Local Case-Control Sampling

basierend auf Fithian und Hastie 2014

Nils Kober

29. Juni 2021

Local Case-Control Sampling

Einführung und Motivation

- Für große Datensätze ist klassische logistische Regression (zu) rechenaufwenig.
- Ziel des Subsamplings: Verringerung des Rechenaufwands unter Beibehaltung einer möglichst hohen Effizienz des Schätzers (d.h. geringe Varianz)
- $\rightarrow$ Ermöglicht weitere statistische Verfahren (Bagging, Boosting, ...)

Grundlegende Idee

- Uniform sampling:
  - Jeder Datensatz hat gleiche Akzeptanzwahrscheinlichkeit.
  - Nicht optimiert für unausgewogene Datenmengen
- Case-Control (CC) sampling:
  - Akzeptanzwahrscheinlichkeit von (x, y) hängt nur von y ab.
  - Nutzt *marginal inbalance* aus, aber keine *conditional imbalance*.
  - Nicht konsistent, falls das Modell falsch spezifiziert ist.
- Local Case-Control (LCC) Sampling
  - Akzeptanzwahrscheinlichkeit hängt von x und y ab.
  - Behält besonders die Datensätze, bei denen das Label $Y = y$ gegeben die Features $X = x$ unerwartet ist.
  - Nutzt beide Arten der *imbalance* aus.
  - Auch konsistent, wenn das Modell falsch spezifiziert ist.

Problemstellung und Notation (1)

- Wahrscheinlichkeitsräume

$$\mathcal{X} \subset \mathbb{R}^p, \quad \mathcal{Y} = \{0, 1\} \tag{1}$$

  mit gemeinsamem Wahrscheinlichkeitsmaß $\mathbb{P}^{X,Y}$
- iid Stichprobe

$$(x_1, y_1), ..., (x_n, y_n) \in \mathcal{X} \times \mathcal{Y} \tag{2}$$

- $\mathbb{P}^X$ sei Lebesque-stetig.

Problemstellung und Notation (2)

- Bedingte Wahrscheinlichkeit & log-odds:

$$p(x) = \mathbb{P}(Y = 1 | X = x) \tag{3}$$

$$f(x) = \log\left(\frac{p(x)}{1 - p(x)}\right) \tag{4}$$

- Logistische Regressionsmodell (augmented notation):

$$f_\theta(x) = \theta^\top x \text{ mit } \theta \in \mathbb{R}^p \tag{5}$$

- Modell muss nicht korrekt spezifiziert sein (außer in Theorem 8).
- Zusätzliche Voraussetzung:

$$\nexists v \in \mathbb{R}^p : \quad \mathbb{E}\left|v^\top X\right| = 0 \tag{6}$$

Local Case-Control Sampling (LCC)

Erstschätzung (pilot estimate):

$$\tilde{p}_\lambda(x) = \frac{exp(\lambda^\top x)}{1 + exp(\lambda^\top x)} \tag{7}$$

Akzeptanzwahrscheinlichkeit:

$$a_\lambda(x, y) = |y - \tilde{p}_\lambda(x)| = \begin{cases} 1 - \tilde{p}_\lambda(x), & y = 1 \\ \tilde{p}_\lambda(x), & y = 0 \end{cases} \tag{8}$$

$$= y(1 - \tilde{p}_\lambda(x)) + (1 - y)\tilde{p}_\lambda(x) \tag{9}$$

Algorithmus:

1. Generiere $z_i \sim Ber(a_\lambda(x_i, y_i))$
2. $S \leftarrow \{(x_i, y_i) | z_i = 1\}$
3. Verwende logistische Regression auf S und erhalte Schätzung $\hat{\theta}_S$
4. Passe Schätzung an: $\hat{\theta} \leftarrow \hat{\theta}_S + \lambda$

Verteilung auf Subsample

Sei $\mathbb{P}_\lambda^{X,Y}$ die Verteilung von X und Y auf dem Subsample S.

Dann ist $\mathbb{P}_\lambda^{X,Y}$ stetig bzgl. $\mathbb{P}^{X,Y}$ mit Dichte

$$\frac{\mathrm{d}\mathbb{P}_\lambda^{X,Y}}{\mathrm{d}\mathbb{P}^{X,Y}} = \frac{a_\lambda(x,y)}{\bar{a}(\lambda)}, \tag{10}$$

wobei $\bar{a}(\lambda) = \mathbb{E}\left[a_\lambda(X,Y)\right]$

Beweis: im Vortrag

Offset in LCC

Warum $\hat{\theta} \leftarrow \hat{\theta}_S + \lambda$?

Sei

$$g(x) = \log \frac{\mathbb{P}_\lambda(Y = 1|X = x)}{\mathbb{P}_\lambda(Y = 0|X = x)} \tag{11}$$

$$= \log \frac{\mathbb{P}(Y = 1|X = x, Z = 1)}{\mathbb{P}(Y = 0|X = x, Z = 1)} \tag{12}$$

$$f(x) = \log \frac{\mathbb{P}(Y = 1|X = x)}{\mathbb{P}(Y = 0|X = x)} \tag{13}$$

Dann gilt:

$$f(x) = g(x) + \lambda^\top x \tag{14}$$

Beweis: im Vortrag

Wahl der Erstschätzung

Die Erstschätzung λ sollte konsistent sein, denn nur so ist $\widehat{\theta}$ konsistent.
Möglichkeiten:

- Generierung aus den Daten (x_i, y_i) bzw. einer Teilmenge davon
  $\rightarrow$ z.B. mit weighted case-control sampling
- Verwendung der Schätzung zu einem früheren Zeitpunkt (bei Zeitreihen)

Risikofunktionen (1)

Verlustfunktion (negative log-likelihood):

$$\rho(\theta; x, y) = -y\theta^\top x + \log(1 + \exp(\theta^\top x)) \tag{15}$$

Risikofunktion für $\mathbb{P}$:

$$R(\theta) = \int \rho(\theta; x, y) \, \mathrm{d}\mathbb{P}^{X,Y}(x, y) \qquad \text{in gesamter Stichprobe} \tag{16}$$

Risikofunktion für $\mathbb{P}_\lambda$:

$$R_\lambda(\theta) = \int \rho(\theta; x, y) \, \mathrm{d}\mathbb{P}_\lambda^{X,Y}(x, y) \qquad \text{in Subsample } S \tag{17}$$

$$Q_\lambda(\theta) = R_\lambda(\theta - \lambda) \qquad \text{in gesamter Stichprobe} \tag{18}$$

Risikofunktionen (2)

Erwartete Akzeptanzwahrscheinlichkeit:

$$\bar{a}(\lambda) = \mathbb{E}\left[a_\lambda(X, Y)\right] \tag{19}$$

Empirisches Risiko:

$$\widehat{R}_\lambda^{(0)}(\theta) = \left(\sum_{i=1}^{n} z_i\right)^{-1} \sum_{i=1}^{n} z_i \rho(\theta; x_i, y_i) \qquad \text{in Subsample } S \tag{20}$$

$$\widehat{R}_\lambda(\theta) = \frac{1}{n\bar{a}(\lambda)} \sum_{i=1}^{n} z_i \rho(\theta; x_i, y_i) \qquad \text{in Subsample } S \tag{21}$$

$$\widehat{Q}_\lambda(\theta) = \widehat{R}_\lambda(\theta - \lambda) \qquad \text{in gesamter Stichprobe} \tag{22}$$

Weitere Notation

$$\widehat{\theta} = \underset{\theta \in \mathbb{R}^p}{\operatorname{argmin}}\, \widehat{Q}_\lambda(\theta) \qquad \text{(LCC Schätzung)} \tag{23}$$

$$\bar{\theta}(\lambda) = \underset{\theta \in \mathbb{R}^p}{\operatorname{argmin}}\, Q_\lambda(\theta) \tag{24}$$

$$\theta^* = \underset{\theta \in \mathbb{R}^p}{\operatorname{argmin}}\, R(\theta) = \bar{\theta}(0) \tag{25}$$

Hauptresultate

Die folgenden Resultate gelten unter gewissen Voraussetzungen:

- Der LCC Schätzer $\widehat{\theta}_n$ ist konsistent, falls die Erstschätzung λ_n konsistent ist.
- Der LCC Schätzer (genauer: $\sqrt{n}(\widehat{\theta}_n - \bar{\theta}(\lambda_n))$ ist asymptotisch normalverteilt.
- Falls das logistische Regressionsmodell korrekt spezifiziert ist mit Parameter θ_0, ist $\sqrt{n}(\widehat{\theta}_n - \theta_0)$ asymptotisch normalverteilt mit der doppelten asymptotischen Kovarianzmatrix des MLE.

(Strikte) Konvexität

$R_\lambda(\theta)$ und $Q_\lambda(\theta)$ sind konvex in θ, da der Integrand $\rho(\theta; X, Y)$ bzw. $\rho(\theta - \lambda; X, Y)$ konvex ist (folgt mit Monotonie und Linearität des Integrals).

Für die strikte Konvexität genügt eine der folgenden Bedingungen (wird in Fithian und Hastie 2014 nicht vorausgesetzt):

- $p(x)$ hat höchstens abzählbar viele Unstetigkeitsstellen.
- $\mathbb{E} \|X\|^2 < \infty$

Nicht-Separabilität

Definition

Die Verteilung der Daten / die Klassen (X, Y) heißt nicht-separabel, falls kein $v \in \mathbb{R}^p$ existiert, so dass

$$\mathbb{P}(Y = 0, v^\top X > 0) = \mathbb{P}(Y = 1, v^\top X < 0) = 0. \tag{26}$$

$\Rightarrow$ Es gibt keine Hyperebene, welche die Klassen $Y = 1$ und $Y = 0$ fast sicher trennt.

Eindeutiges Minimum von $R_\lambda(\theta)$

Lemma (Lemma 1 in Fithian und Hastie 2014)

Angenommen die Daten seien nicht-separabel. Dann hat $R_\lambda(\theta)$ für alle $\lambda \in \mathbb{R}^p$ einen eindeutigen (globalen) Minimierer.

Beweisidee:

- $R_\lambda(\theta)$ ist strikt konvex in θ, hat also maximal einen Minimierer.
- Für die Existenz genügt es zu zeigen, dass $R_\lambda(\theta) \to \infty$ für $\|\theta\| \to \infty$ (da $R_\lambda(\theta)$ konvex und $\mathbb{R}^p$ σ-kompakt).

Proposition 2

Proposition (Proposition 2 in Fithian und Hastie 2014)

Angenommen $\mathbb{E}\,\|X\| < \infty$ *und die Daten seien nicht separabel. Sei* $\theta^* = \underset{\theta \in \mathbb{R}^p}{\operatorname{argmin}}\, R(\theta) = \bar{\theta}(0)$. *Dann gilt:*

$$\theta^* = \underset{\theta \in \mathbb{R}^p}{\operatorname{argmin}}\, Q_{\theta^*}(\theta) = \bar{\theta}(\theta^*). \tag{27}$$

Beweis: im Vortrag

Beweis Proposition 2 - Notizen

$$Q_\lambda(\theta) = \int \rho(\theta - \lambda; x, y) \, \mathrm{d}\mathbb{P}_\lambda^{X,Y}(x, y) \tag{28}$$

$$\rho(\theta - \lambda; x, y) = y(\theta - \lambda)^\top x - \log(1 + \exp((\theta - \lambda)^\top x)) \tag{29}$$

Differentiationslemma siehe Klenke 2013, Satz 6.28.

Punktweise Konvergenz

Proposition (Proposition 3 in Fithian und Hastie 2014)

Sei $\mathbb{E}\,\|X\| < \infty$ *und gelte* $\lambda_n \xrightarrow{\mathbb{P}} \lambda_\infty$. *Dann gilt für alle* $\theta \in \mathbb{R}^p$

$$\widehat{Q}_{\lambda_n}(\theta) \xrightarrow{\mathbb{P}} Q_{\lambda_\infty}(\theta). \tag{30}$$

Beweis: siehe Fithian und Hastie 2014

Gleichmäßige Konvergenz

Proposition (Proposition 4 in Fithian und Hastie 2014)

Sei $\mathbb{E}\,\|X\| < \infty$ *und gelte* $\lambda_n \xrightarrow{\mathbb{P}} \lambda_\infty$. *Weiter sei* $\Theta \subset \mathbb{R}^p$ *kompakt. Dann gilt*

$$\sup_{\theta \in \Theta} \left| \widehat{Q}_{\lambda_n}(\theta) - Q_{\lambda_\infty}(\theta) \right| \xrightarrow{\mathbb{P}} 0. \tag{31}$$

Beweisidee:

1. Zeige, dass $F_n(\theta) = \widehat{Q}_{\lambda_n}(\theta) - Q_{\lambda_\infty}(\theta)$ mit Wahrscheinlichkeit, die gegen 1 konvergiert, gleichgradig Lipschitz-stetig ist (d.h. Lipschitz-Konstante unabhängig von n).
2. Mithilfe einer endlichen offenen Überdeckung von Θ kann gleichmäßige Konvergenz in Wahrscheinlichkeit gefolgert werden.

Alternativer Beweis: Satz von Arzelá-Ascoli (benötigt etwas Vorarbeit)

Konsistenz

Theorem (Theorem 5 in Fithian und Hastie 2014)

Angenommen $\mathbb{E}\,\|X\| < \infty$ *und die Daten seien nicht separabel. Weiter gelte* $\lambda_n \xrightarrow{\mathbb{P}} \theta^*$. *Dann ist der LCC-Schätzer konsistent, d.h.*

$$\widehat{\theta}_n \xrightarrow{\mathbb{P}} \theta^*. \tag{32}$$

Beweis: im Vortrag

Beweis Theorem 5 - Notizen

$$\varepsilon := \inf_{\theta \in \partial\Theta} Q_{\theta^*}(\theta) - Q_{\theta^*}(\theta^*) > 0 \tag{I}$$

$$\mathbb{P}\left(\sup_{\theta \in \Theta} \left| \widehat{Q}_{\lambda_n}(\theta) - Q_{\theta^*}(\theta^*) \right| \right) \geq \frac{\varepsilon}{4}) \to 0 \quad \text{für } n \to \infty \tag{II}$$

$$Q_{\lambda_n}(\mu_n) - \inf_{\theta \in \partial\Theta} \widehat{Q}_{\lambda_n}(\theta) < \frac{\varepsilon}{4} \tag{III}$$

Asymptotische Normalität

Ziel: $\sqrt{n}(\widehat{\theta}_n - \bar{\theta}(\lambda_n))$ ist asymptotisch normalverteilt.

Voraussetzungen:

- Erstschätzungen $(\lambda_n)_{n\in\mathbb{N}}$ sind unabhängig von den Daten $((x_i, y_i)_{i\in\mathbb{N}})$ $\rightarrow$ z.B. Erstschätzungen durch Daten von einem früheren Zeitpunkt oder einer getrennten Stichprobe
- $E\|X\|^2 < \infty$

Asymptotische Normalität - Vorarbeit

$$J(\theta, \lambda) = \mathrm{Var}_{\mathbb{P}_\lambda} \left[\nabla_\theta - \rho(\theta - \lambda; X, Y) \right] \tag{33}$$

$$H(\theta, \lambda) = -\nabla_\theta^2 Q_\lambda(\theta) \tag{34}$$

$$\begin{aligned} &= -\bar{a}(\lambda)^{-1} \int \frac{e^{(\theta-\lambda)^\top x}}{\left(1 + e^{(\theta-\lambda)^\top x}\right)^2} \\ &\qquad \left(\frac{e^{\lambda^\top x} + e^{f(x)}}{\left(1 + e^{\lambda^\top x}\right)\left(1 + e^{f(x)}\right)} \right) x x^\top \, \mathrm{d}\mathbb{P}^X(x) \end{aligned} \tag{35}$$

$H(\theta, \lambda)$ ist positiv definit, da per Voraussetzung $\mathbb{E}\left|v^\top X\right| \neq 0$ für alle $v \in \mathbb{R}^p$.

Asymptotische Normalität

Theorem (Theorem 6 in Fithian und Hastie 2014)

Angenommen $\mathbb{E}\,\|X\|^2 < \infty$. *Sei* $(\lambda_n)_{n\in\mathbb{N}}$ *unabhängig von* $((x_i, y_i)_{i\in\mathbb{N}})$. *Falls* $\lambda_n \xrightarrow{\mathbb{P}} \theta^*$, *dann gilt*

$$\sqrt{n}(\widehat{\theta}_n - \bar{\theta}(\lambda_n)) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \bar{a}(\theta^*)^{-1}\Sigma), \tag{36}$$

wobei $\Sigma = H(\theta^*, \theta^*)^{-1} J(\theta^*, \theta^*) H(\theta^*, \theta^*)^{-1}$.

Beweis: Siehe Fithian und Hastie 2014

Aysmptotische Varianz bei korrektem Modell

Theorem (Theorem 8 in Fithian und Hastie 2014)

Angenommen das logistische Regressionsmodell sei korrekt mit Parameter $\theta_0 \in \mathbb{R}^p$, *d.h.* $f(x) = \log \frac{p(x)}{1-p(x)} = \theta_0^\top x$. *Sei* Σ_{full} *die asymptotische Varianz des ML-Schätzers für die gesamte Stichprobe (genauer:* $\sqrt{n}(\beta_n - \theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma_{full})$ *für MLE* β_n*). Gelte* $\mathbb{E}\|X\|^2 < \infty$, *sei* $(\lambda_n)_{n\in\mathbb{N}}$ *unabhängig von* $((x_i, y_i)_{i\in\mathbb{N}})$. *Falls* $\lambda_n \xrightarrow{\mathbb{P}} \theta_0$, *dann*

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 2\Sigma_{full}). \tag{37}$$

Beweis: im Vortrag

Beweis Theorem 8 - Notizen (1)

Theorem 6:

$$\sqrt{n}(\widehat{\theta}_n - \bar{\theta}(\lambda_n)) \xrightarrow{\mathcal{D}} \bar{a}(\theta^*)^{-1} H(\theta^*, \theta^*)^{-1} J(\theta^*, \theta^*) H(\theta^*, \theta^*)^{-1} \tag{I}$$

Unter Regularitätsvoraussetzungen (die hier erfüllt sind) ist ein MLE asymptotisch effizient (Czado und Schmidt 2011, 4.26. & 4.27.)

$$\Sigma_{full} = I(\theta_0)^{-1}. \tag{II}$$

Es gilt für die Fisher-Information von $\mathbb{P}_\lambda$

$$I_\lambda(\theta) = -\nabla_\theta^2 Q_\lambda(\theta). \tag{III}$$

Beweis Theorem 8 - Notizen (2)

$$J(\theta, \lambda) = \mathrm{Var}_{\mathbb{P}_\lambda}\left[\nabla_\theta - \rho(\theta - \lambda; X, Y)\right] \tag{IV}$$

$$\begin{aligned} H(\theta, \lambda) &= -\nabla_\theta^2 Q_\lambda(\theta) \\ &= -\bar{a}(\lambda)^{-1} \int \frac{e^{(\theta-\lambda)^\top x}}{\left(1 + e^{(\theta-\lambda)^\top x}\right)^2} \\ &\qquad \left(\frac{e^{\lambda^\top x} + e^{f(x)}}{\left(1 + e^{\lambda^\top x}\right)\left(1 + e^{f(x)}\right)} \right) x x^\top \, \mathrm{d}\mathbb{P}^X(x) \end{aligned} \tag{V}$$

$$H(\theta_0, 0) = \int \frac{\exp(\theta_0^\top x)}{(1 + \exp(\theta_0^\top x))^2} x x^\top \, \mathrm{d}\mathbb{P}^X(x) \tag{VI}$$

Theorem 8 - Deutung

Die asymptotische Kovarianzmatrix in Theorem 8 ist unabhängig von der Akzeptanzwahrscheinlichkeit $a_\lambda(x, y)$ bzw. $\bar{a}(\lambda)$.

Erwartete Subsample-Größe:

$$n\bar{a}(\lambda) = n\mathbb{E}a_\lambda(X, Y) = n\mathbb{E}\left[Y(1 - \tilde{p}_\lambda(X)) + (1 - Y)\tilde{p}_\lambda(X)\right] \tag{38}$$

$$= n\mathbb{E}\left[p(X)(1 - \tilde{p}_\lambda(X)) + (1 - p(X))\tilde{p}_\lambda(X)\right] \tag{39}$$

$\rightarrow$ Subsample klein bei starker *imbalance*
$\rightarrow$ Günstiges Verhältnis aus Subsample-Größe und Varianz bei starker *imbalance*

Zusammenfassung

- Der LCC Schätzer nutzt *marginal* und *conditional imbalance* in den Daten aus.
- Falls die Erstschätzung konsistent ist, so auch die LCC Schätzung (sogar wenn Modell nicht korrekt spezifiziert ist).
- Die LCC Schätzungen sind in einem gewissen Sinne aymptotisch normalverteilt.
- Im korrekt spezifizierten Fall ist die Effizienz des LCC Schätzers halb so groß wie die des MLE bei potentiell deutlich kleinerem Subsample und Rechenaufwand.

Danke für eure Aufmerksamkeit!

Quellen

- Czado, Claudia und Thorsten Schmidt (2011). *Mathematische Statistik*. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN: 978-3-642-17260-1 978-3-642-17261-8. DOI: 10.1007/978-3-642-17261-8. URL: http://link.springer.com/10.1007/978-3-642-17261-8 (besucht am 21. 06. 2021).
- Fithian, William und Trevor Hastie (Okt. 2014). „Local case-control sampling: Efficient subsampling in imbalanced data sets“. In: *The Annals of Statistics* 42.5, S. 1693–1724. ISSN: 0090-5364. DOI: 10.1214/14-AOS1220. URL: http://projecteuclid.org/euclid.aos/1410440622.
- Klenke, Achim (2013). *Wahrscheinlichkeitstheorie*. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN: 978-3-642-36017-6 978-3-642-36018-3. DOI: 10.1007/978-3-642-36018-3. URL: http://link.springer.com/10.1007/978-3-642-36018-3 (besucht am 27. 06. 2021).