und innerhalb der Klasse 2. der Variabilität innerhalb von Gruppen versus der Gesamtvariabilität 3. der Gesamt- versus der gruppeninternen Variabilität | |
| **Ressourcen (Videos, Verweislinks)** |  | |
| **Verwandtes Material** |  | |
| **Verwandte PPT** |  | |
| **Literaturverzeichnis** | Boedeker, P., & Kearns, N. T. (2019). Lineare Diskriminanzanalyse zur Vorhersage der Gruppenzugehörigkeit: A user-friendly primer. Advances in Methods and Practices in Psychological Science, 2, 250-263. | |
| **Zur Verfügung gestellt von** | [Uniovi] | |