

SS 2007, Stochastik IV: Übungsblatt 8

Aufgabe 1 (Quadratische Diskriminanzanalyse)

Daten und Verteilungsannahmen

Betrachtet werden $p \geq 2$ Kriteriumsvariablen Y_1, Y_2, \dots, Y_p , für welche jeweils aus $g = 2$ Gruppen (Grundgesamtheiten) G_1 und G_2 , Stichproben der Umfänge n_1 und n_2 vorliegen. Angenommen wird, dass für jede der beiden Gruppen der Zufallsvektor $Y = (Y_1, Y_2, \dots, Y_p)'$ der p Kriteriumsvariablen p -dimensional normalverteilt ist, $Y|G_1 \sim N_p(\mu_1, \Sigma_1)$ und $Y|G_2 \sim N_p(\mu_2, \Sigma_2)$, mit jeweiligen Erwartungswertvektoren μ_1 und μ_2 , und (positiv definiten) Kovarianzmatrizen Σ_1 und Σ_2 . Mittels der Dichten notiert:

$$f_{G_i}(y) = (2\pi)^{-p/2} |\Sigma_i|^{-1/2} \exp[-0.5(y - \mu_i)' \Sigma_i^{-1} (y - \mu_i)] \quad (y \in \mathbb{R}^p, i = 1, 2).$$

Kosten

Weiters wird von einer Kostenmatrix, wie sie auf der Folie Seite 184 gegeben ist, ausgegangen. Im Fall einer Fehlklassifikation vom Typ

Wahre Gruppe G_1 / Klassifikation G_2

fallen die Kosten $c(2|1) \in \mathbb{R}_{>0}$ an, im Fall einer Fehlklassifikation vom Typ

Wahre Gruppe G_2 / Klassifikation G_1

die Kosten $c(1|2) \in \mathbb{R}_{>0}$. Bei korrekten Klassifikationen fallen keine Kosten an.

A-priori Wahrscheinlichkeiten

Die zwei Gruppen G_1 und G_2 treten mit beliebig vorgegebenen a-priori Wahrscheinlichkeiten $\pi_1, \pi_2 \in (0, 1)$ auf. Diese könnten zum Beispiel die Proportionen der beiden Gruppen in der Gesamtpopulation repräsentieren.

Herleitung Entscheidungsregel

Vorhersageproblem: Beobachtet werden für einen Fall dessen Ausprägungen in den Kriteriumsvariablen Y_i ($1 \leq i \leq p$), kurz $y = (y_1, y_2, \dots, y_p)'$. Vorhergesagt wird dessen Gruppenzugehörigkeit.

Entscheidet man sich für Gruppe G_1 betragen die erwarteten (bedingten) Kosten:

$$\begin{aligned}
& E(\text{Kosten} | Y = y, \text{Klassifikation } G_1) \\
&= c(1|1) \cdot P(G = G_1 | Y = y) + c(1|2) \cdot P(G = G_2 | Y = y) \\
&\stackrel{c(1|1)=0}{=} c(1|2) \cdot P(G = G_2 | Y = y) \\
&\stackrel{\text{Bayes-Regel}}{=} c(1|2) \cdot f_{G_2}(y) \pi_2 P(Y = y)^{-1}.
\end{aligned}$$

Entscheidet man sich für Gruppe G_2 betragen die erwarteten (bedingten) Kosten:

$$\begin{aligned}
& E(\text{Kosten} | Y = y, \text{Klassifikation } G_2) \\
&= c(2|2) \cdot P(G = G_2 | Y = y) + c(2|1) \cdot P(G = G_1 | Y = y) \\
&\stackrel{c(2|2)=0}{=} c(2|1) \cdot P(G = G_1 | Y = y) \\
&\stackrel{\text{Bayes-Regel}}{=} c(2|1) \cdot f_{G_1}(y) \pi_1 P(Y = y)^{-1}.
\end{aligned}$$

Für den ‘log-ratio’ dieser erwarteten Kosten gilt:

$$\begin{aligned}
\Theta &:= \ln \left[\frac{E(\text{Kosten} | Y = y, \text{Klassifikation } G_1)}{E(\text{Kosten} | Y = y, \text{Klassifikation } G_2)} \right] \\
&= \ln \left[\frac{c(1|2) f_{G_2}(y) \pi_2}{c(2|1) f_{G_1}(y) \pi_1} \right] \\
&= \ln \left[\frac{f_{G_2}(y)}{f_{G_1}(y)} \right] + \ln \left[\frac{c(1|2) \pi_2}{c(2|1) \pi_1} \right].
\end{aligned}$$

Setzt man die jeweiligen Dichten (p -dimensional normal) ein, so ergibt sich

$$\begin{aligned}
\Theta &= \ln \left[\frac{|\Sigma_2|^{-1/2} \exp \left[-0.5(y - \mu_2)' \Sigma_2^{-1} (y - \mu_2) \right]}{|\Sigma_1|^{-1/2} \exp \left[-0.5(y - \mu_1)' \Sigma_1^{-1} (y - \mu_1) \right]} \right] + \ln \left[\frac{c(1|2) \pi_2}{c(2|1) \pi_1} \right] \\
&= 0.5 \ln \left[\frac{|\Sigma_1|}{|\Sigma_2|} \right] + \ln \left[\underbrace{\frac{\exp \left[-0.5(y - \mu_2)' \Sigma_2^{-1} (y - \mu_2) \right]}{\exp \left[-0.5(y - \mu_1)' \Sigma_1^{-1} (y - \mu_1) \right]}}_{=: \gamma} \right] + \ln \left[\frac{c(1|2) \pi_2}{c(2|1) \pi_1} \right],
\end{aligned}$$

wobei γ (nach direkter Umformung) ausgedrückt werden kann als

$$\gamma = \exp \left[0.5y'(\Sigma_1^{-1} - \Sigma_2^{-1})y - (\mu_1' \Sigma_1^{-1} - \mu_2' \Sigma_2^{-1})y + 0.5(\mu_1' \Sigma_1^{-1} \mu_1 - \mu_2' \Sigma_2^{-1} \mu_2) \right].$$

Dies impliziert

$$\begin{aligned}\Theta &= 0.5 \ln \left[\frac{|\Sigma_1|}{|\Sigma_2|} \right] \\ &\quad + [0.5y'(\Sigma_1^{-1} - \Sigma_2^{-1})y - (\mu_1'\Sigma_1^{-1} - \mu_2'\Sigma_2^{-1})y + 0.5(\mu_1'\Sigma_1^{-1}\mu_1 - \mu_2'\Sigma_2^{-1}\mu_2)] \\ &\quad + \ln \left[\frac{c(1|2)\pi_2}{c(2|1)\pi_1} \right].\end{aligned}$$

Entscheidungsregel (in unbekanntem Populationsgrößen): Als Vorhersage der Gruppenzugehörigkeit für den Fall mit entsprechendem Ausprägungsvektor y fällt die

‘Entscheidung auf Gruppe G_1 ’

$$\begin{aligned}&:\iff E(\text{Kosten} | Y = y, \text{Klassifikation } G_1) < E(\text{Kosten} | Y = y, \text{Klassifikation } G_2) \\ &\iff \Theta < 0 \\ &\iff -0.5 \ln \left[\frac{|\Sigma_1|}{|\Sigma_2|} \right] \\ &\quad - [0.5y'(\Sigma_1^{-1} - \Sigma_2^{-1})y - (\mu_1'\Sigma_1^{-1} - \mu_2'\Sigma_2^{-1})y + 0.5(\mu_1'\Sigma_1^{-1}\mu_1 - \mu_2'\Sigma_2^{-1}\mu_2)] \\ &\quad > \ln \left[\frac{c(1|2)\pi_2}{c(2|1)\pi_1} \right] \\ &\iff -0.5y'(\Sigma_1^{-1} - \Sigma_2^{-1})y + (\mu_1'\Sigma_1^{-1} - \mu_2'\Sigma_2^{-1})y \\ &\quad - 0.5 \left[\ln \left[\frac{|\Sigma_1|}{|\Sigma_2|} \right] + (\mu_1'\Sigma_1^{-1}\mu_1 - \mu_2'\Sigma_2^{-1}\mu_2) \right] \\ &\quad > \ln \left[\frac{c(1|2)\pi_2}{c(2|1)\pi_1} \right],\end{aligned}$$

‘Entscheidung auf Gruppe G_2 ’

$$\begin{aligned}&:\iff E(\text{Kosten} | Y = y, \text{Klassifikation } G_1) > E(\text{Kosten} | Y = y, \text{Klassifikation } G_2) \\ &\iff \Theta > 0 \\ &\iff -0.5 \ln \left[\frac{|\Sigma_1|}{|\Sigma_2|} \right] \\ &\quad - [0.5y'(\Sigma_1^{-1} - \Sigma_2^{-1})y - (\mu_1'\Sigma_1^{-1} - \mu_2'\Sigma_2^{-1})y + 0.5(\mu_1'\Sigma_1^{-1}\mu_1 - \mu_2'\Sigma_2^{-1}\mu_2)] \\ &\quad < \ln \left[\frac{c(1|2)\pi_2}{c(2|1)\pi_1} \right] \\ &\iff -0.5y'(\Sigma_1^{-1} - \Sigma_2^{-1})y + (\mu_1'\Sigma_1^{-1} - \mu_2'\Sigma_2^{-1})y \\ &\quad - 0.5 \left[\ln \left[\frac{|\Sigma_1|}{|\Sigma_2|} \right] + (\mu_1'\Sigma_1^{-1}\mu_1 - \mu_2'\Sigma_2^{-1}\mu_2) \right]\end{aligned}$$

$$< \ln \left[\frac{c(1|2)\pi_2}{c(2|1)\pi_1} \right].$$

Entscheidungsregel (in bekannten Schätzgrößen): Schätzt man

die Erwartungswertvektoren μ_1 und μ_2 durch die entsprechenden Stichprobe Gruppenmittelwertvektoren

$$\begin{aligned}\bar{y}_1 &:= (\bar{y}_{11}, \bar{y}_{21}, \dots, \bar{y}_{p_1})', \\ \bar{y}_2 &:= (\bar{y}_{12}, \bar{y}_{22}, \dots, \bar{y}_{p_2})'\end{aligned}$$

(\bar{y}_{ij} empirischer Mittelwert zur Variablen Y_i gebildet nur über die Fälle der Gruppe j),

und die Kovarianzmatrizen Σ_1 und Σ_2 durch die entsprechenden Stichprobe Gruppenkovarianzmatrizen S_1 und S_2 (S_j empirische Kovarianzmatrix zum Vektor Y gebildet nur über die Fälle der Gruppe j),

so ergibt sich als Vorhersage der Gruppenzugehörigkeit für den Fall mit entsprechendem Ausprägungsvektor y folgende Entscheidungsregel:

‘Entscheidung auf Gruppe G_1 ’

$$\begin{aligned}\iff & -0.5y'(S_1^{-1} - S_2^{-1})y + (\bar{y}_1'S_1^{-1} - \bar{y}_2'S_2^{-1})y \\ & -0.5 \left[\ln \left[\frac{|S_1|}{|S_2|} \right] + (\bar{y}_1'S_1^{-1}\bar{y}_1 - \bar{y}_2'S_2^{-1}\bar{y}_2) \right] \\ & > \ln \left[\frac{c(1|2)\pi_2}{c(2|1)\pi_1} \right],\end{aligned}$$

‘Entscheidung auf Gruppe G_2 ’

$$\begin{aligned}\iff & -0.5y'(S_1^{-1} - S_2^{-1})y + (\bar{y}_1'S_1^{-1} - \bar{y}_2'S_2^{-1})y \\ & -0.5 \left[\ln \left[\frac{|S_1|}{|S_2|} \right] + (\bar{y}_1'S_1^{-1}\bar{y}_1 - \bar{y}_2'S_2^{-1}\bar{y}_2) \right] \\ & < \ln \left[\frac{c(1|2)\pi_2}{c(2|1)\pi_1} \right].\end{aligned}$$