Statistik IV für Nebenfachstudierende

5 Diskriminanzanalyse III

Prof. Dr. Andreas Mayr

Institut für Statistik, LMU München

Sommersemester 2017

Geschätzte Zuordnungsregel

- Bisher: Wir haben optimale Zuordnung bei bekannten Verteilung analysiert:

$$f(\underline{x}|r)p(r)$$

- Jetzt: Wir schätzen die relevanten Größen aus einer Lernstichprobe:

$$(Y_1, \underline{x}_1), ..., (Y_n, \underline{x}_n)$$

$$\mathcal{L} = (Y_i, \underline{x}_i), \quad i = 1, ..., n$$

- Daraus werden geschätzte Diskriminanzfunktionen $\widehat{d}_r(\underline{x})$ erzeugt, z.B.:

$\widehat{d}_r(\underline{x})$

Unmittelbar Posteriori schätzen

- $P(r \mid x)$ schätzen aus durch $\widehat{P}(r \mid x)$, z.B. durch mehrkategorielles Logit Modell:

$$\widehat{P}(r \mid x) = \frac{\exp(\underline{x}^\top \hat{\beta}_r)}{\sum_j \exp(\underline{x}^\top \hat{\beta}_j)}$$

Merkmalsdichte und a priori schätzen

- $\hat{p}(r)$ und $\hat{f}(\underline{x} \mid r)$ aus der Lernstichprobe $\mathcal{L}$ schätzen.
- $\hat{p}(r)$ durch relative Häufigkeiten r-te Klasse
- $\hat{f}(\underline{x} \mid r)$ entweder nicht-parametrisch oder parametrisch:
  - nicht-parametrisch: z.B. durch Dichteschätzer
  - parametrisch: z.B. $\underline{x} \sim N(\hat{\mu}_r, \hat{\underline{\underline{\Sigma}}})$

- Die geschätzte Regel $\hat{\delta}$ ist im Allgemeinen schlechter als die beste Regel δ^*_{Bayes}:

$$\varepsilon_{\hat{\delta}} = \sum_{r=1}^{k} \int_{\delta(x)\neq r} f(\underline{x} \mid r) p(r) dx$$

- Fehlerrate $\varepsilon_{\hat{\delta}}$ ist Zufallsvariable,weil $\hat{\delta}$ Zufallsvariable ist (durch Lernstichprobe geschätzt).
- Es gilt:

$$\varepsilon_{\delta^*} \leq \varepsilon_{\hat{\delta}}$$

- Um die Fehlerrate sinnvoll abschätzen zu können, sollte man nicht die Lernstichprobe verwenden, auf der $\hat{\delta}$ optimiert wurde.

  Probleme:

  - Unterschätzung des Fehlers
  - Overfitting

Komplexität: linear / quadratisch

- Bei parametrischer Schätzung
  - Linear: z.B. Logit Modelle mit $\eta_r = \underline{x}^\top \beta_r$
  - Quadratisch: z.B. Logit Modelle mit $\eta_r = \underline{x}^\top \beta_r + x_1^2 a_1 + x_2^2 a_2$
- Es gilt

  ε (wahrer Bayes quadratisch) $\leq \varepsilon$ (wahrer Bayes linear)
- Aber

  ε (gesch. Bayes quadratisch) $\quad$ ε (gesch. Bayes linear)

Komplexität: Anzahl Variablen

Fisher-Diskriminanzanalyse

Daten als Ausgangspunkt:

Gesucht ist die Linearkombination

$$Y = \underline{\mathbf{a}}^{\top} \underline{x}$$

welche die Daten optimal trennt, bei normierten $\underline{\mathbf{a}}$, meist

$||\underline{\mathbf{a}}|| = 1$

Bild

- Es kommt auf die Richtung der Projektion $\underline{\mathbf{a}}$ an:

Was wird optimiert?

- Projektion $y = \mathbf{a}^\top \underline{x}$ mit $\| \mathbf{a} \| = 1$, sodass das folgende Kriterium maximiert wird

$$Q(\mathbf{a}) = \frac{(\bar{y}_1 - \bar{y}_2)^2}{s_1^2 + s_2^2} \ ,$$

wobei

$$\begin{aligned} \bar{y}_i &= \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{a}^\top \underline{x}_{ij} = \mathbf{a}^\top \bar{\underline{x}}_i \quad \text{Klassenmittelpunkt,} \\ s_i^2 &= \sum (\mathbf{a}^\top x_{ij} - \mathbf{a}\bar{x}_i)^2 \quad \text{Quadr. Abweichungen } i\text{–te Klasse.} \end{aligned}$$

Was ist die Lösung?

- Ableiten und Nullsetzen liefert:

$$\mathbf{a} \quad = \quad \mathbf{W}^{-1}(\underline{\bar{x}}_1 - \underline{\bar{x}}_2)$$

mit

$$\mathbf{W} \quad = \quad (n_1 + n_2 - 2)\underline{\underline{\mathbf{S}}} = \sum_{i=1}^{2}\sum_{j=1}^{n_i}(\underline{x}_{ij} - \underline{\bar{x}}_i)(\underline{x}_{ij} - \underline{\bar{x}}_i)^\top.$$

Zusammenhang zur Zuordnung unter NV?

- Fisher-Diskriminanzanalyse liefert ähnliche Lösung wie bei Annahme von Normalverteilung für $\underline{x}$.
- Aber:
  - Fisher trifft keine Verteilungsannahme, Grundidee geht nur auf beste lineare Trennung zurück.
  - Ist somit robust gegenüber Verletzung der NV-Annahme.

Fisher Diskriminanzanalyse für mehr als zwei Klassen?

- Mehr–Klassen–Fall

  Kriterium:

$$Q(\mathbf{a}) = \frac{\sum_{i=1}^{g} n_i(\bar{y}_i - \bar{y})^2}{\sum_{i=1}^{g} s_i^2} \quad \rightarrow \quad \max$$

Fisher Diskriminanzanalyse für mehr als zwei Klassen?

- Mehr–Klassen–Fall

  Kriterium:

$$Q(\mathbf{a}) = \frac{\sum_{i=1}^{g} n_i(\bar{y}_i - \bar{y})^2}{\sum_{i=1}^{g} s_i^2} \quad \rightarrow \quad \max$$

Klassifikation mit q Projektionen

$$\delta(x) = r \Leftrightarrow \sum_{j=1}^{q} (\mathbf{a}^\top \underline{x} - \mathbf{a}^\top \underline{\bar{x}}_r)^2 = \min \sum_{i=1}^{q} (\mathbf{a}^\top \underline{x} - \mathbf{a}^\top \underline{x}_s)$$

...ist Minimum– Distanz–Klassifizierung

Ist identisch zur ML-Zuordnung unter der Annahme $\underline{x}|Y = r \sim N(\mu_r, \underline{\underline{\Sigma}})$ und

- μ_r durch $\underline{\bar{x}}_r$ ersetzen.
- $\underline{\underline{\Sigma}}$ durch $\underline{\underline{\mathbf{S}}}$ ersetzen.

Nicht-parametrische Verfahren

- Weder lineare noch quadratische Terme um Zuordnungsregeln zu finden, sondern *nicht-parametrisch*.
  - Nicht-parametrisch bedeutet z.B. ohne Annahme einer Verteilung.
  - Wird aber auch manchmal verwendet um zu verdeutlichen dass es sich nicht um eine klassische lineare oder quadratische Linearkombination handelt.
- Wir stellen zwei Verfahren dar:
  1. Nächste Nachbarn Regel
  2. Klassifikationsbäume

Nächste Nachbarn Regel

- Grundidee:

  Für eine neue Beobachtung $\tilde{x}$ wähle die Klasse, welche der möglichst *ähnlicher* x aus der Lernstichprobe entspricht.

Was sind nächste Nachbarn?

- Die Ähnlichkeit wird über ein Distanzmaß definiert:

  $d(x, \tilde{x}) \quad x, \tilde{x} \in \mathbb{R}^p$

- Es werden die k nächsten Nachbarn in der Lernstichprobe betrachtet, für welche die Eigenschaft gilt:

$$d(\underline{x}_{(1)}, \tilde{x}) \leq d(\underline{x}_{(2)}, \tilde{x}) \leq \ldots \leq d(\underline{x}_{(k)}, \tilde{x})$$

- $Y_{(1)}, Y_{(2)}, \ldots, Y_{(k)}$ sind die dazugehörigen Klassen für die k Nachbarn.

- Wähle

$$\hat{\delta}(\tilde{x}) = r \Leftrightarrow \text{häufigste Klasse in} \{Y_{(1)}, Y_{(2)}, \ldots, Y_{(k)}\}$$

Eigenschaften

- Funktioniert in allen Datensituationen, hängt von keiner Verteilung ab, keine Schätzungen.
- *Tuning parameter* ist die Anzahl der Nachbarn k:
  - optimale Anzahl liegt irgendwo zwischen 1 (sehr datennah) und n (x spielen keine Rolle).
- Nachteil: Keine fixe Zuordnungsregel, muss jedes Mal mit kompletter Lernstichprobe neu bestimmt werden

Klassifikationsbäume

- Sukzessive Partitionierung des Merkmalsraums (basierend auf dichotome Entscheidungen auf Grund einzelner x
- z.B. $x > c$ vs. $x \leq c$ mit Cutpoint c.
- Ziel ist die Bildung von möglichst homogener Untergruppen (homogen bzgl. Y)
- Diese Entscheidungsregeln werden anschließend auch für neue Beobachtungen angewandt.

Beispiel

Schätzung der Fehlerraten

- Lernstichprobe wird verwendet um Zuordnungsregel zu schätzen.
- Entscheidungen wie linear vs. quadratisch werden auf Lernstichprobe getroffen.
- Typischerweise resultiert das in einer Zuordnungsregel die für die Lernstichprobe relativ oft korrekt zuordnet.
- Eigentlich interessiert aber die Diskriminierung bei neuen Beobachtungen.

- Verwendung der Lernstichprobe ist somit zu optimistisch.
- Die eigentliche Fehlerrate für neue Beobachtungen wird wohl größer sein.
- Möglichkeiten:
  - Zusätzliche Test-Daten
  - Test-Daten aus Lernstichprobe bilden, z.B. Kreuzvalidierung, Bootstrapping, Subsampling.

Kreuzvalidierung

Bootstrapping

Subsampling