Prediction of the attrition for 1470 IBM employees

Anne-Sophie Van de Velde

Ecole Polytechnique Fédérale de Lausanne

245951

June 17, 2019

Overview

- Objective of the study and basic notation.

  We want to **predict if an employee is feeling an attrition** given a set of explanatory variables. The employees that feel an attrition are in the class $\mathcal{C}_1$ with associated target variable $t = 1$. Otherwise, they are in class $\mathcal{C}_2$ and $t = 0$. Thus our task is to predict a binary response and we will always model the **likelihood** of the data given the parameters as:

$$p(X, \mathbf{t}|\omega) = \prod_{i=1}^{N} f(\omega^T x_n)^{t_n} \times \{1 - f(\omega^T x_n)\}^{1-t_n}. \quad (1)$$

  Here we used: N the number of observations; X is matrix with line $x_n^T \in \mathbb{R}^d$, the vector of explanatory variables associated to the n^{th} oberservation; $\mathbf{t} = (t_1, \ldots, t_n)^T \in \mathbb{R}^d$ the vector of responses (*i.e.* $t_n \in \{0, 1\}$); $\omega \in \mathbb{R}^d$ the **random parameter** we want to estimate; $f : \mathbb{R} \to \mathbb{R}$ is the **activation function**.

- We study three different models here, namely:
  1. Gaussian prior and logistic activation function
  2. Gaussian prior and inverse probit activation function
  3. Student-t prior and logistic activation function

- We want to find a satisfactory approximation method.
  1. Laplace approximation
  2. Metropolis-Hastings random walk
  3. Metropolis-Adjusted Langevin

Dataset presentation

	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeNumber	...
0	41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sciences	1	1	...
1	49	No	Travel_Frequently	279	Research & Development	8	1	Life Sciences	1	2	...
2	37	Yes	Travel_Rarely	1373	Research & Development	2	2	Other	1	4	...
3	33	No	Travel_Frequently	1392	Research & Development	3	4	Life Sciences	1	5	...
4	27	No	Travel_Rarely	591	Research & Development	2	1	Medical	1	7	...
5	32	No	Travel_Frequently	1005	Research & Development	2	2	Life Sciences	1	8	...
6	59	No	Travel_Rarely	1324	Research & Development	3	3	Medical	1	10	...

- 1470 observations, 35 covariates (categorical and continuous) per observation.
- The target variable is the Attrition column
- After dropping highly correlated features and doing some categorical encoding, we arrive at a feature matrix of dimension $(1470, 39)$ and a target variable of size $(1470, 1)$.
- We separate our data into a train set to train our models and a test set to test their performance.

We implement a **Bayesian logistic regression** using a Laplace approximation for the posterior distribution of the fitted parameters ω.

- The **activation function** is the sigmoid function, *i.e.* the function f in (1) is

$$f(a) = \sigma(a) = \frac{\exp(a)}{1 + \exp(a)}. \tag{2}$$

- Since we seek for a Gaussian approximation of the posterior, it is natural to take a **multivariate Gaussian prior** of the parameters ω, **i.e.**

$$\omega \sim \mathcal{N}(m_0, S_0). \tag{3}$$

- Thus, using (1), the **unnormalized log posterior** is:

$$\begin{aligned} \log p(\omega|X, \mathbf{t}) = &- \frac{1}{2}(\omega - m_0)^T S_0^{-1}(\omega - m_0) \\ &+ \sum_{n=1}^{N} \{t_n \log \sigma(\omega^T x_n) + (1 - t_n) \log(1 - \sigma(\omega^T x_n))\} + \text{const.} \end{aligned} \tag{4}$$

  We work with the logarithm posterior to avoid underflow.

- For a Gaussian approximation of the unnormalized posterior, we need
  1. the **MAP** solution ω_{MAP} of (4), found with a **Gardient Descent** algorithm;
  2. then the covariance function of the approximation is given by

$$S_N^{-1} = -\partial_\omega^2 \ln p(\omega_{MAP}|X, \mathbf{t}) = S_0^{-1} + \sum_{n=1}^{N} \sigma(\omega^T x_n)(1 - \sigma(\omega_{MAP}^T x_n)) x_n x_n^T. \tag{5}$$

- The our normalized posterior will be approximate by

$$q(\omega) = \mathcal{N}(\omega|\omega_{MAP}, S_N). \tag{6}$$

Gradient Descent algorithm

Theoretical algorithm

Input ω_0

For $n = 0, \ldots, M$:

1. Compute $\nabla f(\omega_i)$
2. Compute $\omega_{n+1} = \omega_n - \gamma \times \nabla f(\omega_i)$
3. Stop if $f(\omega_n) - f(\omega_{n+1}) < \epsilon$

Output: ω_{n+1}

In practice

- $\omega_0 = 0 \in \mathbb{R}^{39}$
- $f(\omega) = -\log p(\omega|X, \mathbf{t})$, with $m_0 = 0 \in \mathbb{R}^{39}$, $S_0 = 9 \times I \in \mathbb{R}^{39\times39}$
- $\frac{\partial f(\omega)}{\partial \omega_n} = \omega_n + (\sigma(\omega^T x_n) - t_n)x_n$
- $\gamma = 0,0001$
- $\epsilon = 0,0001$
- $M = 10000$

Results of the algorithm

- Computational time: 42 minutes, stopping after 7601 iterations (≈ 0.3 sec/iteration).
- Biggest coefficients:
    1. OverTime_Yes ≈ 1.16,
    2. BusinessTravel_Travel_Frequently ≈ 0.97

Prediction

Prediction: theoretical aspects

The predictive distribution for $\mathcal{C}_1$ given a new feature vector x is obtained by **marginalization**:

$$p(\mathcal{C}_1|x, X, \mathbf{t}) = \int p(\mathcal{C}_1|x, \omega)p(\omega|X, \mathbf{t})\ \mathrm{d}\omega \approx \int \sigma(\omega^T x)q(\omega)\ \mathrm{d}\omega, \tag{7}$$

and with some calculations that can be found in [Bishop, 2006], we arrive at

$$p(\mathcal{C}_1|x, t) \approx \sigma\left(\mu_a \times \left(1 + \frac{\pi\sigma_a^2}{8}\right)^{1/2}\right), \tag{8}$$

where

$$\begin{aligned} \mu_a &= \omega_{MAP}^T x, \\ \sigma_a^2 &= x^T S_N x. \end{aligned} \tag{9}$$

Then, we assign as a label to x: $\begin{cases} 1 & \text{if } p(\mathcal{C}_1|x, t) > 0.5 \\ 0 & \text{else.} \end{cases}$

Accuracy on test and train set

- Test: 88.78%
- Train: 87.59%

Conclusion

- Very large prior, leading to
- Satisfactory results
- High computational cost
- Results easily interpretable

- The **activation function** is the Log Weibull function, *i.e.* the function f in (1) is

$$f(a) = 1 - \exp(-\exp(a)). \tag{10}$$

- **Multivariate gaussian prior** of the parameters ω, **i.e.**

$$\omega \sim \mathcal{N}(m_0, S_0). \tag{11}$$

- Thus, using (1), the **unnormalized log posterior** is:

$$\begin{aligned} \log p(\omega|X, \mathbf{t}) = & -\frac{1}{2}(\omega - m_0)^T S_0^{-1}(\omega - m_0) \\ & + \sum_{n=1}^{N}\{t_n \log(1 - \exp(-\exp(\omega^T x_n))) - (1 - t_n)\exp(\omega^T x_n)\} + \text{const.} \end{aligned} \tag{12}$$

We work with the logarithm posterior to avoid underflow.

Metropolis-Hastings random walk algorithm

Theoretical algorithm

Input ω_0

For $n = 0, \dots, M$:

1. Generate $\omega^* \sim q(\omega|\omega_n)$, where $q(\cdot|\cdot)$ is symmetric
2. Compute $R = \min\left\{1, \frac{p(\omega^*|X,\mathbf{t})}{p(\omega_n|X,\mathbf{t})}\right\}$
3. Generate $U \sim \mathcal{U}(0,1)$
   - if $U \leq R : \omega_{n+1} = \omega^*$,
   - else: $\omega_{n+1} = \omega_n$.

Output: array of accepted parameters

In practice

- $\omega_0 = 0 \in \mathbb{R}^{39}$
- $q(\omega|\omega_n) = \mathcal{N}(\omega|\omega_n, 0.000001 \times I)$
- $\log p(\omega|X,\mathbf{t})$ with $m_0 = 0 \in \mathbb{R}^{39}$ and $S_0 = 9 \times I_{39}$
- we work with logarithm to avoid underflow
- $M = 60000$

Results of the algorithm

- Computational time: 1h 43min
- 54310 different accepted coefficients
- 5690 rejected coefficients

Prediction

Prediction: theoretical aspects

The predictive distribution for $\mathcal{C}_1$ given a new feature vector x is obtained by **marginalization**:

$$\begin{aligned} p(\mathcal{C}_1|x, X, \mathbf{t}) &= \int p(\mathcal{C}_1|x, \omega)p(\omega|X, \mathbf{t}) \, \mathrm{d}\omega \\ &\approx \frac{\sum_{i=1}^{M} p(\mathcal{C}_1|x, \omega_i)}{M} \end{aligned} \tag{13}$$

We assign as a label to x: $\begin{cases} 1 & \text{if } p(\mathcal{C}_1|x, t) > 0.5 \\ 0 & \text{else.} \end{cases}$

Accuracy on test and train set

Considering a burn-in period of 25%, we get the following accuracies:

- Test: 88.10%
- Train: 87.5%

Conclusion

- Very large prior, leading to
- Satisfactory results
- Really cheap computationally: ≈ 0.1 sec/iteration

Model

- The **activation function** is the sigmoid function, *i.e.* the function f in (1) is

$$f(a) = \sigma(a) = \frac{\exp(a)}{1 + \exp(a)}. \tag{14}$$

- **Multivariate student prior** of the parameters ω, **i.e.**

$$\omega \sim t_\nu(\mu, \Sigma) \tag{15}$$

- Thus, using (1), the **unnormalized log posterior** is:

$$\begin{aligned} \log p(\omega | X, \mathbf{t}) = &- \frac{(\nu + p)}{2} \times \log\left(1 + \frac{1}{\nu}(\omega - \mu)^T \Sigma^{-1} (\omega - \mu)\right) \\ &+ \sum_{n=1}^{N} \{ t_n \log \sigma(\omega^T x_n) + (1 - t_n) \log(1 - \sigma(\omega^T x_n)) \} + \text{const}, \end{aligned} \tag{16}$$

  where p is the dimension of the parameter space.
  We work with the logarithm posterior to avoid underflow.

Metropolis-Adjusted langevin algorithm

Theoretical algorithm

Input ω_0

For $n = 0, \ldots, M$:

1. Compute $\nabla f(\omega)$
2. Generate $\omega^* = \omega_n + \frac{\tau}{2}\nabla f(\omega_n) + \sqrt{2\tau}\eta_n$, where $\tau > 0$ small, $\eta_n \sim \mathcal{N}(0, I)$ indep.
3. Compute $R = \min\left\{1, \frac{p(\omega^*|X,\mathbf{t})q(\omega_n|\omega^*)}{p(\omega_n|X,\mathbf{t})q(\omega^*|\omega_n)}\right\}$
4. Generate $U \sim \mathcal{U}(0,1)$
    - if $U \leq R : \omega_{n+1} = \omega^*$,
    - else: $\omega_{n+1} = \omega_n$.

Output: array of accepted parameters

In practice

- $\omega_0 = 0 \in \mathbb{R}^{39}$
- $f(\omega) = \log p(\omega|X, \mathbf{t})$, with $\nu = 3, p = 39, \mu = 0 \in \mathbb{R}^{39}, \Sigma = 9 \times I_{39}$
- $\log q(x|y) = -\frac{1}{4\tau}\|x - y - \tau\nabla \log p(y|X, \mathbf{t})\|_2^2$
- we work with the log to avoid underflow
- $\tau = 0.0001$
- $M = 1000$

Results of the algorithm

- Computational time: 15 minutes
- 849 different accepted coefficients
- 261 rejected coefficients

Prediction

Prediction: theoretical aspects

The predictive distribution for $\mathcal{C}_1$ given a new feature vector x is obtained by **marginalization**:

$$
\begin{aligned}
p(\mathcal{C}_1|x, X, \mathbf{t}) &= \int p(\mathcal{C}_1|x, \omega) p(\omega|X, \mathbf{t})\, \mathrm{d}\omega \\
&\approx \frac{\sum_{i=1}^{M} p(\mathcal{C}_1|x, \omega_i)}{M}
\end{aligned}
\tag{17}
$$

We assign as a label to x: $\begin{cases} 1 & \text{if } p(\mathcal{C}_1|x, t) > 0.5 \\ 0 & \text{else.} \end{cases}$

Accuracy on test and train set

Considering a burn-in period of 25%, we get the following accuracies:

- Test: 86.05%
- Train: 83.33%

Conclusion

- Very large prior
- Non satisfactory results: we always predict class $\mathcal{C}_2$
- Costly computationally: ≈ 0.3 sec/iteration

Comparaison of the models

Model	Accuracy on test set	Computational cost	Conclusion
• Laplace approximation • Symmetric Gaussian prior • Sigmoid activation function	88.78 %	• $\approx$ 0.3 sec/iteration • 7601 iterations • 42 minutes in total	• Good results, easily interpretable • Really costly • Simplicity of prior + properties of sigmoid $\Rightarrow$ easy computation of the Hessian
• MH random walk • Gaussian prior • Log Weibull activation function	88.10 %	• $\approx$ 0.1 sec/iteration • 3000 iterations • 1h 43min in total (87.7% in 5min)	• Satisfactory results • Really cheap • Approximation using only first order • Difficulties in predicting class $\mathcal{C}_1$
• MALA • Student-t prior • Sigmoid activation function	86.05 %	• $\approx$ 0.3 sec/iteration • 1000 iterations • 5.5 minutes in total	• Really bad results • Costly • We fail to predict class $\mathcal{C}_1$

Remarks

- For the choice of the prior, we based our choice on the paper [Ghush, Li, Mitra, 2018].
- In general, we see that is **really hard to predict whether an employee is feeling an sentiment of attrition**.
- Metropolis-Adjusted Langevin fails to predict any attrition.
- Preference to the **Laplace approximation for the prediction results**, but to **the Metropolis-Hastings random walk for a trade-off between good results and computational time**.

Conclusion

Bayesian logistic regression and Laplace approximation

- Our most satisfactory model in term of prediction is the **Bayesian logistic regression** using a **Laplace approximation**.
- The simplicity of the unnormalized posterior permits not to worry about the calimity of multimodality.
- This is a powerful approximation, using the second-order derivative of the unnormalized posterior.
- Looking more closely at ω_{MAP}, what seems to be important to keep an eye on for the company is:
  1. the overtime that an employee does: with a coeff of 1.15 it seems to be the most important factor regarding attrition;
  2. employees traveling frequently also seems to feel a bigger attrition: 0.97;
  3. however, being a Research Director seems to avoid this feeling: -0.95.
- Men (0.36) tend to feel a bigger attrition than female (0.18).

Possible ameliorations

- One could use a **Stochastic Gradient Descent** to reduce computational time.
- We could implement the **Wolfe-Powell** algorithm for γ.
- We could perform a more careful exploratory analysis to be able to choose **less general priors**.
- We could implement a **Metropolis-within-Gibbs** to reduce computational cost of the Metropolis-Hastings algorithm.
- We could run more iterations (*e.g.* on a more powerful computer, or on a cluster).

References

Christopher M. Bishop (2006)
Pattern Recognition and Machine Learning (Information Science and Statistics)

Joyee Ghosh, Yingbo Li, Robin Mitra (2018)
On the Use of Cauchy Prior Distributions for Bayesian Logistic Regression
2018 International Society for Bayesian Analysis

The End