SSY316 - Python 2

Fengyuan Liu Pär Aronsson

November 2023

1 Logistic Regression via Maximum Likelihood

The first part of this assignment consisted of learning about two different logistic regression models, the first finding the maximum likelihood with the help of the Newton-Raphson algorithm and another finding the MAP using Laplace approximation, we will discuss the two in the subsections below.

1.1 NewtonRaphsonLogistReg

The following function was given to us in equation 1.1 which defines the logistic function σ. We are also given the negative loss function which also plays a part when finding the optimal with Newton-Raphson which can be seen in equation 1.1.

The general idea of the Newton-Raphson algorithm is to try and find the optimal slope of a function by starting at a random position for the root and then literately finding a better approximation based on the coefficients at that position. This algorithm then tries to find a better root until a threshold has been reached or if we have reached the number of iterations.

Logistic Regression model The logistic regression model uses the logistic function $\sigma(\cdot) : (-\infty, \infty) \to (0, 1)$ (also known as sigmoid)

$$\sigma(x) = \left(\frac{1}{1 + e^{-x}} \right)$$

Note that the inverse of the logistic function $\sigma(\cdot)^{-1}$, in the sense of their composition equaling the identity map, is the logit function

$$\text{logit}(a) = \log \left(\frac{a}{1 - a} \right)$$

Specifically, in the context of logistic regression, the function we need is the following

$$\sigma(x, \beta) = \left(\frac{1}{1 + e^{-x\beta}} \right)$$

Negative loss function

$$\log p(y|\beta, x) = \sum_{i=1}^{n} (y_i \log f_i + (1 - y_i) \log(1 - f_i))$$

1.2 NewtonRaphsonBayesLogReg

The next method was using Laplace approximation which works similarly though using Laplace approximation we now focus on finding the best distribution as to draw our variables from rather than finding the variables themselves. The posteriors and prior were given to us in equation 1.2.

The prior and Posterior given for Laplace approximation

$$\beta \sim \mathcal{N}(m_0, S_0^{-1}) \quad \text{(prior)}$$

$$p(\beta|x, y) \sim \mathcal{N}(\beta, S^{-1}) \quad \text{(posterior)}$$

$$\beta_{\text{MAP}} = \arg\min_{\beta} \log p(\beta|y, x), \quad \log p(\beta|y, x) = \log p(y|\beta, x) + \log p(\beta)$$

$$S = -\nabla^2 \log p(\beta|y, x)\Big|_{\beta=\beta_{\text{MAP}}}$$

1.3 Activity 1

The task for activity 1 was to compute the outputs of both these models given that we also take "year" into consideration and the answers we got can be seen below:

	intercept	age	nodes_detected	year
NewtonRaphsonLogistReg	2.980	0.0121	-0.100	-0.035
NewtonRaphsonBayesLogReg	2.869	0.012	-0.097	-0.033

Table 1: Results from activity 1

2 Evaluate predictive performance for Logistic regression

The next part of the assignment evaluated the predictive performance and the metrics considered were: Area under curve (AUC), Log evidence, Bayesian information criterion (BIC), and Negative log-likelihood. These metrics give insights as to how well a model fits its data and how well it can predict new data.

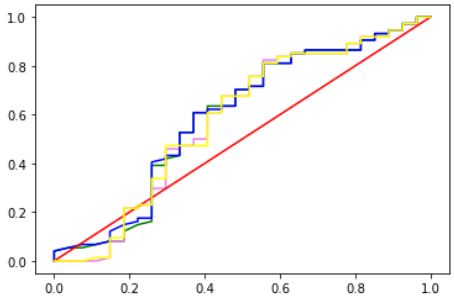
2.1 Activity 2

These metrics were then used for activity 2 where we wanted to compare two models, one using "age" and "nodes_detected" named m1 and another also taking "year" into consideration named m2.

The results are displayed in the table and figure below. What we gathered from this activity was that the model using more predictors scored slightly better which could mean that "year" provides some valuable information.

	BIC	Area under ROC	Monte Carlo
m1	-116.787	0.590	0.582
m2	-119.227	0.581	0.582

Table 2: Caption

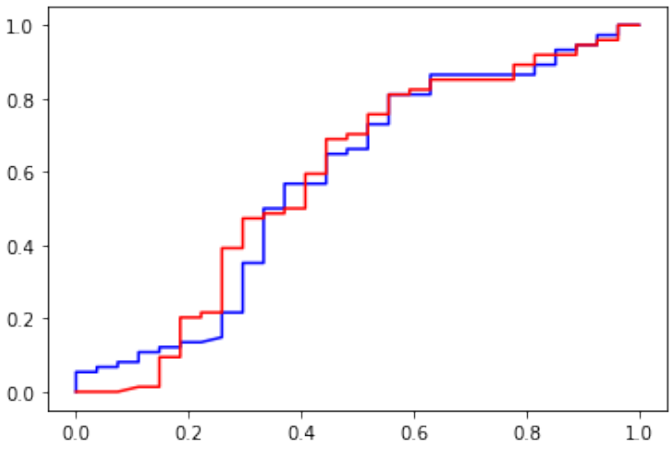


2.2 Activity 3

The dataset was divided into a training set and a test set with a 2:1 ratio using a random state to ensure reproducibility. The response variable, survival status, was converted to a binary format for compatibility with the predictive models.

Table 3: Logistic Regression Model Fitting Iterations

Iteration	Negative Log Likelihood	Difference
1	142.09517021748883	–
2	109.65341319399893	32.4417570234899
3	108.97450892987667	0.67890426412226
4	108.95849230528565	0.01601662459102
5	108.9572879760047	0.00120432928095



The logistic regression model's ROC curve (blue line) and the LDA model's ROC curve (red line) were plotted on the same graph. Both models demonstrated a degree of overlap in their ROC curves, suggesting comparable performance in distinguishing between the classes. The curves' progression toward the top left corner of the plot suggests a reasonable level of discriminative ability for both models.

Discussion:

Upon examining the performance metrics and iterative convergence of the two logistic regression models, it is evident that the inclusion of additional predictors does not always equate to an enhanced model. Model M1, which utilizes a more parsimonious set of predictors (age and nodes detected), outperforms Model M2 (which also includes year) in terms of log-evidence and the Bayesian Information Criterion (BIC). This suggests that M2, despite its complexity, does not significantly improve predictive power, which can be an indication of overfitting, especially when the additional covariate (year) does not contribute substantial explanatory power to the model.

The Maximum Likelihood Estimation (MLE) approach in these models aims to find point estimates that maximize the likelihood function, which can be powerful for predictions but may overlook the uncertainty inherent in the parameter estimates. In contrast, our Bayesian approach offers a broader perspective by seeking the entire posterior distribution of parameters. This method acknowledges the uncertainty and variability in the estimates, potentially providing a more robust and interpretable model, especially in the presence of prior knowledge or when dealing with small datasets.

The Receiver Operating Characteristic (ROC) curves further illustrate the models' classification capabilities. While both models demonstrate reasonable discriminative ability, as shown by the area under the curve (AUC), the marginal improvement from M1 to M2 does not justify the complexity added by an extra covariate. This is a classic example of the principle of parsimony, where simpler models are preferred unless more complex models provide a substantially better fit.

In conclusion, our analysis underscores the importance of model selection criteria that balance fit and complexity, such as BIC, and the value of Bayesian methods for capturing the uncertainty in parameter estimation. Future work may explore the impact of different priors or the inclusion of interaction terms between covariates to potentially improve model performance without unnecessarily increasing complexity.