MUSTERERKENNUNG

Vorlesung im Sommersemester 2017

Prof. E.G. Schukat-Talamazzini

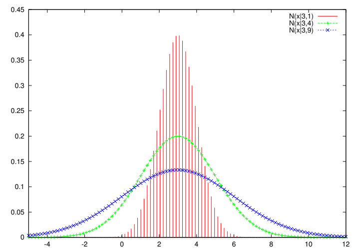
Stand: 6. März 2017

Teil VII

Normalverteilungsklassifikatoren

Univariate Normalverteilungsdichte

$$\mathcal{N}(x \mid \mu, \sigma^2) \stackrel{\text{def}}{=} \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$



Definition

Eine stetige Zufallsvariable $\mathbb{X}$ heißt (univariat) **normalverteilt** mit Mittelwert $\mu \in \mathbb{R}$ und Varianz $\sigma^2 \neq 0$, wenn gilt:

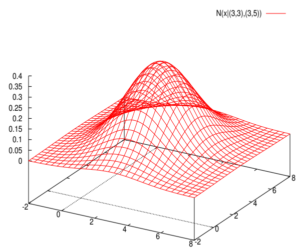
$$f_{\mathbb{X}}(x) = \mathcal{N}(x \mid \mu, \sigma^2)$$

Bemerkung

Unter der Annahme klassenweise *statistisch unabhängiger* und *normalverteilter* Merkmale läßt sich die (naive!) Bayesregel mit Hilfe von $K \cdot D$ univariaten NV-Dichten realisieren.

Bivariat unkorrelierte Normalverteilungsdichte

$$\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \sigma_1^2, \sigma_2^2) \stackrel{\text{def}}{=} \frac{1}{2\pi\sigma_1\sigma_2} \cdot \exp\left\{ -\frac{1}{2} \cdot \left(\frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} \right) \right\}$$



Definition

Eine stetiger Zufallsvektor $\mathbb{X} = (\mathbb{X}_1, \mathbb{X}_2)$ heißt **bivariat unkorreliert normalverteilt** mit Mittelwertvektor $\boldsymbol{\mu} \in \mathbb{R}^2$ und Varianzen $\sigma_1^2, \sigma_2^2 > 0$, wenn gilt:

$$f_{\mathbb{X}}(x_1, x_2) \;=\; \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \sigma_1^2, \sigma_2^2)$$

Bemerkung

Für Normalverteilungen sind *Unkorreliertheit* und *Unabhängigkeit* äquivalent. Obige Dichte entspricht also dem Produkt $\mathcal{N}(x_1 \mid \mu_1, \sigma_1^2) \cdot \mathcal{N}(x_2 \mid \mu_2, \sigma_2^2)$ der univariaten NV-Dichten (Randverteilungen).

Multivariate Normalverteilungsdichte

Definition

Ein Zufallsvektor $\mathbb{X} = (\mathbb{X}_1, \ldots, \mathbb{X}_D)^\top$ heißt **multivariat normalverteilt**, falls er der D-dimensionalen Verteilungsdichtefunktion

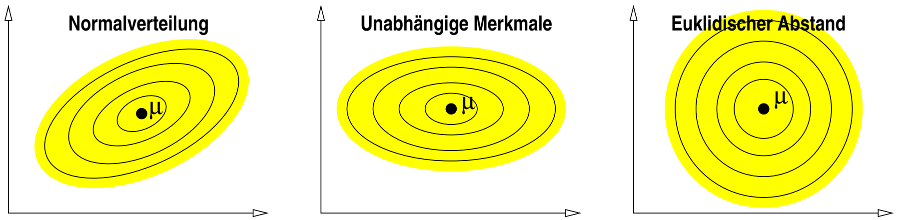
$$\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{S}) \stackrel{\text{def}}{=} \frac{1}{\sqrt{\det(2\pi\boldsymbol{S})}} \cdot \exp\left\{ -\frac{1}{2} \cdot (\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{S}^{-1} (\boldsymbol{x}-\boldsymbol{\mu}) \right\}$$

gehorcht. Es ist $\boldsymbol{\mu} \in \mathbb{R}^D$ der **Erwartungswertvektor** der Verteilung; die positiv-definite, symmetrische Matrix $\boldsymbol{S} \in \mathbb{R}^{D \times D}$ heißt **Kovarianzmatrix** der Normalverteilung.

Bemerkungen

1. Die Isolinien (Hyperebenen gleicher Dichtewerte) der multivariaten NV-Dichte besitzen die Form von *Hyperellipsoiden*.
2. Die Richtungen und Radien ihrer Achsen entnehmen wir den Eigenvektoren und Eigenwerten der Diagonalisierung $\boldsymbol{S} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^\top$.

Parameterreduzierte Normalverteilungsdichten



Symmetrisch positiv-definit	**Diagonalmatrix**	**Einheitsmatrix skaliert**
[redacted] $\begin{pmatrix} & \sigma_{12} & \cdots & \sigma_{1D} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2D} \\ \vdots & & \ddots & \vdots \\ \sigma_{D1} & \sigma_{D2} & \cdots & \sigma_{DD} \end{pmatrix}$	$\begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_D^2 \end{pmatrix}$	$\begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix}$
allgemeines Hyperellipsoid	Trägheitsachsen parallel zu Koordinatenachsen	skalierte Hypersphäre
$(D+1) \cdot D/2$ Parameter	D Parameter	1 Parameter

Ist $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{S})$ ein gutes Verteilungsmodell ?

Das kommt ganz auf die Anwendung & den Lerndatenvorrat an

Das NV-Modell ist zu simpel für unsere Daten

- Unimodale Dichtelandschaft — ? Löwe/Löwin
- Elliptische Symmetrie — ? nichtnegative Merkmale
- Exponentielles Abklingverhalten — ? Ausreißer

Das NV-Modell ist zu komplex für unseren Klassifikator

- Speicheraufwand $\mathrm{O}(D^2 \cdot K)$ — ? Bilder, Microarrays
- Rechenaufwand $\mathrm{O}(D^2 \cdot K)$ — ? Echtzeitanwendungen
- Robustheit der Schätzung $\hat{\boldsymbol{S}} = \boldsymbol{S}(\omega)$ — ? Rang und Inversenbildung

Multivariate Normalverteilungsdichte

Normalverteilungsklassifikatoren

Maximum-Likelihood Parameterschätzung

Maximum-a posteriori- und Bayesschätzung

Graphische Gaußsche Modelle

Mathematische Hilfsmittel

Normalverteilungsklassifikator

Definition

Einen Klassifikator mit den Prüfgrößen

$$u_\kappa(\boldsymbol{x}) \;=\; \mathrm{P}(\boldsymbol{x}, \Omega_\kappa) \;=\; p_\kappa \cdot \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_\kappa, \boldsymbol{S}_\kappa) \;, \qquad \boldsymbol{x} \in \mathbb{R}^D$$

für $\kappa = 1, \ldots, K$ bezeichnet man als D-dimensionalen **Normalverteilungsklassifikator** mit den Verteilungsparametern $[p_\kappa, \boldsymbol{\mu}_\kappa, \boldsymbol{S}_\kappa]_{\kappa=1..K}$.

Bemerkung

In der Praxis verwendet man einfachheitshalber die dazu *antitonen* Prüfgrößen

$$u_\kappa(\boldsymbol{x}) \;=\; -2 \cdot \log\left(\mathrm{P}(\boldsymbol{x}, \Omega_\kappa)\right) \;,$$

die quadratische Funktionen der Mustermerkmale sind.

Entscheidungsregel: ➪ Prüfgröße minimieren (Minuszeichen)

Prüfgrößen der NV-Bayesregel

Normalverteilungsklassifikator mit uneingeschränkten Kovarianzmatrizen $\boldsymbol{S}_\kappa$

$$u_\kappa(\boldsymbol{x}) \;=\; \underbrace{-2\log p_\kappa + \log|2\pi \boldsymbol{S}_\kappa|}_{\gamma_\kappa} \;+\; \underbrace{(\boldsymbol{x}-\boldsymbol{\mu}_\kappa)^\top \cdot \boldsymbol{S}_\kappa^{-1} \cdot (\boldsymbol{x}-\boldsymbol{\mu}_\kappa)}_{\text{Mahalanobisabstand } \|\boldsymbol{x}-\boldsymbol{\mu}_\kappa\|^2_{\boldsymbol{S}_\kappa}}$$

Bemerkungen

1. Je Klasse $1 + D + \binom{D+1}{2}$ Parameter ➪ $\mathrm{O}(D^2 K)$
2. Je Muster und Klasse $3D^2$ Addit./Multiplik. ➪ $\mathrm{O}(D^2 K)$

$$\tilde{\boldsymbol{x}}^\top \boldsymbol{S}_\kappa^{-1} \tilde{\boldsymbol{x}} \;=\; \sum_{i=1}^{D}\sum_{j=1}^{D} \tilde{x}_i c_{\kappa ij} \tilde{x}_j \;, \qquad \boldsymbol{C}_\kappa = \boldsymbol{S}_\kappa^{-1}$$

3. Für den Abstandsausdruck lohnt sich die folgende Betrachtung:

$$(\boldsymbol{x}-\boldsymbol{\mu}_\kappa)^\top \boldsymbol{S}_\kappa^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_\kappa) \;=\; \underbrace{\boldsymbol{x}^\top \boldsymbol{S}_\kappa^{-1}\boldsymbol{x}}_{\mathrm{spur}\left(\boldsymbol{S}_\kappa^{-1}\cdot \boldsymbol{x}\boldsymbol{x}^\top\right)} \underbrace{-2\boldsymbol{\mu}_\kappa^\top \boldsymbol{S}_\kappa^{-1}}_{\boldsymbol{a}_\kappa^\top}\boldsymbol{x} + \underbrace{\boldsymbol{\mu}_\kappa^\top \boldsymbol{S}_\kappa^{-1}\boldsymbol{\mu}_\kappa}_{c_\kappa}$$

Prüfgrößen der naiven NV-Bayesregel

Normalverteilungsklassifikator mit diagonalen Kovarianzmatrizen $\boldsymbol{S}_\kappa$

$$u_\kappa(\boldsymbol{x}) \;=\; \gamma_\kappa \;+\; \sum_{d=1}^{D}\left(\frac{x_d - \mu_{\kappa,d}}{\sigma_{\kappa,d}}\right)^2$$

mit der Konstanten

$$\gamma_\kappa \;=\; -2\log p_\kappa + D\cdot\log(2\pi) + \sum_d \log\sigma^2_{\kappa,d}$$

Bemerkungen

1. Je Klasse $1 + D + D$ Parameter ➪ $\mathrm{O}(DK)$
2. Je Muster und Klasse $4D$ Addit./Multipl./Divis. ➪ $\mathrm{O}(DK)$
3. Keine Merkmalkorrelationen — keine „schrägen" Klassengebiete!

Prüfgrößen der sphärischen NV-Bayesregel

Normalverteilungsklassifikator mit skalierter Einheitskovarianz $\boldsymbol{S}_\kappa = \sigma_\kappa^2 \boldsymbol{E}$

$$u_\kappa(\boldsymbol{x}) \;=\; \gamma_\kappa \;+\; \|\boldsymbol{x} - \boldsymbol{\mu}_\kappa\|^2 \;/\; \sigma_\kappa^2$$

mit der Konstanten

$$\gamma_\kappa \;=\; -2\log p_\kappa + D\cdot\log(2\pi) + 2D\cdot\log\sigma_\kappa$$

Bemerkungen

1. Je Klasse $1 + D + 1$ Parameter ➪ $\mathrm{O}(DK)$
2. Je Muster und Klasse $3D$ Addit./Multipl./Divis. ➪ $\mathrm{O}(DK)$
3. Klassengebiete $\hat{=}$ Hyperkugeln unterschiedlicher Radien

Prüfgrößen des Minimum-Abstand-Klassifikators

Normalverteilungsklassifikator mit Einheitskovarianz $\boldsymbol{S}_\kappa = \boldsymbol{E}$

$$u_\kappa(\boldsymbol{x}) \;=\; \gamma_\kappa \;+\; \|\boldsymbol{x} - \boldsymbol{\mu}_\kappa\|^2$$

mit der Konstanten

$$\gamma_\kappa \;=\; -2\log p_\kappa + D\cdot\log(2\pi)$$

Bemerkungen

1. Je Klasse $1 + D + 0$ Parameter ➪ $\mathrm{O}(DK)$
2. Je Muster und Klasse $2D$ Addit./Multipl./Divis. ➪ $\mathrm{O}(DK)$
3. Klassengebiete $\hat{=}$ Hyperkugeln identischer Radien
4. *Modifizierter MAK* — incl. Klassengewicht γ_κ
5. *Gewöhnlicher MAK* — excl. Klassengewicht γ_κ

Prüfgrößen des Mahalanobis-Klassifikators

Normalverteilungsklassifikator mit klassenunabhängiger Kovarianz $\boldsymbol{S}_\kappa = \boldsymbol{S}_0$

$$u_\kappa(\boldsymbol{x}) \;=\; \gamma_\kappa \;+\; \underbrace{(\boldsymbol{x} - \boldsymbol{\mu}_\kappa)^\top \cdot \boldsymbol{S}_0^{-1} \cdot (\boldsymbol{x} - \boldsymbol{\mu}_\kappa)}_{\textit{Mahalanobisabstand}\ \|\boldsymbol{x}-\boldsymbol{\mu}_\kappa\|^2_{\boldsymbol{S}_0}}$$

mit der Konstanten

$$\gamma_\kappa \;=\; -2\log p_\kappa + D\cdot\log(2\pi) + \log|\boldsymbol{S}_0|$$

Bemerkungen

1. Je Klasse $1 + D$ Parameter zzgl. $\boldsymbol{S}_0$ ➪ $\mathrm{O}(DK + D^2)$
2. Je Klasse $2D$ Addit./Multiplik. zzgl. quadr. Form ➪ $\mathrm{O}(DK + D^2)$
3. Für den Abstandsausdruck lohnt sich die folgende Betrachtung:

$$(\boldsymbol{x} - \boldsymbol{\mu}_\kappa)^\top \boldsymbol{S}_0^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_\kappa) \;=\; \underbrace{\boldsymbol{x}^\top \boldsymbol{S}_0^{-1} \boldsymbol{x}}_{\mathrm{spur}(\boldsymbol{S}_0^{-1}\cdot\boldsymbol{x}\boldsymbol{x}^\top)} \; \underbrace{-2\boldsymbol{\mu}_\kappa^\top \boldsymbol{S}_0^{-1}}_{\boldsymbol{a}_\kappa^\top} \boldsymbol{x} + \underbrace{\boldsymbol{\mu}_\kappa^\top \boldsymbol{S}_0^{-1} \boldsymbol{\mu}_\kappa}_{c_\kappa}$$

Prüfgrößen des Richter-Klassifikators

Normalverteilungsklassifikator mit isotrop skalierter Kovarianz $\boldsymbol{S}_\kappa = \alpha_\kappa \boldsymbol{S}_0$

$$u_\kappa(\boldsymbol{x}) \;=\; \gamma_\kappa \;+\; \underbrace{\alpha_\kappa^{-1} \cdot (\boldsymbol{x} - \boldsymbol{\mu}_\kappa)^\top \cdot \boldsymbol{S}_0^{-1} \cdot (\boldsymbol{x} - \boldsymbol{\mu}_\kappa)}_{\alpha_\kappa^{-1}\cdot\|\boldsymbol{x}-\boldsymbol{\mu}_\kappa\|^2_{\boldsymbol{S}_0}}$$

mit der Konstanten

$$\gamma_\kappa \;=\; -2\log p_\kappa + D\cdot\log(2\pi) + D\cdot\log\alpha_\kappa + \log|\boldsymbol{S}_0|$$

Bemerkungen

1. Je Klasse $1 + D + 1$ Parameter zzgl. $\boldsymbol{S}_0$ ➪ $\mathrm{O}(DK + D^2)$
2. Je Klasse $2D$ Addit./Multiplik. zzgl. quadr. Form ➪ $\mathrm{O}(DK + D^2)$
3. Für den Abstandsausdruck lohnt sich die folgende Betrachtung:

$$\boldsymbol{x}^\top \boldsymbol{S}_\kappa^{-1} \boldsymbol{x} \;=\; \underbrace{\alpha_\kappa^{-1} \cdot \mathrm{spur}\left(\boldsymbol{S}_0^{-1} \cdot \boldsymbol{x}\boldsymbol{x}^\top\right)}_{c_{\boldsymbol{x}}}$$

Prüfgrößen des Eigenraumklassifikators

Normalverteilungsklassifikator mit achsenparallelen Kovarianzen $\boldsymbol{S}_\kappa = \boldsymbol{U}\boldsymbol{D}_\kappa\boldsymbol{U}^\top$

$$u_\kappa(\boldsymbol{x}) \;=\; \gamma_\kappa \;+\; \underbrace{(\boldsymbol{U}^\top(\boldsymbol{x}-\boldsymbol{\mu}_\kappa))^\top \cdot \boldsymbol{D}_\kappa^{-1} \cdot (\boldsymbol{U}^\top(\boldsymbol{x}-\boldsymbol{\mu}_\kappa))}_{\left\|\boldsymbol{U}^\top(\boldsymbol{x}-\boldsymbol{\mu}_\kappa)\right\|^2_{\boldsymbol{D}_\kappa}}$$

mit der Konstanten

$$\gamma_\kappa \;=\; -2\log p_\kappa + D\cdot\log(2\pi) + \sum_d \log\lambda_{\kappa d}$$

Bemerkungen

1. Je Klasse $1+D+D$ Parameter zzgl. $\boldsymbol{U}$ ➪ $\mathrm{O}(DK+D^2)$
2. Je Klasse $4D$ Operationen für $\|\cdot\|^2_{\boldsymbol{D}_\kappa}$ zzgl. D^2 für $\boldsymbol{U}^\top\boldsymbol{x}$ ➪ $\mathrm{O}(DK+D^2)$
3. Für den Abstandsausdruck lohnt sich die folgende Betrachtung:

$$\boldsymbol{x}^\top\boldsymbol{S}_\kappa^{-1}\boldsymbol{x} \;=\; \boldsymbol{x}^\top\boldsymbol{U}\boldsymbol{D}_\kappa^{-1}\boldsymbol{U}^\top\boldsymbol{x} \;=\; (\boldsymbol{U}^\top\boldsymbol{x})^\top\boldsymbol{D}_\kappa^{-1}(\boldsymbol{U}^\top\boldsymbol{x}) \;=\; \sum_{d=1}^{D}(\boldsymbol{u}_d^\top\boldsymbol{x})^2\,/\;\lambda_{\kappa d}$$

4. Es kommt auch eine *unvollständige* Entwicklung in Betracht, bei der Trägheitsachsen mit kleinen Eigenwerten ignoriert werden …

Parameterschätzung für Wahrscheinlichkeitsmodelle

Verteilungsmodell ➪ Lerndaten ➪ Parameterschätzwert

Parametrische Verteilungsdichtefamilie

Die Wertetupel $\boldsymbol{x} \in \mathbb{R}^D$ eines Zufallsvektors $\mathbb{X}$ seien gemäß

$$\{f(x|\boldsymbol{\theta}) \mid \boldsymbol{\theta} \in \mathcal{M}\}$$

verteilt; jede Verteilungsdichte der Familie ist durch ein Feld $\boldsymbol{\theta}$ von Parametern aus einer Mannigfaltigkeit $\mathcal{M}$ charakterisiert.

Repräsentative Lernstichprobe

Die unbekannte Verteilung von $\mathbb{X}$ ist durch eine Stichprobe ω repräsentiert, deren Elemente $\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_T\}$ unabhängig und identisch gemäß $f(\cdot|\boldsymbol{\theta})$ verteilt gezogen wurden.

Problem

Wie lautet der beste Schätzwert $\hat{\boldsymbol{\theta}}$ für die unbekannten Parameter $\boldsymbol{\theta}^*$?

Maximum-Likelihood Schätzung

Lemma

Die (logarithmierte) Ziehungswahrscheinlichkeit für den unabhängig und identisch mittels $f(\cdot|\boldsymbol{\theta})$ gezogenen Datensatz ω beträgt

$$\ell_{\boldsymbol{\theta}}(\omega) \;=\; \log\prod_{x\in\omega} f(x|\boldsymbol{\theta}) \;=\; \sum_{x\in\omega}\log f(x|\boldsymbol{\theta}) \;.$$

*Die Größe $\ell_{\boldsymbol{\theta}}(\omega)$ heißt **Likelihoodfunktion** von $\boldsymbol{\theta}$.*

Definition

Die **Maximum-Likelihood-Schätzung** (MLS) der Parameter einer Dichtefamilie $[f(x|\boldsymbol{\theta})]$ maximiert die parameterbedingte Stichprobenwahrscheinlichkeit, d.h. es gilt

$$\hat{\boldsymbol{\theta}}_{\text{ML}} \;=\; \underset{\boldsymbol{\theta}}{\text{argmax}}\prod_{x\in\omega} f(x|\boldsymbol{\theta}) \;=\; \underset{\boldsymbol{\theta}}{\text{argmax}}\sum_{x\in\omega}\log f(x|\boldsymbol{\theta}) \;.$$

Bemerkung

Der ML-Schätzwert $\hat{\boldsymbol{\theta}}_{\text{ML}}$ ist von allen Parameterwerten derjenige, zu dem die vorliegenden Daten ω am besten passen.

Maximum-Likelihood Schätzung

Satz

Der ML-Schätzer ist ***erwartungstreu****, d.h.: ist eine Zufallsvariable $\mathbb{X}$ gemäß $f(x|\boldsymbol{\theta}^\star)$ verteilt, so ist der Erwartungswert des ML-Schätzers für eine Stichprobe unabhängiger Realisierungen von $\mathbb{X}$ gleich $\boldsymbol{\theta}^\star$.*

Bemerkungen

1. Für eine repräsentative Lernstichprobe zunehmenden Umfangs strebt der ML-Schätzwert gegen den *wahren* Parametervektor.
2. Über das Verhalten des ML-Schätzwertes bei Verwendung einer *individuellen, endlichen* Probe trifft der Satz keinerlei verbindliche Aussage.
3. Gehorcht der Datenerzeugungsprozeß nicht tatsächlich für irgendeinen festen Parameterwert $\boldsymbol{\theta} \in \mathcal{M}$ dem postulierten Verteilungsgesetz $f(x|\boldsymbol{\theta})$, so besitzen *selbst die asymptotischen* ML-Parameter $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ keine Aussagekraft.

ML-Schätzung für den NV-Klassifikator

Erzeugungswahrscheinlichkeit

einer unabhängig und identisch verteilten, etikettierten Stichprobe

$$\mathrm{P}(\bigcup_\kappa \omega_\kappa) = \prod_{\kappa=1}^{K} \mathrm{P}(\omega_\kappa) = \prod_{\kappa=1}^{K} \prod_{\boldsymbol{x}\in\omega_\kappa} \mathrm{P}(\Omega_\kappa)\cdot \mathrm{P}(\boldsymbol{x}|\Omega_\kappa)$$

Logarithmierte ML-Zielgröße

Parametrisiert durch $(p_\kappa, \boldsymbol{\theta}_\kappa)$, $\kappa = 1, \ldots, K$

$$\log \prod_{\kappa=1}^{K} \prod_{\boldsymbol{x}\in\omega_\kappa} p_\kappa \cdot f(\boldsymbol{x}|\boldsymbol{\theta}_\kappa) = \sum_{\kappa=1}^{K} T_\kappa \log p_\kappa + \sum_{\kappa=1}^{K} \left(\sum_{\boldsymbol{x}\in\omega_\kappa} \log f(\boldsymbol{x}|\boldsymbol{\theta}_\kappa) \right)$$

zerfällt in $(K+1)$ voneinander unabhängige Optimierungsprobleme

Beweis.

[Diskrete Verteilung $(p_1, \ldots, p_K)$ der Musterklassen]

Die ML-Zielfunktion lautet zunächst

$$\ell'_{\boldsymbol{p}}(\omega) = \log \prod_{\kappa=1}^{K} p_\kappa^{T_\kappa} = \sum_{\kappa=1}^{K} T_\kappa \log p_\kappa$$

und ist aber unter Berücksichtigung der Normierungsbedingung $\sum_\kappa p_\kappa = 1$ zu maximieren; die Bedingung wird mit einem Lagrange-Multiplikator inkorporiert:

$$\ell_{\boldsymbol{p}}(\omega) = \sum_{\kappa=1}^{K} T_\kappa \log p_\kappa - \lambda \cdot (\sum_\kappa p_\kappa - 1)$$

Wir bilden nun die partiellen Ableitungen

$$\frac{\partial \ell_{\boldsymbol{p}}(\omega)}{\partial p_\kappa} = T_\kappa \frac{1}{p_\kappa} - \lambda \quad \text{und} \quad \frac{\partial \ell_{\boldsymbol{p}}(\omega)}{\partial \lambda} = 1 - \sum_\kappa p_\kappa$$

Nullsetzen der Ableitungen ergibt

$$\frac{T_\kappa}{p_\kappa} = \lambda \quad \Rightarrow \quad p_\kappa = \frac{T_\kappa}{\lambda}$$

und wegen

$$1 = \sum_\kappa p_\kappa = \sum_\kappa \frac{T_\kappa}{\lambda} = \frac{1}{\lambda} \sum_\kappa T_\kappa = \frac{1}{\lambda} \cdot T$$

folgt $\lambda = T$ und daher $p_\kappa = {}^{T_\kappa}/_{T}$ für alle $\kappa = 1, \ldots, K$. □

ML-Schätzung für den NV-Klassifikator

mit vollbesetzten klassenabhängigen Kovarianzmatrizen

Satz

Die Maximum-Likelihood-Parameter eines Normalverteilungsklassifikators bezüglich einer etikettierten Stichprobe $[\omega_\kappa]$ lauten

$$\begin{aligned}
\hat{p}_\kappa &= T_\kappa \Big/ \sum_{\lambda=1}^{K} T_\lambda \\
\hat{\boldsymbol{\mu}}_\kappa &= \frac{1}{T_\kappa} \sum_{\boldsymbol{x}\in\omega_\kappa} \boldsymbol{x} \\
\hat{\boldsymbol{S}}_\kappa &= \frac{1}{T_\kappa} \sum_{\boldsymbol{x}\in\omega_\kappa} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_\kappa)(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_\kappa)^\top \\
&= \frac{1}{T_\kappa} \sum_{\boldsymbol{x}\in\omega_\kappa} \boldsymbol{x}\boldsymbol{x}^\top - \hat{\boldsymbol{\mu}}_\kappa \hat{\boldsymbol{\mu}}_\kappa^\top
\end{aligned}$$

Beweis.

[Parameter μ einer univariaten Gaußdichte]

$$f_{\mathbb{X}}(x) \;=\; \mathcal{N}(x \mid \mu, \sigma^2) \;=\; \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Die ML-Zielfunktion $\ell_{\mu,\sigma^2}(\omega) \;=\; -2 \cdot \log \prod_{x\in\omega} \mathcal{N}(x \mid \mu, \sigma^2)$ lautet

$$\ell_{\mu,\sigma^2}(\omega) \;=\; -2 \cdot \sum_{x\in\omega} \left(-\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right) \;=\; T \cdot \log(2\pi\sigma^2) + \frac{1}{\sigma^2}\sum_{x\in\omega}(x-\mu)^2$$

Partielle Ableitung nach μ:

$$\frac{\partial\ell(\omega)}{\partial\mu} \;=\; \frac{1}{\sigma^2}\sum_{x\in\omega} 2\cdot(x-\mu)\cdot(-1) \;=\; -\frac{2}{\sigma^2}\left(\sum_{x\in\omega} x - \sum_{x\in\omega}\mu\right)$$

Nullsetzen ergibt

$$\sum_{x\in\omega} x \;=\; \sum_{x\in\omega}\mu \;=\; T\cdot\mu \quad\Rightarrow\quad \hat{\mu} \;=\; \frac{1}{T}\sum_{x\in\omega} x$$

□

Beweis.

[Parameter σ^2 einer univariaten Gaußdichte bei bekanntem Wert μ]

$$f_{\mathbb{X}}(x) \;=\; \mathcal{N}(x \mid \mu, \sigma^2) \;=\; \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Die ML-Zielfunktion $\ell_{\mu,\sigma^2}(\omega) \;=\; -2 \cdot \log \prod_{x\in\omega} \mathcal{N}(x \mid \mu, \sigma^2)$ lautet

$$\ell_{\mu,\sigma^2}(\omega) \;=\; -2 \cdot \sum_{x\in\omega} \left(-\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right) \;=\; T \cdot \log(2\pi\sigma^2) + \frac{1}{\sigma^2}\sum_{x\in\omega}(x-\mu)^2$$

Partielle Ableitung nach σ^2:

$$\frac{\partial\ell(\omega)}{\partial\sigma^2} \;=\; T\cdot\frac{1}{2\pi\sigma^2}\cdot 2\pi - \frac{1}{\sigma^4}\sum_{x\in\omega}(x-\mu)^2 \;=\; \frac{1}{\sigma^2}\left(T - \frac{1}{\sigma^2}\sum_{x\in\omega}(x-\mu)^2\right)$$

Nullsetzen ergibt

$$T \;=\; \frac{1}{\sigma^2}\sum_{x\in\omega}(x-\mu)^2 \quad\Rightarrow\quad \hat{\sigma}^2 \;=\; \frac{1}{T}\sum_{x\in\omega}(x-\mu)^2$$

□

Bemerkung

In der Praxis ist mit σ^2 natürlich auch μ unbekannt und es muß unter Zuhilfenahme des ML-Schätzwertes $\hat{\mu}$ optimiert werden. Eine Rechnung ähnlich der obigen ergibt die Varianzschätzformel

$$\hat{\sigma}^2 \;=\; \frac{1}{T-1}\sum_{x\in\omega}(x-\hat{\mu})^2 \;.$$

Beweis.

[Parameter $\boldsymbol{\mu}$ einer multivariaten Gaußdichte]

$$\mathcal{N}(x \mid \boldsymbol{\mu}, \boldsymbol{S}) \;=\; |2\pi\boldsymbol{S}|^{-1/2} \cdot \exp\left(-\tfrac{1}{2}(x-\boldsymbol{\mu})^\top \boldsymbol{S}^{-1}(x-\boldsymbol{\mu})\right)$$

Die ML-Zielfunktion lautet

$$\begin{aligned}
\ell_{\boldsymbol{\mu},\boldsymbol{S}}(\omega) \;=\; -2\cdot\log\prod_{x\in\omega}\mathcal{N}(x \mid \boldsymbol{\mu}, \boldsymbol{S}) \;&=\; -2\sum_{x\in\omega}\left(-\frac{1}{2}\log|2\pi\boldsymbol{S}| - \frac{1}{2}(x-\boldsymbol{\mu})^\top\boldsymbol{S}^{-1}(x-\boldsymbol{\mu})\right) \\
&=\; T\log|2\pi\boldsymbol{S}| + \sum_{x\in\omega}(x-\boldsymbol{\mu})^\top\boldsymbol{S}^{-1}(x-\boldsymbol{\mu}) \\
&=\; T\log|2\pi\boldsymbol{S}| + \sum_{x\in\omega}\left(x^\top\boldsymbol{S}^{-1}x - 2x^\top\boldsymbol{S}^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}^\top\boldsymbol{S}^{-1}\boldsymbol{\mu}\right)
\end{aligned}$$

Partielle Ableitung nach $\boldsymbol{\mu}$ (Gradientenvektor):

$$\begin{aligned}
\nabla_{\boldsymbol{\mu}}\ell_{\boldsymbol{\mu},\boldsymbol{S}}(\omega) \;&=\; 0 - 0 + \sum_{x\in\omega}\nabla_{\boldsymbol{\mu}}\left(x^\top\boldsymbol{S}^{-1}x - 2x^\top\boldsymbol{S}^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}^\top\boldsymbol{S}^{-1}\boldsymbol{\mu}\right) \\
&=\; \sum_{x\in\omega}\left(0 - 2\cdot\boldsymbol{S}^{-1}x + 2\cdot\boldsymbol{S}^{-1}\boldsymbol{\mu}\right) \;=\; 2\cdot\boldsymbol{S}^{-1}\sum_{x\in\omega}(\boldsymbol{\mu}-x) \;=\; 2\cdot\boldsymbol{S}^{-1}\left(T\boldsymbol{\mu} - \sum_{x\in\omega}x\right)
\end{aligned}$$

Nullsetzen und Multiplikation mit $\tfrac{1}{2}\cdot\boldsymbol{S}$ ergibt

$$T\boldsymbol{\mu} \;=\; \sum_{x\in\omega}x \quad\Rightarrow\quad \boldsymbol{\mu} \;=\; {}^{1}\!/_{T}\sum_{x\in\omega}x$$

□

Beweis.

[Parameter $\boldsymbol{S}$ einer multivariaten Gaußdichte]

Die ML-Zielfunktion lautet

$$\begin{aligned}
\ell_{\boldsymbol{\mu},\boldsymbol{S}}(\omega) \;&=\; T\log|2\pi\boldsymbol{S}| + \sum_{x\in\omega}(x-\boldsymbol{\mu})^\top\boldsymbol{S}^{-1}(x-\boldsymbol{\mu}) \\
&=\; TD\log(2\pi) - T\log|\boldsymbol{S}^{-1}| + \sum_{x\in\omega}\mathrm{spur}\left(\boldsymbol{S}^{-1}(x-\boldsymbol{\mu})(x-\boldsymbol{\mu})^\top\right) \\
&=\; TD\log(2\pi) - T\log|\boldsymbol{S}^{-1}| + \underbrace{\mathrm{spur}\left(\boldsymbol{S}^{-1}\cdot\sum_{x\in\omega}(x-\boldsymbol{\mu})(x-\boldsymbol{\mu})^\top\right)}_{T\cdot\mathrm{spur}\left(\boldsymbol{S}^{-1}\cdot\hat{\boldsymbol{S}}\right)}
\end{aligned}$$

Wir reformulieren die Zielgröße unter Verwendung der *inversen* Kovarianzmatrix $\boldsymbol{Q} = \boldsymbol{S}^{-1}$:

$$\ell_{\boldsymbol{\mu},\boldsymbol{Q}}(\omega) \;=\; TD\log(2\pi) - T\log|\boldsymbol{Q}| + T\cdot\mathrm{spur}\left(\boldsymbol{Q}\cdot\hat{\boldsymbol{S}}\right)$$

Und nun leiten wir partiell nach der *inversen* Kovarianzmatrix ab:

$$\nabla_{\boldsymbol{Q}}\ell_{\boldsymbol{\mu},\boldsymbol{Q}}(\omega) \;=\; 0 - T\cdot\boldsymbol{Q}^{-1} + T\cdot\hat{\boldsymbol{S}} \;=\; T\cdot\left(\hat{\boldsymbol{S}} - \boldsymbol{Q}^{-1}\right) \;=\; T\cdot\left(\hat{\boldsymbol{S}} - \boldsymbol{S}\right)$$

Nach dem Nullsetzen ergibt sich folglich

$$\boldsymbol{S} \;=\; \hat{\boldsymbol{S}} \;=\; \frac{1}{T}\sum_{x\in\omega}(x-\boldsymbol{\mu})(x-\boldsymbol{\mu})^\top$$

□

ML-Schätzung für den NV-Klassifikator

Diagonale Kovarianzmatrizen & Mahalanobis-Klassifikator

Diagonale Kovarianzen

Die ML-Zielgröße zerfällt auf Grund der Unabhängigkeitsannahme in $(1 + K \cdot D)$ unabhängige Optimierungsterme.

$$\hat{\sigma}^2_{\kappa,d} \;=\; \frac{1}{T_\kappa} \sum_{\boldsymbol{x} \in \omega_\kappa} (x_d - \mu_{\kappa,d})^2$$

Mahalanobis-Klassifikator

Bei klassenübergreifenden Kovarianzstatistiken zerfällt $\ell_{\boldsymbol{\theta}}(\cdot)$ nicht mehr vollständig in klassenspezifische Optimierungsausdrücke!

$$\hat{\boldsymbol{S}}_0 \;=\; \boldsymbol{S}_W([\omega_\kappa]) \;=\; \frac{1}{T} \sum_{\kappa=1}^{K} \sum_{\boldsymbol{x} \in \omega_\kappa} (\boldsymbol{x} - \boldsymbol{\mu}_\kappa)(\boldsymbol{x} - \boldsymbol{\mu}_\kappa)^\top$$

Einphasige Berechnung von $\hat{\boldsymbol{S}}_0$ ist möglich: $\boldsymbol{S}_W = \boldsymbol{S} - \boldsymbol{S}_B$

ML-Schätzung für den NV-Klassifikator

Richter-Modell: ähnliche Klassenkovarianzen $\boldsymbol{S}_\kappa \;=\; \alpha_\kappa \boldsymbol{S}_0$

Iterationsanfang

Berechne Probenstatistiken und initiale Skalierungsfaktoren:

$$\hat{p}_\kappa \;=\; \frac{T_\kappa}{T} \qquad \hat{\boldsymbol{\mu}}_\kappa \;=\; \frac{1}{T_\kappa} \sum_{\boldsymbol{x} \in \omega_\kappa} \boldsymbol{x}$$

$$\alpha_\kappa^{(0)} \;=\; 1 \qquad \hat{\boldsymbol{S}}_\kappa \;=\; \frac{1}{T_\kappa} \sum_{\boldsymbol{x} \in \omega_\kappa} \boldsymbol{x}\boldsymbol{x}^\top - \hat{\boldsymbol{\mu}}_\kappa \hat{\boldsymbol{\mu}}_\kappa^\top$$

Iterationsschritt

Berechne Kovarianzprototyp und Skalierungsfaktoren für $i = 1, 2, \ldots$:

$$\boldsymbol{S}_0^{(i)} \;=\; \sum_{\kappa=1}^{K} \hat{p}_\kappa \cdot (\alpha_\kappa^{(i-1)})^{-1} \cdot \hat{\boldsymbol{S}}_\kappa$$

$$\alpha_\kappa^{(i)} \;=\; \frac{1}{D} \cdot \mathrm{spur}\left(\hat{\boldsymbol{S}}_\kappa \cdot (\boldsymbol{S}_0^{(i)})^{-1} \right)$$

ML-Schätzung und Lernstichprobenumfang

Problem

In der NVK-Prüfgröße treten die ***Inversen*** *und die* ***reziproken Determinanten*** *aller $\hat{\boldsymbol{S}}_\kappa$ auf!*

1. *Der Varianz-MLS $\hat{\sigma}_{\kappa,d}$ wird Null, sobald $|\omega_\kappa| \leq 1$ ist.*
2. *Der Kovarianz-MLS $\hat{\boldsymbol{S}}_\kappa$ wird singulär, sobald $|\omega_\kappa| \leq D$ ist.*
3. *Selbst für Klassen mit $|\omega_\kappa| > D$ besitzt $\hat{\boldsymbol{S}}_\kappa$ häufig schlechte Kondition.*

➩ *Schwierigkeiten für kleine T, große D, große K.*

Lösung

Verringerung der ***Modellkapazität*** *(Anzahl freier Parameter)*

1. *Fixierung und/oder Verklebung von Parametern*
2. *Strukturierung von Variablenabhängigkeiten*
3. *Wissensbasierte Engführung des Parameterraums*

Maximum-a posteriori Schätzung

Verteilungsparameter $\boldsymbol{\theta}$ als Werte einer Zufallsvariablen Θ

Bayesscher Denkansatz

Die wahren Verteilungsparameter $\boldsymbol{\theta}^*$ des Prozesses sind nicht nur **unbekannt**, sie sind sogar **stochastisch**.

Ihre Verteilungsdichte $f_\Theta(\cdot)$ repräsentiert unser **Vorwissen** über ihre möglichen Werte(kombinationen).

Lemma

*Sind die Parameter der Verteilungsfamilie $\{f_{\mathbb{X}}(\cdot|\boldsymbol{\theta})\}_{\boldsymbol{\theta}\in\mathcal{M}}$ selbst gemäß **a priori Dichte** $f_\Theta(\boldsymbol{\theta})$ verteilt, so lautet — für den unabhängig und identisch gezogenen Datensatz ω — die datenbedingte **a posteriori Dichte** der Parameter*

$$\mathrm{P}(\boldsymbol{\theta}|\omega) \;=\; \frac{\mathrm{P}(\boldsymbol{\theta})\cdot\mathrm{P}(\omega|\boldsymbol{\theta})}{\mathrm{P}(\omega)} \;=\; \frac{f_\Theta(\boldsymbol{\theta})\cdot\prod\limits_{\boldsymbol{x}\in\omega} f_{\mathbb{X}}(\boldsymbol{x}|\boldsymbol{\theta})}{\mathrm{P}(\omega)}\;.$$

Maximum-a posteriori Schätzung

Die im Lichte der Datenprobe wahrscheinlichsten Verteilungsparameter

Definition

Die **Maximum-a posteriori-Schätzung** (MAP) der Parameter einer Dichtefamilie $[f(\boldsymbol{x}|\boldsymbol{\theta})]$ unter Annahme der **a priori**-Verteilungsdichte $f_\Theta(\boldsymbol{\theta})$ für $\boldsymbol{\theta}$ maximiert die stichprobenbedingte Wahrscheinlichkeit des gesuchten Parameterfeldes, d.h. es gilt:

$$\hat{\boldsymbol{\theta}}_{\mathrm{MAP}} \;=\; \underset{\boldsymbol{\theta}}{\mathrm{argmax}} \left(f_\Theta(\boldsymbol{\theta})\cdot\prod_{\boldsymbol{x}\in\omega} f_{\mathbb{X}}(\boldsymbol{x}|\boldsymbol{\theta})\right)$$

Bemerkungen

1. Der MAP-Schätzwert $\hat{\boldsymbol{\theta}}_{\mathrm{MAP}}$ ist von allen Parameterwerten derjenige, der zu den vorliegenden Daten ω am besten paßt.
2. Hand aufs Herz — niemand (außer dem *Capo di tutti capi*) kennt diese mysteriöse Dichte $f_\Theta(\cdot)$.

Wissenswertes über die Maximum-a posteriori Schätzung

Spezialfall Maximum-Likelihood

Unter Gleichverteilungsannahme für $f_\Theta(\cdot)$ mutiert die MAP-Schätzung in eine ML-Schätzung.

Asymptotisches Schätzverhalten

Für große Stichproben ($|\omega|\to\infty$) strebt $\hat{\boldsymbol{\theta}}_{\mathrm{MAP}}$ gegen $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$.

Methode der konjugierten Dichtefamilien

Die *analytische* Optimierung der MAP-Zielfunktion erfordert eine geeignete Form der a priori-Dichte:

$$f_\Theta(\boldsymbol{\theta}) \mathrel{\hat{=}} C\cdot \prod_{\boldsymbol{z}\in\omega_{\text{prior}}} f_{\mathbb{X}}(\boldsymbol{z}|\boldsymbol{\theta})$$

Mit dieser Wahl gilt nämlich

$$\hat{\boldsymbol{\theta}}_{\mathrm{MAP}}(\omega) \;=\; \hat{\boldsymbol{\theta}}_{\mathrm{ML}}(\omega\cup\omega_{\text{prior}})$$

und das Problem der $f_\Theta(\cdot)$-Findung ist auf elegante Art gelöst!

MAP-Schätzung für diskrete Verteilungen

Wahrscheinlichkeitsparameter $p_1+p_2+\ldots+p_K=1$ für K Ereignisse

Definition

Der Zufallsvektor $\Theta=(\Theta_1,\ldots,\Theta_K)^\top\in[0,1]^K$ mit $\sum_\ell \Theta_\ell = 1$ heißt **Dirichlet-verteilt** mit den **Hyperparametern** $r_1,\ldots,r_K>-1$ genau dann, wenn gilt:

$$f_\Theta(\boldsymbol{p}) \;=\; \mathcal{D}(\boldsymbol{p}|\boldsymbol{r}) \;=\; C\cdot\prod_{\ell=1}^{K} p_\ell^{r_\ell}$$

Bemerkungen

1. Für $\boldsymbol{r}=\boldsymbol{0}$ ist $\mathcal{D}(\boldsymbol{p}|\boldsymbol{r})$ eine **Gleichverteilung**.
2. Für $\boldsymbol{r}=\boldsymbol{1}$ nimmt $\mathcal{D}(\boldsymbol{p}|\boldsymbol{r})$ ihr Dichtemaximum bei der **Gleichverteilung** $p_\ell\equiv {}^1\!/_K$ an.
3. Allgemein nimmt $\mathcal{D}(\boldsymbol{p}|\boldsymbol{r})$ ihr Dichtemaximum bei der Verteilung $\boldsymbol{p}\propto\boldsymbol{r}$ an, also für die Wahrscheinlichkeiten $p_\ell = r_\ell/R$, $R=\sum_i r_i$.
4. Der Dichtegipfel ist umso steiler, je größer der Skalenfaktor R ist.

MAP-Schätzung für diskrete Verteilungen

Satz

Gehorchen die kanonischen Parameter $p_1, \ldots, p_K$ einer diskreten Wahrscheinlichkeitsverteilung der Dirichletverteilung mit Hyperparametern $\boldsymbol{r} \in \mathbb{R}^K$, so lautet der MAP-Schätzwert für eine Stichprobe mit den absoluten Ereignishäufigkeiten $T_1 + T_2 + \ldots + T_K = T$

$$\hat{p}_\ell = \frac{T_\ell + r_\ell}{T + R} , \qquad R = \sum_{\ell=1}^{K} r_\ell .$$

Bemerkungen

1. Eine MAP-Schätzung mit Vorwissen $\mathcal{D}(\cdot|\boldsymbol{r})$ bewirkt die Aufstockung der Lerndaten ω um eine **virtuelle Datenprobe** ω_{prior} mit den Ereignishäufigkeiten r_ℓ; diese Werte müssen allerdings nicht unbedingt ganzzahlig sein.
2. Der Spezialfall einer gleichverteilten oder **uninformativen** Dirichletdichte ($r_\ell \equiv r_0$) ergibt die MAP-Schätzwerte (**Laplaceschätzformel** im Fall $r_0 = 1$)

$$\hat{p}_\ell = {}^{(T_\ell + r_0)}\!/\!_{(T + K \cdot r_0)} , \qquad \ell = 1, 2, \ldots, K .$$

Beweis.

Es beträgt die Stichprobenwahrscheinlichkeit

$$\mathrm{P}(\omega|\boldsymbol{p}) = \prod_{\kappa=1}^{K} p_\kappa^{T_\kappa}$$

und die a posteriori Parameterwahrscheinlichkeit (bei festen Hyperparametern)

$$\mathrm{P}(\boldsymbol{p}|\omega) \propto \mathrm{P}(\omega|\boldsymbol{p}) \cdot f_\Theta(\boldsymbol{p}) \propto \prod_{\kappa=1}^{K} p_\kappa^{T_\kappa} \cdot \prod_{\kappa=1}^{K} p_\kappa^{r_\kappa} \propto \prod_{\kappa=1}^{K} p_\kappa^{(T_\kappa + r_\kappa)}$$

Das Maximum nimmt $\mathrm{P}(\boldsymbol{p}|\omega)$ bekanntlich für diejenige Verteilung an, die proportional zu den Exponenten ist:

$$\hat{p}_\kappa = \frac{T_\kappa + r_\kappa}{T + R} , \qquad R = \sum_\kappa r_\kappa$$

□

Der MAP-Schätzwert ist ein gewichtetes Mittel („Konvexkombination") aus ML-Schätzwert und dem *Modus*

$$\rho_\kappa = r_\kappa / R , \qquad \kappa = 1, \ldots, K$$

der a priori-Dichte:

$$\hat{p}_\kappa = \frac{T_\kappa + r_\kappa}{T + R} = \frac{T_\kappa}{T + R} + \frac{r_\kappa}{T + R} = \underbrace{\frac{T_\kappa}{T}}_{\hat{p}_\kappa^{\mathrm{ML}}} \cdot \underbrace{\frac{T}{T + R}}_{\lambda} + \underbrace{\frac{r_\kappa}{R}}_{\rho_\kappa} \cdot \underbrace{\frac{R}{T + R}}_{(1-\lambda)}$$

MAP-Schätzung für die multivariate NV-Dichte

Definition

Eine Zufallsmatrix $\mathbb{S}$ über der Mannigfaltigkeit aller symmetrischen, positiv-definiten $(D \times D)$-Matrizen heißt **Wishart-verteilt** genau denn, wenn

$$f_{\mathbb{S}}(\boldsymbol{S}) = \mathcal{W}(\boldsymbol{S} \mid \alpha, \boldsymbol{V}) = \frac{1}{2^{\frac{\alpha D}{2}} |\boldsymbol{V}|^{\frac{\alpha}{2}} \Gamma_D(\frac{\alpha}{2})} \cdot |\boldsymbol{S}|^{\frac{\alpha - D - 1}{2}} \cdot \exp\left(-{}^1\!/\!_2 \cdot \operatorname{spur}\left(\boldsymbol{V}^{-1}\boldsymbol{S}\right)\right)$$

gilt mit den Hyperparametern $\alpha > D - 1$ und $\boldsymbol{V} \in \mathbb{R}^{D \times D}$ positiv-definit.

Lemma

Für die multivariate NV-Dichte $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{S})$ bildet das Produkt

$$f_\Theta(\boldsymbol{\mu}, \boldsymbol{S}) = \mathcal{N}(\boldsymbol{\mu} \mid \boldsymbol{m}, \tau^{-1}\boldsymbol{S}) \cdot \mathcal{W}(\boldsymbol{S}^{-1} \mid \alpha, \boldsymbol{V})$$

eine konjugierte Dichtefamilie mit den Hyperparametern $\boldsymbol{m} \in \mathbb{R}^D$, $\tau > 0$, $\alpha > D - 1$ und positiv-definiter Matrix $\boldsymbol{V} \in \mathbb{R}^{D \times D}$.

MAP-Schätzung für den NV-Klassifikator

Satz

Die Lerndaten $\omega_1, \ldots, \omega_K \subset \mathbb{R}^D$ eines numerischen Klassifikationsproblems seien klassenweise normalverteilt mit den unbekannten Parametern $(p_\kappa, \boldsymbol{\mu}_\kappa, \boldsymbol{S}_\kappa)$, $\kappa = 1, \ldots, K$. Die a priori Verteilung der Parameter sei definiert durch

$$f_\Theta(\boldsymbol{\theta}) = \mathcal{D}(\boldsymbol{p}|\boldsymbol{r}) \cdot \prod_{\kappa=1}^{K} \mathcal{N}(\boldsymbol{\mu}_\kappa \mid \boldsymbol{m}_\kappa, \tau_\kappa^{-1}\boldsymbol{S}_\kappa) \cdot \prod_{\kappa=1}^{K} \mathcal{W}(\boldsymbol{S}_\kappa^{-1} \mid \alpha_\kappa, \boldsymbol{V}_\kappa) .$$

Dann lauten die Maximum-a posteriori-Parameter:

$$\hat{p}_\kappa = \frac{r_\kappa + T_\kappa}{R + T} , \qquad R = \sum_\kappa r_\kappa$$

$$\hat{\boldsymbol{\mu}}_\kappa = \frac{1}{\tau_\kappa + T_\kappa} \left(\tau_\kappa \boldsymbol{m}_\kappa + \sum_{\boldsymbol{x} \in \omega_\kappa} \boldsymbol{x} \right)$$

$$\hat{\boldsymbol{S}}_\kappa = \frac{\boldsymbol{V}_\kappa + \tau_\kappa (\hat{\boldsymbol{\mu}}_\kappa - \boldsymbol{m}_\kappa)(\hat{\boldsymbol{\mu}}_\kappa - \boldsymbol{m}_\kappa)^\top + \sum_{\boldsymbol{x} \in \omega_\kappa} \boldsymbol{x}\boldsymbol{x}^\top - T_\kappa \hat{\boldsymbol{\mu}}_\kappa \hat{\boldsymbol{\mu}}_\kappa^\top}{(\alpha_\kappa - D) + T_\kappa}$$

„Plug-in"-Schätzverfahren

Die Suche nach den *unbekannten*, aber *wahren* Parametern

Traditionelles Induktionsparadigma

Die Verteilungsannahme $\omega \sim f_{\mathbb{X}}(\cdot|\theta)$ ist korrekt.

Es existiert eine wahre Parameterkonfiguration θ^* — wir müssen sie nur finden!

ML-Schätzung

$$\hat{\theta}_{\mathsf{ML}} = \underset{\theta}{\operatorname{argmax}}\, \mathrm{P}(\omega|\theta)$$

Posterior-Mean-Schätzung

$$\hat{\theta}_{\mathsf{PM}} = \mathcal{E}[\Theta|\omega] = \int \theta \cdot \mathrm{P}(\theta|\omega)\, d\theta$$

MAP-Schätzung

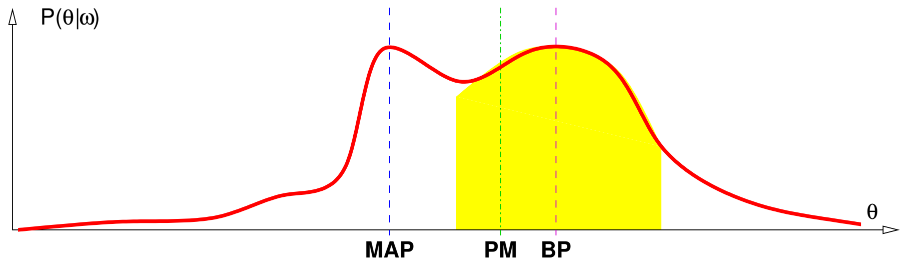
$$\hat{\theta}_{\mathsf{MAP}} = \underset{\theta}{\operatorname{argmax}}\, \mathrm{P}(\theta|\omega)$$

Bayespunkt-Schätzung

$$\hat{\theta}^{(\rho)}_{\mathsf{BP}} = \underset{\theta}{\operatorname{argmax}} \int_{\mathcal{U}_\rho(\theta)} \mathrm{P}(\vartheta|\omega)\, d\vartheta$$

„Plug-in"-Schätzverfahren

Analyse der a posteriori Parameterdichte



MAP Wo liegt der **Gipfel** der Posteriordichte?

PM Wo liegt der **Durchschnitt** der Posteriordichte?

BP Wo liegt das **kleinste Intervall** mit Wahrscheinlichkeitsmasse $\rho > 0$?

Bayes-Schätzung

Der Abschied von der Idee „wahrer" Verteilungsparameter

Bayessches Induktionsparadigma

Die Verteilungsannahme $\omega \sim f_{\mathbb{X}}(\cdot|\theta)$ ist korrekt.

Aber jedes $\boldsymbol{x} \in \omega$ wird unter Verwendung eines eigenen, zufällig ausgewürfelten Modellparameters θ gezogen!

$$\begin{aligned}
\mathrm{P}(\boldsymbol{x}|\omega) &= \int_{\mathcal{M}} \mathrm{P}(\boldsymbol{x}, \boldsymbol{\theta} \mid \omega)\, d\boldsymbol{\theta} \\
&= \int_{\mathcal{M}} \mathrm{P}(\boldsymbol{x} \mid \boldsymbol{\theta}, \omega) \cdot \mathrm{P}(\boldsymbol{\theta} \mid \omega)\, d\boldsymbol{\theta} \\
&= \int_{\mathcal{M}} \underbrace{f_{\mathbb{X}}(\boldsymbol{x}|\boldsymbol{\theta})}_{\text{Modelldichte}} \cdot \underbrace{\frac{f_{\mathbb{X}}(\omega|\boldsymbol{\theta}) \cdot f_{\Theta}(\boldsymbol{\theta})}{f_{\mathbb{X}}(\omega)}}_{\text{a posteriori}}\, d\boldsymbol{\theta}
\end{aligned}$$

Analytisch extrem schwer lösbar — bestenfalls wenn $f_\Theta(\cdot) \equiv c$

Bayesapproximation

Asymptotisch korrekte Näherung unter Gleichverteilungsannahme für $f_\Theta(\cdot)$

Praktikable Näherungslösung für den Bayesschätzer

Unwissen um $f_\Theta(\cdot) \rightsquigarrow$ Gleichverteilung $\rightsquigarrow$ Herauskürzen

Simultan in Zähler und Nenner: Integralbildung $\rightsquigarrow$ Maximumbildung

$$\begin{aligned}
\mathrm{P}(\boldsymbol{x}|\omega) = \frac{\mathrm{P}(\boldsymbol{x}, \omega)}{\mathrm{P}(\omega)} &= \frac{\int f_{\mathbb{X}}(\omega, \boldsymbol{x}|\boldsymbol{\theta}) \cdot f_\Theta(\boldsymbol{\theta})\, d\boldsymbol{\theta}}{\int f_{\mathbb{X}}(\omega|\boldsymbol{\theta}) \cdot f_\Theta(\boldsymbol{\theta})\, d\boldsymbol{\theta}} \\
&\approx \frac{\max_{\boldsymbol{\theta}} f_{\mathbb{X}}(\omega, \boldsymbol{x}|\boldsymbol{\theta})}{\max_{\boldsymbol{\theta}} f_{\mathbb{X}}(\omega|\boldsymbol{\theta})} = \frac{\prod_{\boldsymbol{z} \in \omega, \boldsymbol{x}} f_{\mathbb{X}}(\boldsymbol{z} \mid \hat{\boldsymbol{\theta}}_{\mathsf{ML}}(\omega, \boldsymbol{x}))}{\prod_{\boldsymbol{z} \in \omega} f_{\mathbb{X}}(\boldsymbol{z} \mid \hat{\boldsymbol{\theta}}_{\mathsf{ML}}(\omega))}
\end{aligned}$$

Achtung:

Die Bayesapproximation $\hat{\mathrm{P}}_{\mathsf{BA}}(\boldsymbol{x}|\omega)$ ist i.a. **keine** Dichtefunktion (Normierungseigenschaft)!

Graphische Gaußsche Modelle

Die Bias-Varianz-Problematik

Dichtemodell mit vielen Parametern

NV-Dichten mit voll besetzter Kovarianzmatrix
Alle paarweisen Merkmalabhängigkeiten $\rightsquigarrow \mathrm{O}(KD^2)$
Kleiner Bias — große Varianz

Dichtemodell mit wenigen Parametern

NV-Dichten mit diagonal besetzter Kovarianzmatrix
Alle Merkmale paarweise unabhängig $\rightsquigarrow \mathrm{O}(KD)$
Großer Bias — kleine Varianz

Lösungsidee

Nicht alle, sondern nur die **wichtigen** Merkmalabhängigkeiten werden explizit modelliert.

Gaußsche Bayesnetze

Kettenregel der Wahrscheinlichkeitstheorie

$$\mathrm{P}(x_1, \ldots, x_D) \;=\; \mathrm{P}(x_1) \cdot \mathrm{P}(x_2 | x_1) \cdot \prod_{d=3}^{D} \mathrm{P}(x_d \mid x_1, \ldots, x_{d-1})$$

Das d-te Merkmal ist explizit von $(d-1)$ anderen abhängig.

Beispiel: baumförmige Bayesnetze

$$\mathrm{P}(x_1, \ldots, x_D) \;\approx\; \prod_{d=1}^{D} \mathrm{P}(x_d \mid x_{\pi(d)})$$

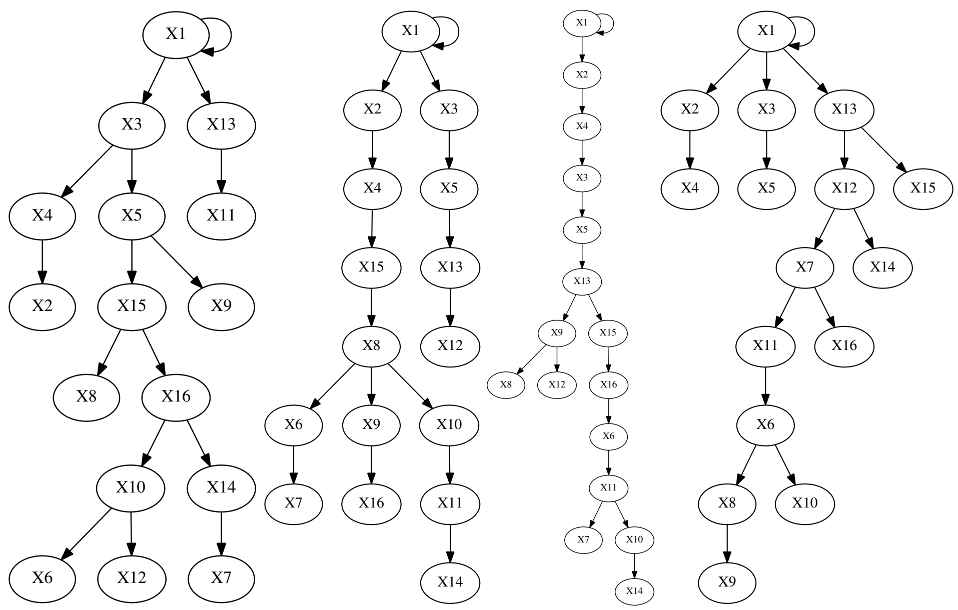
Jedes Merkmal x_d ist explizit nur von **genau einem** anderen abhängig.

Problem

Finde diejenige Abhängigkeitsstruktur, welche die exakteste Näherung der Datenverteilung gewährleistet!

Gaußsche Bayesnetze

Datensatz `letter.lern` (16 Merkmale, Klassen `'A'`, `'B'`, `'C'`, `'D'`)



Gaußsche Markovnetze

Parametrische Struktur der multivariaten NV-Dichte

$$-2 \cdot \log \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{S}) \;=\; |2\pi \boldsymbol{S}| + \sum_{i=1}^{D} \sum_{j=1}^{D} (x_i - \mu_i) \cdot C_{ij} \cdot (x_j - \mu_j)\,, \qquad \boldsymbol{C} := \boldsymbol{S}^{-1}$$

Modellkomplexität $\hat{=}$ Anzahl nicht verschwindender Einträge von $\boldsymbol{S}^{-1}$

Aufgabenstellung der Kovarianzselektion

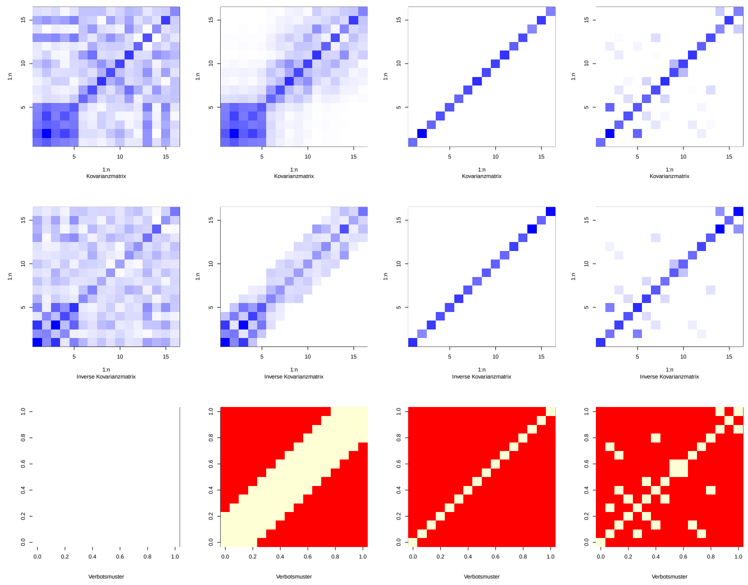
Suche eine Näherungsmatrix $\tilde{\boldsymbol{S}} \approx \hat{\boldsymbol{S}}$, deren Inverse möglich **viele Nulleinträge** aufweist!

Bedingte statistische Unabhängigkeit

Über normalverteilte Daten wissen wir, daß $C_{ij} = 0$ genau dann gilt, wenn die beiden Merkmale x_i und x_j **statistisch unabhängig** sind, sofern wir die Kenntnis der restlichen Merkmale $\{x_1, \ldots, x_D\} \setminus \{x_i, x_j\}$ voraussetzen.

Gaußsche Markovnetze

Dempsters Kovarianzselektion



Beispiel

Datensatz
`letter`
16 Merkmale
alle Klassen

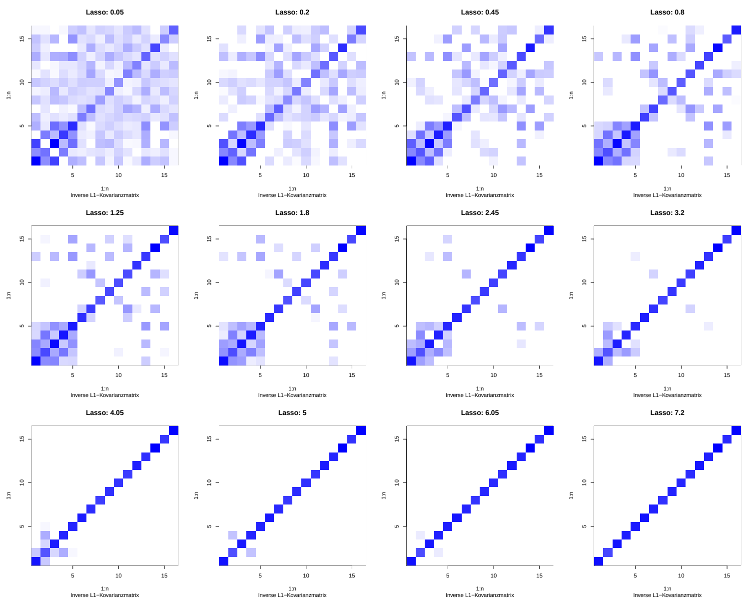
oben:
Kovarianz
$\hat{\boldsymbol{S}} = \boldsymbol{C}^{-1}$

Mitte:
Konzentration
$\boldsymbol{C}$ erfüllt $\boldsymbol{A}$

unten:
Adjazenz $\boldsymbol{A}$
Abhängigkeits-
muster
(gegeben)

Gaußsche Markovnetze

Lasso (regularisierte $\|\cdot\|_1$-Norm Matrixinvertierung)



Beispiel

Datensatz
`letter`
16 Merkmale
alle Klassen

Konzentrations-
matrizen für
unterschiedliche
**Regularisierungs-
parameter**

$$\rho \;=\; \frac{1}{20} \cdot n^2$$
$$n \;=\; 1, 2, \ldots, 12$$

Gradientenvektor und Gradientenmatrix

Extremalwertaufgabe ➪ *„Ableiten & Nullsetzen“*

Definition

Es sei

$$f : \mathbb{R}^n \to \mathbb{R} \qquad \text{bzw.} \qquad g : \mathbb{R}^{n \times m} \to \mathbb{R}$$

ein Vektor- bzw. ein Matrixfunktional. Dann heißen die Felder

$$\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix} \qquad \text{bzw.} \qquad \nabla_{\boldsymbol{Y}} f(\boldsymbol{Y}) = \begin{pmatrix} \frac{\partial g}{\partial y_{11}} & \frac{\partial g}{\partial y_{12}} & \cdots & \frac{\partial g}{\partial y_{1m}} \\ \frac{\partial g}{\partial y_{21}} & \frac{\partial g}{\partial y_{22}} & \cdots & \frac{\partial g}{\partial y_{2m}} \\ \vdots & & \ddots & \vdots \\ \frac{\partial g}{\partial y_{n1}} & \frac{\partial g}{\partial y_{n2}} & \cdots & \frac{\partial g}{\partial y_{nm}} \end{pmatrix}$$

von partiellen Ableitungen nach allen Eingangsvariablen der **Gradientenvektor** von f an der Stelle $\boldsymbol{x}$ bzw. die **Gradientenmatrix** von g an der Stelle $\boldsymbol{Y}$.

Bemerkung

Eine notwendige Bedingung für das Vorliegen eines relativen Maximums von f an der Stelle $\boldsymbol{x} \in \mathbb{R}^n$ ist das Verschwinden ($\nabla_{\boldsymbol{x}} f = \boldsymbol{0}$) des Gradientenvektors in diesem Punkt.

Gradientenvektorberechnung

Beispiel: Linearkombination

Beispielfunktional

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} a_i x_i = \boldsymbol{a}^\top \boldsymbol{x}$$

Berechnung partieller Ableitungen

$$\frac{\partial f(\boldsymbol{x})}{\partial x_k} = \frac{1}{\partial x_k} \left(\sum_{i=1}^{n} a_i x_i \right) = \sum_{i=1}^{n} \frac{\partial}{\partial x_k} (a_i x_i) = \sum_{i=1}^{n} a_i \cdot \frac{\partial}{\partial x_k} (x_i) = a_k$$

Gradientenvektor

$$\nabla_{\boldsymbol{x}} \left(\boldsymbol{a}^\top \boldsymbol{x} \right) = \boldsymbol{a}$$

Gradientenvektorberechnung

Beispiel: Quadratische Form

Beispielfunktional

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} \sum_{j=1}^{n} x_i A_{ij} x_j = \boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{x}$$

Berechnung partieller Ableitungen

$$\frac{\partial f(\boldsymbol{x})}{\partial x_k} = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\partial}{\partial x_k} (x_i A_{ij} x_j) = \sum_{i \neq k} x_i A_{ik} + \sum_{j \neq k} A_{kj} x_j + 2 x_k \cdot A_{kk} = 2 \cdot \sum_{i=1}^{n} x_i A_{ik} = 2 \cdot \boldsymbol{a}_k^\top \boldsymbol{x}$$

Gradientenvektor

$$\nabla_{\boldsymbol{x}} \left(\boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{x} \right) = 2 \cdot \boldsymbol{A} \boldsymbol{x}$$

Gradientenmatrixberechnung

Beispiel: Frobeniusnorm

Beispielfunktional

$$f(\boldsymbol{X}) = \sum_{i=1}^{n} \sum_{j=1}^{m} X_{ij}^2 = \operatorname{spur} \left(\boldsymbol{X}^\top \boldsymbol{X} \right) = \|\boldsymbol{X}\|_{\text{Frob}}^2$$

Berechnung partieller Ableitungen

$$\frac{\partial f(\boldsymbol{X})}{\partial X_{k\ell}} = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\partial}{\partial X_{k\ell}} (X_{ij}^2) = 2 \cdot X_{k\ell}$$

Gradientenmatrix

$$\nabla_{\boldsymbol{X}} \left(\|\boldsymbol{X}\|_{\text{Frob}}^2 \right) = 2 \cdot \boldsymbol{X}$$

Gradientenmatrixberechnung

Beispiel: Determinante

Beispielfunktional

$$f(\boldsymbol{X}) = \det(\boldsymbol{X})$$

Berechnung partieller Ableitungen

$$\frac{\partial f(\boldsymbol{X})}{\partial X_{k\ell}} = \frac{\partial \det(\boldsymbol{X})}{\partial X_{k\ell}} = \det(\boldsymbol{X}) \cdot \left(\boldsymbol{X}^{-1}\right)_{k\ell}$$

Gradientenmatrix

$$\nabla_{\boldsymbol{X}} \left(\det(\boldsymbol{X})\right) = \det(\boldsymbol{X}) \cdot \boldsymbol{X}^{-1}$$

Gradientenmatrixberechnung

Beispiel: Logarithmierte Determinante

Beispielfunktional

$$f(\boldsymbol{X}) = \log\det(\boldsymbol{X})$$

Berechnung partieller Ableitungen

$$\frac{\partial f(\boldsymbol{X})}{\partial X_{k\ell}} = \frac{\partial \log\det(\boldsymbol{X})}{\partial \det(\boldsymbol{X})} \cdot \frac{\partial \det(\boldsymbol{X})}{\partial X_{k\ell}} = \frac{1}{\det(\boldsymbol{X})} \cdot \det(\boldsymbol{X}) \cdot \left(\boldsymbol{X}^{-1}\right)_{k\ell} = \left(\boldsymbol{X}^{-1}\right)_{k\ell}$$

Gradientenmatrix

$$\nabla_{\boldsymbol{X}} \left(\log\det(\boldsymbol{X})\right) = \boldsymbol{X}^{-1}$$

Zusammenfassung (7)

1. Die **multivariate Normalverteilung** beschreibt eine unimodale (Zentrum $\boldsymbol{\mu}$), exponentiell abklingende Dichte mit elliptisch-symmetrischen (Trägheitsachsen von $\boldsymbol{S}$) Isolinien.
2. Die Prüfgrößen der NV-Bayesregel sind **quadratische Polynome** in den Merkmalen $x_1, \ldots, x_D$.
3. Die **Maximum-Likelihood**-Schätzung sucht die Modellparameter mit der größten Datenerzeugungswahrscheinlichkeit.
4. Die ML-Zielgröße ist nach allen Parametern **partiell abzuleiten**; nach **Nullsetzen der Gradienten** ergibt sich günstigenfalls eine **geschlossene Lösung** (LGS) oder wenigstens eine rasch konvergierende **Iterationsformel**.
5. Die **Maximum-a posteriori**-Schätzung verwendet **a priori-Wissen** über die Dichteparameter und ist **robuster** bei (zu) **kleinen Lernenstichproben**.
6. Praktikable MAP-Schätzer bedienen sich der Methode der **konjugierten Parameterdichtefamilien**.
7. Verteilungsmodelle werden robuster, wenn die **Abhängigkeitsstruktur der Merkmale** sachgemäß **ausgedünnt** wird.