Linear classification with missing values

Internship Final Presentation

Angel REYERO

M1 Mathematics & Artificial Intelligence
Institut de Mathématiques d'Orsay (IMO)
Paris-Saclay University

Under the supervision of:
Alexis AYME, Claire BOYER, Aymeric DIEULEVEUT and Erwan SCORNET.

July 20, 2023

Contents

Index

Motivation

Trauma.center	Heart rate	Death	Anticoagulant. therapy	Glascow score
Pitie-Salpêtrière	88	0	No	3
Beaujon	103	0	NA	5
Bicêtre	NA	0	Yes	6
Bicêtre	NA	0	No	NA
Lille	62	0	Yes	6
Lille	NA	0	No	NA
	⋮	⋮	⋮	⋮

Different sources of missing values (`Not Available(NA)`):

- Bugs/ Sensors failures
- Costs
- Sensitive data
- Data merging
- No time to measure in an emergency situation

Notation

Let an observation with missing values $(X_{\mathrm{obs}(M)}, M, Y)$ be:

- Missing value pattern $M \in \{0,1\}^d$ such that

$$M_j = 1 \iff X_j \text{ is missing.}$$

- $\mathrm{obs}(M) := \{j \in \{1,...d\} | M_j = 0\}$.
- $X_{\mathrm{obs}(M)}$ observed covariates.
- $Y \in \{-1,1\}$ the label (*always* observed).

Example:

$$\begin{aligned} X &= (6,3,\mathtt{NA},3,\mathtt{NA}), \\ M &= (0,0,\ 1,0,\ 1), \\ \mathrm{obs}(M) &= (1,2,\quad 4\quad), \\ X_{\mathrm{obs}(M)} &= (6,3,\quad 3\quad). \end{aligned}$$

Supervised learning with missing values: Classification

- Complete data case
  - Dataset: $\mathscr{D}_n = \{(X_i, Y_i), i \in \{1, \dots n\}\}$
  - Misclassification probability:

$$\mathscr{L}_{\mathrm{comp}}\left(\widehat{h}_{\mathrm{comp}}\right) := \mathbb{P}\left(\widehat{h}_{\mathrm{comp}}(X) \neq Y\right).$$

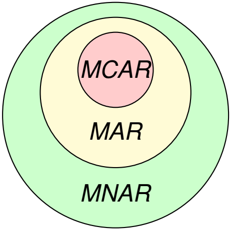
- Incomplete data case
  - Dataset: $\mathscr{D}_n = \{(X_{i,\mathrm{obs}(M_i)}, M_i, Y_i), i \in \{1, \dots n\}\}$
  - Misclassification probability:

$$\mathscr{L}\left(\widehat{h}\right) := \mathbb{P}\left(\widehat{h}(X_{\mathrm{obs}(M)}, M) \neq Y\right).$$

Missing values mechanism

Assumptions on $M|X, Y$ Rubin [1976]:

- **MCAR** (Missing completely at random).
$$M \perp\!\!\!\perp X, Y.$$
- **MAR** (Missing at random).
$$\forall m \in \{0,1\}^d,$$
$$\mathbb{P}(M = m|X, Y) = \mathbb{P}\left(M = m|X_{obs(m)}\right).$$
- **MNAR** (Missing not at random). M depends on the full vectors X and Y.



Index

Pattern-by-pattern Bayes predictor

Proposition

The Bayes predictor is given by

$$h^\star(X_{\mathrm{obs}(M)}, M) := \mathrm{sign}(\mathbb{E}\,[\,Y|X_{\mathrm{obs}(M)}, M\,]).$$

Following the same idea as Ayme et al. [2022], we can decompose it pattern-by-pattern as

$$h^\star(Z) = h^\star(X_{\mathrm{obs}(M)}, M) = \sum_{m\in\{0,1\}^d} h^\star_m(X_{\mathrm{obs}(m)})\mathbb{1}_{M=m}$$

with

$$h^\star_m(X_{\mathrm{obs}(m)}) := \mathrm{sign}(\mathbb{E}\,[\,Y|X_{\mathrm{obs}(m)}, M=m\,]).$$

Problematic: prediction vs model inference

- Dempster et al. [1977]: EM algorithm to compute MLE from incomplete data.
- (!) Missing values in the training set <u>and</u> in the test set
- ✗ Estimating the underlying model does not help for prediction purposes.

$$\mathbb{E}[Y|X] = f_{\beta}(X) \qquad \not\Rightarrow \qquad \widehat{Y} \neq f_{\widehat{\beta}}(X_{\mathrm{obs}(M)})$$

- We need to design predictors handling missing entries.
- ✓ Decompose predictors specifically to the missing patterns.
- (!) The pattern-by-pattern Bayes classifier may not conserve the model structure on the observed covariates.

$$\mathbb{E}[Y|X] = f_{\beta}(X) \qquad \overset{?}{\Rightarrow} \qquad \mathbb{E}[Y|X_{\mathrm{obs}}, M] = f_{\beta'}(X_{\mathrm{obs}(M)}, M)$$

  - ✓ (Linear model) Morvan et al. [2020]
  - ✗ (Logistic model) This work.

Index

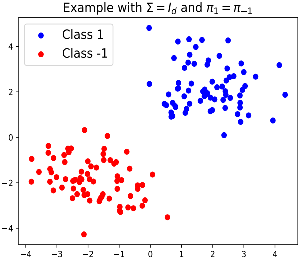
LDA in the complete data case

Assumption (LDA): $X|Y = k \sim \mathcal{N}(\mu_k, \Sigma)$.
Notation $\pi_k := \mathbb{P}(Y = k)$.

Proposition

The Bayes predictor reads as

$$h^\star(x) = \operatorname{sign}\left((\mu_1 - \mu_{-1})^\top \Sigma^{-1} \left(x - \frac{\mu_1 + \mu_{-1}}{2} \right) - \log\left(\frac{\pi_{-1}}{\pi_1} \right) \right). \tag{1}$$



LDA in the complete data case vs LDA in the missing data case

Proposition [Internship contribution]

Under the **LDA** model with **MCAR** inputs, the pattern-by-pattern Bayes classifier is given by

$$h_m^\star(x_{\mathrm{obs}(m)}) = \operatorname{sign}\left(\left(\mu_{1,\mathrm{obs}(m)} - \mu_{-1,\mathrm{obs}(m)}\right)^\top \Sigma_{\mathrm{obs}(m)}^{-1} \left(x_{\mathrm{obs}(m)} - \frac{\mu_{1,\mathrm{obs}(m)} + \mu_{-1,\mathrm{obs}(m)}}{2} \right) - \log\left(\frac{\pi_{-1}}{\pi_1}\right)\right)$$

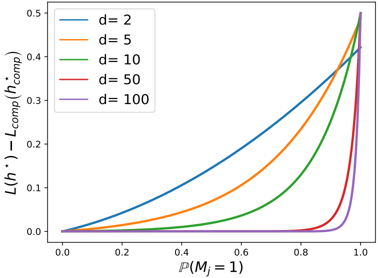
They are the projected parameters!

Approximation error introduced by missing data

The objective is to establish an upper bound for $\mathbf{L}(\mathbf{h}^\star) - \mathbf{L}_{\text{comp}}(\mathbf{h}^\star_{\text{comp}})$.

Assumptions made:

- Balanced classes $(\pi_1 = \pi_{-1} = 1/2)$.
- $\forall j \in \{1, \dots d\}, \eta_j := \mathbb{P}(M_j = 1) = \eta$.
- $\forall j \in \{1, \dots d\}, (\mu_1 - \mu_{-1})_j = \pm \mu$.
- MCAR.



Observation

Exponential decay of the **approximation error** introduced by the missing values with d!

Bound on $L(h^\star) - L_{\mathrm{comp}}(h^\star_{\mathrm{comp}})$

Denote $\lambda := \mu/\sqrt{\lambda_{\max}(\Sigma)}$ the signal-to-noise ratio (SNR).

Proposition [Internship contribution]

Under the previous assumptions, we have that

$$L(h^\star) - L_{\mathrm{comp}}(h^\star_{\mathrm{comp}}) \lesssim \mu \sqrt{\frac{d}{\lambda_{\min}(\Sigma)}} \epsilon(\eta, \lambda)^d,$$

with $\epsilon(\eta, \lambda) := \eta + e^{-\frac{\lambda^2}{8}}(1 - \eta) < 1$.

Observation

Exponential decay of the bound with d.

Bound accuracy on the approximation error

$$L(h^\star) - L_{\text{comp}}(h^\star_{\text{comp}}) \lesssim \mu \sqrt{\frac{d}{\lambda_{\min}(\Sigma)}} \epsilon(\eta, \lambda)^d.$$

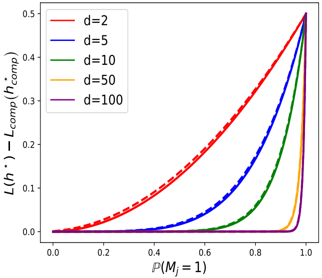


Figure 1: The tightness of the provided upper bound. Continuous lines represent the true difference, while the dashed lines represent the established bound divided by a constant.

LDA estimation

Given a dataset $\mathscr{D}_n^\star = (X_{\mathrm{obs}(M_i),i}, M_i, Y_i)_{i=1,\dots n}$, the objective is to estimate the parameters of our p-b-p LDA. Suppose the Σ known and $\pi_1 = \pi_{-1} = 1/2$.

Definition

For each class $k \in \{-1, 1\}$ and $j \in [d]$,

$$\widehat{\mu}_{k,j} := \frac{\sum_{i=1}^n X_{i,j} \mathbb{1}_{Y_i=k} \mathbb{1}_{M_{i,j}=0}}{\sum_{i=1}^n \mathbb{1}_{Y_i=k} \mathbb{1}_{M_{i,j}=0}} = \frac{\sum_{i=1}^n (X_i \odot (1 - M_i))_j \mathbb{1}_{Y_i=k} \mathbb{1}_{M_{i,j}=0}}{\sum_{i=1}^n \mathbb{1}_{Y_i=k} \mathbb{1}_{M_{i,j}=0}}. \tag{2}$$

Denote $\widehat{h}$ as the p-b-p classifier estimated by (2).

LDA estimation convergence rate

Denote $\rho := \max_{i\in[n]} \Sigma_{i,i}/\lambda_{\min}(\Sigma)$ the greatest value of the diagonal of the covariance over its smallest eigenvalue.
The objective is to establish an upper bound on the **estimation error** $(\mathbf{L}(\widehat{\mathbf{h}}) - \mathbf{L}(\mathbf{h}^\star))$.

Theorem [Internship contribution]

Under **LDA** model with **MCAR** inputs, we have that for a n large enough

$$L(\widehat{h}) - L(h^\star) \lesssim \sqrt{\rho d/n}$$

- Observe that if $\Sigma = \sigma^2 I_d$ then $\rho = 1$ and $L(\widehat{h}) - L(h^\star) \lesssim \sqrt{d/n}$.
- (!) Not an informative bound in high dimensions! ($d \gg n$)

LDA estimation error

Observation

Under **LDA** model with **MCAR** and $\Sigma = \sigma^2 I_d$ inputs:

- *The estimation error* $(L(\widehat{h}) - L(h^\star))$ is of the order of $\sqrt{d/n}$.
- The *approximation error* due to the missing values $(L(h^\star) - L_{\mathrm{comp}}(h^\star_{\mathrm{comp}}))$ is of the order of $\lambda\eta\sqrt{d}\epsilon(\eta,\lambda)^d$.

Then, for d verifying

$$-\frac{\log(\sqrt{n}\lambda\eta)}{\log(\epsilon(\eta,\lambda))} \lesssim d,$$

the error introduced by the missing values is negligible compared with the estimation error.

Approximation error $\lesssim$ *Estimation error*

Estimates for sparsity assumptions

Assumption $\Sigma = \sigma^2 I_d$.
Assumption (Sparsity) $\mathrm{card}\left(\{j \in [d], \mu_{1,j} - \mu_{-1,j} \neq 0\}\right) = s \ll d$.

Definition

Given a dataset $\mathcal{D}_n^\star$, we estimate the mean as

$$\widetilde{\mu}_{k,j} := \widehat{\mu}_{k,j} \mathbb{1}_{\widehat{\mu}_{k,j} > \tau_{k,j}}, \qquad \text{where} \qquad \tau_{k,j} := 2\sigma \sqrt{\frac{\log(d)}{N_{k,j}}}, \tag{3}$$

$\widehat{\mu}_{k,j}$ is defined in (2) and $N_{k,j} := \sum_{i=1}^n \mathbb{1}_{Y_i = k} \mathbb{1}_{M_{i,j} = 0}$.

Notation: let $\widetilde{h}$ be the Bayes classifier estimated by (3).

Observation

More confidence for the coordinates that have been observed more frequently!

LDA estimation under sparsity Assumptions

We can mitigate the curse of dimensionality with sparsity:

Theorem [Internship contribution]

Under the previous assumptions, for n large enough, we have that

$$L(\widetilde{h}) - L(h^\star) \lesssim \sqrt{s \log(d)/n}.$$

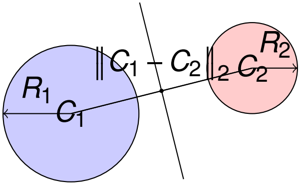
Index

Data modeling

Consider two separated balls B_1 and B_2 resp. centered at C_1, C_2 and of respective radius R_1, R_2 in the normed vector space $(\mathbb{R}^d, \|\cdot\|_p)$ with $p > 0$.

Assumptions made:

- $(C_1 - C_2)_j$ i.i.d.
- $(R_1|(C_1, C_2), R_2|(C_1, C_2))$ $\sim \mathscr{U}(0, \frac{1}{2}\left\|C_1 - C_2\right\|_p)^{\otimes 2}$.
- $M \sim \mathscr{U}(\{m \in \{0,1\}^d, \|m\|_0 = s\})$.



Linear separability

Observation

To ensure the convergence of the *p-b-p perceptron*, we need the linear separability.

Asymptotic separability of two balls with the same radius:

Proposition [Internship contribution]

Under the previous assumptions and $R_2 = R_1$, then,

$$\lim_{d \to +\infty} \mathbb{P}\left(B_{1,\mathrm{obs}(M)} \cap B_{2,\mathrm{obs}(M)} = \emptyset\right) = \sqrt[p]{1-\gamma}, \tag{4}$$

with $\gamma := \lim_{d\to\infty} s/d$.

Observation

More separability with norms of higher order!

Linear separability

Finite distance separability of two Euclidean balls with the different radius:

Proposition [Internship contribution]

Given two fixed centers c_1 and c_2, $R_1, R_2 \sim \mathscr{U}(0, \frac{1}{2}\|c_1 - c_2\|_p)^{\otimes 2}$, $(R_1 \perp\!\!\!\perp R_2)$, $(M \perp\!\!\!\perp R_1, R_2)$ and $\mathbb{P}(M_j = 1) = \eta$, then

$$\mathbb{P}\left(B_{1,obs(M)} \cap B_{2,obs(M)} = \emptyset\right) \geq 1 - \eta.$$

- (!) This result is worse than the previous result as we loose the square root, but it is a finite distance result.

Index

Take-home message

Scarcity of methods for prediction with missing values

$\Rightarrow$ **p-b-p decomposition**

- On the **LDA**(with MCAR):
  - It accepts p-b-p decomposition
  - With d large enough the missing values error is negligible
  - Estimated with a rate convergence of (with $\Sigma = \sigma^2 I_d$):
    - $\sqrt{d/n}$
    - $\sqrt{s \log(d)/n}$ (with sparsity assumption)
- On the **logistic regression**:
  - It does not accept p-b-p decomposition under very general assumptions
- On the **perceptron**:
  - P-b-p linear separability is ensured with a high probability if there is a small probability of being missing.

Index

References

Alexis Ayme, Claire Boyer, Aymeric Dieuleveut, and Erwan Scornet. Near-optimal rate of consistency for linear models with missing values. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 1211–1243. PMLR, 17–23 Jul 2022. URL `https://proceedings.mlr.press/v162/ayme22a.html`.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977. ISSN 00359246. URL `http://www.jstor.org/stable/2984875`.

References

Marine Le Morvan, Nicolas Prost, Julie Josse, Erwan Scornet, and Gael Varoquaux. Linear predictor on linearly-generated data with missing values: non consistency and solutions. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 3165–3174. PMLR, 26–28 Aug 2020. URL `https://proceedings.mlr.press/v108/morvan20a.html`.

Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3): 581–592, 1976. ISSN 00063444. URL `http://www.jstor.org/stable/2335739`.

References

T. Tony Cai and Linjun Zhang. High Dimensional Linear Discriminant Analysis: Optimality, Adaptive Algorithm and Missing Data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81 (4):675–705, 06 2019. ISSN 1369-7412. doi: 10.1111/rssb.12326. URL `https://doi.org/10.1111/rssb.12326`.

Thank You, Questions?

Index

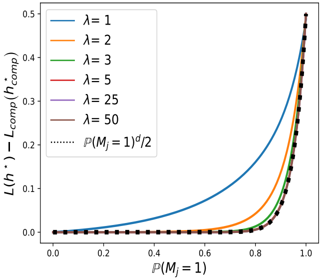
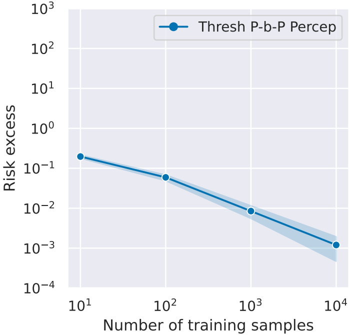


Figure 2: Convergence of the error introduced by the missing values as the signal-to-noise ratio explodes.

P-b-p perceptron experiences



(a) MCAR $d = 5$

(b) MCAR $d = 10$

Figure 3: Excess risk w.r.t. the number of training samples. The curve represents the averaged excess risk over 100 repetitions within a 95% confidence interval.