Confidence sets for binary classification problems

Author:
Noémi Takács

Supervisor:
Ambrus Tamás
SZTAKI, ELTE

Math Project ELTE, May 2024

Table of contents

Topic of the project

Previous semester:

- confidence intervals for mean estimates
- preparation for binary classification

Aim of the second semester:

- estimate the regression function (f_*) in binary classification
- construct confidence sets around the estimation

Binary classification

It is given an i.i.d. sample: $(X, Y) \in \mathbb{X} \times \mathbb{Y}$, where $\mathbb{X} \subseteq \mathbb{R}^d$ and $\mathbb{Y} = \{1, -1\}$.

Definition

The (measurable) $g : \mathbb{X} \to \mathbb{Y}$ classifier is Bayes optimal, if it minimizes the Bayes risk $R(g) \doteq E[L(Y, g(X))]$, where L is a nonnegative measurable loss function.

If $L(Y, g(X)) = \mathbb{I}(Y \neq g(X))$, where $\mathbb{I}$ is the indicator function, then the Bayes optimal classifier will be the sign of the regression function,

$$f_*(x) \doteq E[Y|X = x] = 2 \cdot P(Y = 1|X = x) - 1.$$

$$\boxed{P(Y = 1|X = x) = \frac{P(Y = 1) \cdot f_{X|Y=1}(x)}{P(Y = 1) \cdot f_{X|Y=1}(x) + P(Y = -1) \cdot f_{X|Y=-1}(x)}.}$$

Generalization of SPS method

- → distribution-free
- → non-asymptotic

Assumptions:

- $\mathbb{X} \subseteq \mathbb{R}^d$ and the $\{(X_j, Y_j)\}_{j=1}^n$ sample is i.i.d.;
- for the regression function a parameterised family $\mathcal{F}$ is given, which contains f_*;
- the parameterisation is injective, such that for all $\theta_1 \neq \theta_2 \in \Theta$:

$$\|f_{\theta_1} - f_{\theta_2}\|_P^2 \doteq \int_{\mathbb{X}} (f_{\theta_1}(x) - f_{\theta_2}(x))^2 \mathrm{d}P_X(x) \neq 0.$$

Main idea:

- for a given θ generate $m-1$ alternative samples $\mathcal{D}_i(\theta)$
- compare the original $\mathcal{D}_0$ and the alternative samples with a ranking function
- construct the confidence set based on the rank of $\mathcal{D}_0$

The confidence set

Theorem

Under the mentioned, mild statistical assumptions, the confidence set contains f_* with exact user-chosen inclusion probability.

Remark

1. If $\theta = \theta^*$, then $\mathcal{D}_0$ and $\mathcal{D}_i(\theta^*)$ comes from the same distribution.
2. If $\theta \neq \theta^*$, then the distribution of $\mathcal{D}_i(\theta)$ differs from that $\mathcal{D}_0$.

Hypothesis testing:

$$H_0 : f_* = f_\theta$$
$$H_1 : f_* \neq f_\theta$$

Acceptance region: the confidence set

Generated mixed Laplace distributions I.

Parameters used for the generations:

- $P(Y=1) = P(Y=-1) = 0.5$
- $\mu = Y$
- $\lambda = 1$

Tested parameters: $p = P(Y=1)$ and λ. Fixed location parameters.

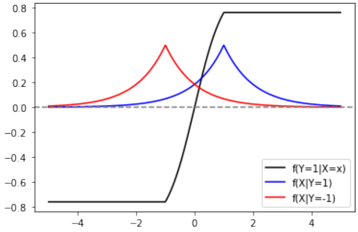

Figure: The two Laplace PDF and the real regression function

First experiment:

- $\mu_1 = 1$, $\mu_2 = -1$ (as known)
- $n = 500$
- $m = 20 \to$ confidence levels: $5\%, \ldots, 95\%, 100\%$

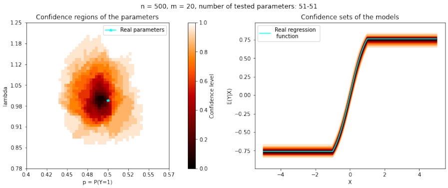

Figure: Different level confidence sets in the parameter and model space

Generated mixed Laplace distributions II.

Second experiment:

- comparison of the known values and the ML estimates of the location parameters
- $n = 20, 60, 100, 200, 300$
- $m = 10 \rightarrow$ confidence levels: $10\%, \dots, 90\%, 100\%$
- repeated 5 times
- calculate the mean of the ranks for all tested parameter pairs

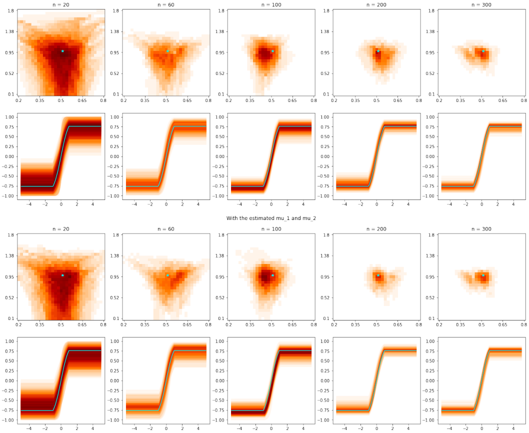

Figure: Mean of the ranks in the parameter and the model space

Bank churn

- $\mathbb{Y} = \{1, 0\}$
- P(leaving the bank) $\sim$ age
- two model class: using normal and lognormal distributions

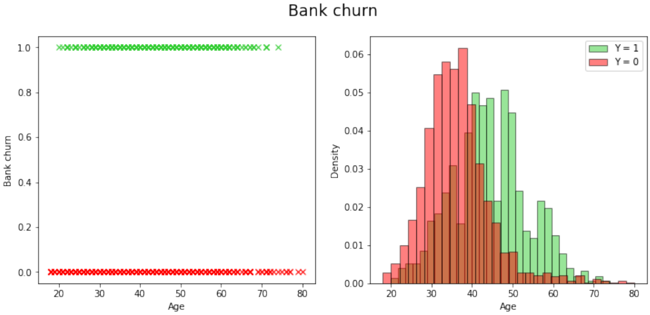

Figure: Illustrations of the observations

Compare model classes

- tested parameters: p and σ_2^2
- fixed parameters: μ_1, μ_2 and σ_1^2 with ML estimates
- $n = 4000$
- $m = 10 \rightarrow$ confidence levels: $10\%, \ldots, 90\%, 100\%$

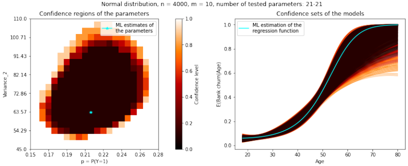

Figure: Results of using normal distributions

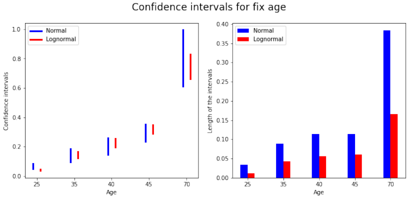

Figure: Confidence intervals and their length for the two model classes

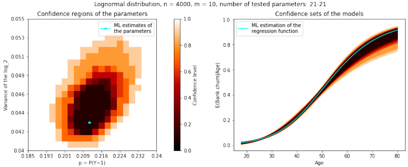

Figure: Results of using lognormal distributions

Monte Carlo sampling

- $n = 200$
- multiple sampling (5 times)
- using lognormal distributions
- tested parameters:
  - p and σ_2^2
  - p and μ_2
- $m = 10 \rightarrow$ confidence levels: $10\%, \ldots, 90\%, 100\%$

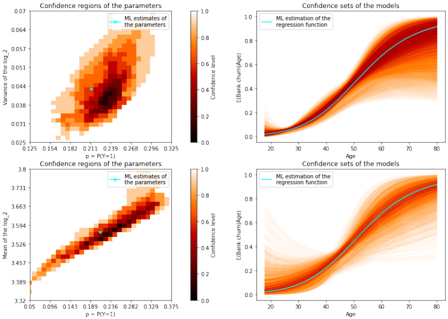

Figure: Mean of the ranks: σ_2^2 vs. μ_2

Further plans

- Multivariate cases (higher dimension, more parameters)
- Represent the results
- Extract confidence intervals from the sets

References

- Tamás, A., Csáji, B. Cs. (2020). Szto-chasztikus garanciák bináris klasszifikációhoz. **Alkalmazott matematikai lapok**, 37, 365–379.
- Tamás, A., Csáji, B. Cs. (2022). Exact Distribution-Free Hypothesis Tests for the Regression Function of Binary Classification via Conditional Kernel Mean Embeddings. **IEEE CONTROL SYSTEMS LETTERS**, VOL. 6, 860 - 865
- willian oliveira gibin, and SIMARPREET SINGH. accessed: 04.22.2024. (2024). Bank Churn Prediction [Data set]. **Kaggle** https://doi.org/10.34740/KAGGLE/DSV/7466166

Thank you for your attention!