Wykorzystanie metod optymalizacyjnych do modelowania empirycznych krzywych ROC dla bankowych modeli scoringowych za pomocą wybranych funkcji teoretycznych

Błażej Kochański

Idea

Nazwa krzywej ROC (receiver operating curve) pochodzi z dziedziny rozpoznawania sygnałów. Krzywa jest używana, żeby oceniać jakość klasyfikatorów binarnych w wielu dziedzinach (oprócz przetwarzanie sygnałów, medycyna, credit scoring, inne obszary wykorzystania uczenia maszynowego)

Istnieje kilka (albo i kilkanaście) formuł na rysowanie teoretycznej krzywej ROC. Przykłady to krzywe określane angielskimi terminami *normal, binormal, bifractal, bibeta*, *bigamma,* etc. Konieczność modelowania w kontekście biostatystycznym najczęściej wynika z braku dużej liczby obserwacji i potrzeby wyznaczania np. przedziałów ufności dla miar opartych na ROC (np. pole powierzchni pod krzywą ROC: AUROC, lub częściej w przypadku scoringu współczynnik Giniego: $Gini = 2 \cdot AUROC - 1$). W kontekście kredytów bankowych (credit scoring) konieczność modelowania wynika z innych powodów – np. chęć symulowania krzywej ROC dla modelu, który jeszcze nie istnieje.

Zadaniem projektu jest wykorzystanie metod optymalizacji dostępnych w bibliotekach Julia, Python lub R do dopasowania wybranych modelowych krzywych do empirycznych danych ROC.

1 Modele krzywej ROC

Niektóre modele krzywej ROC oparte są o równanie:

$$y = F_B\left(F_G^{-1}(x)\right), \tag{1}$$

gdzie F_B to dystrybuanta oceny punktowej (ang. *score*) dla złych klientów, zaś F_G^{-1} to dystrybuanta wartości scoringu dla klientów dobrych. Parametrycznie:

$$\begin{aligned} y &= F_B(s) \\ x &= F_G(s), \end{aligned} \tag{2}$$

gdzie *s* to ocena punktowa (lub jej monotoniczne odwzorowanie).

1. ***Bibeta:***

$$y = F_{\alpha_B,\beta_B}\left(F^{-1}_{\alpha_G,\beta_G}(x)\right), \tag{3}$$

gdzie F_{α_B,β_B} i F_{α_G,β_G} to dwie dystrybuanty.

2. **Uproszczona wersja modelu *bibeta*** (Chen & Hu, 2016):

Jest to model *bibeta*, gdzie $\alpha_B = 1$ and $\beta_G = 1$:

$$y = 1 - \left(1 - x^{\frac{1}{\alpha_G}}\right)^{\beta_B}, \tag{4}$$

3. ***Bigamma:***

(Dorfman et al., 1997):

$$y = G_{\alpha_B,\beta_B}\left(G^{-1}_{\alpha_G,\beta_G}(x)\right). \tag{5}$$

4. Binormal:

$$y = F_{\mu_B,\sigma_B}\left(F^{-1}_{\mu_G,\sigma_G}(x)\right). \tag{6}$$

Po transformacjach:

$$y = \Phi\left(a + b\Phi^{-1}(x)\right), \tag{7}$$

gdzie Φ to dystrybuanta rozkładu normalnego standardowego, a Φ^{-1} to funkcja do niej odwrotna, zaś *a* and *b* mają następujące wzory:

$$a = \frac{\mu_G - \mu_B}{\sigma_G}$$

i

$$b = \frac{\sigma_B}{\sigma_G}.$$

Do moich celów przyda się jeszcze jedna transformacja funkcji *binormal*, która jako jawnie przyjmuje parametr γ równy współczynnikowi Giniego modelu scoringowego:

$$y = \Phi\left(\Phi^{-1}\left(\frac{\gamma+1}{2}\right)\sqrt{1+b^2} + b\Phi^{-1}(x)\right) \tag{8}$$

5 *Bilogistic:*

$$y = \left(1 + \exp(\alpha_1 \ln\left(\frac{1}{x} - 1\right) - \alpha_0)\right)^{-1} \tag{10}$$

6. Power function:

$$y = x^{\theta}, \theta < 1 \tag{11}$$

7. Bifractal:

$$y = \beta\left(1 - (1-x)^{\frac{1+\gamma}{1-\gamma}}\right) + (1-\beta)x^{\frac{1-\gamma}{1+\gamma}}, \qquad (14)$$

8. Normal:

We wczesnej literaturze pojawia się również model normal, który można sprowadzić do modelu binormal z parametrem b=1.

2 Dane empiryczne, dla których stosujemy krzywe

Dane empiryczne pochodzą z publicznie dostępnych artykułów naukowych (Řezáč & Řezáč, 2011, Wójcicki & Migut, 2010, Tobback & Martens, 2017) i prezentacji branżowych (Conolly, 2015, Jennings, 2017) oraz uzyskanych przez autora zanonimizowanych krzywych ROC z polskich instytucji finansowych.

3 Podejście do optymalizacji

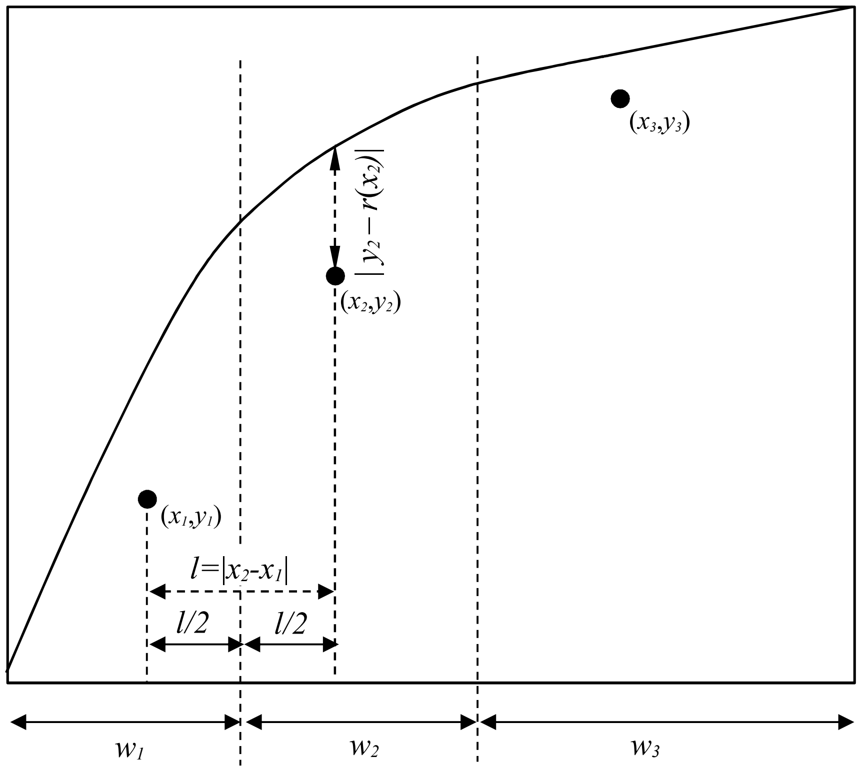
Na podstawie przeglądu dostępnych metod optymalizacji najwłaściwszym podejściem wydaje się zastosowanie jednej z bezderywatowych metod z ograniczeniami na parametry. Na ten moment zidentyfikowaną metodą dostępną w Julia jest metoda BOBYQA. (Powell, 2009, Bates et al., 2014).

Proponowana funkcja celu to:

$$f_{obj} = \sum_i |y_i - r(x_i)| \cdot w_i, \qquad (17)$$

gdzie $|a|$ oznacza wartość bezwzględną a, x_i i y_i to współrzędne punktów pochodzących z danych o empirycznej krzywej ROC, $r(x_i)$ to wartość modelowej funkcji ROC dla punktu x_i a w_i to wagi wyznaczone tak, jak na rysunku 1.

Rysunek 1: Wyznaczanie wag dla funkcji celu.



Planowany zakres projektu:

Zadania do wykonania do połowy projektu:

1. Implementacja poszczególnych funkcji (bibeta, binormal itd.) w Julia.
2. Przeprowadzenie optymalizacji dla wszystkich modeli i wszystkich wybranych zbiorów danych.

Zadania do wykonania w drugiej części projektu:

3. Wykrycie możliwości usprawnień (inne funkcje optymalizacyjne).
4. Przeprowadzenie usprawnień.
5. Przygotowanie raportu z badania.

Literatura

Bandos, A. I., Guo, B., & Gur, D. (2017). Estimating the Area Under ROC Curve When the Fitted Binormal Curves Demonstrate Improper Shape. *Academic Radiology, 24*(2), 209–219.

Bates, D., Mullen, K. M., Nash, J. C. and Varadhan, R. (2014). minqa: Derivativefree optimization algorithms by quadratic approximation. R package version 1.2.4. (Available from https://CRAN.R-project.org/package=minqa.)

Chen, W., & Hu, N. (2016). Proper bibeta ROC model: algorithm, software, and performance evaluation. In C. K. Abbey & M. A. Kupinski (Eds.), *Medical Imaging 2016: Image Perception, Observer Performance, and Technology Assessment* (Vol. 9787, p. 0E). San Diego, CA.

Conolly, S. (2017). Personality and risk: a new chapter for credit assessment. In *Credit Scoring and Credit Control XV Conference - Presented Papers*. Retrieved April 27, 2018, from https://www.business-school.ed.ac.uk/crc-conference/accepted-papers/

Dorfman, D. D., Berbaum, K. S., Metz, C. E., Lenth, R. V., Hanley, J. A., & Abu Dagga, H. (1997). Proper receiver operating characteristic analysis: the bigamma model. *Academic Radiology, 4*(2), 138-149.

Gönen, M., Heller, G. (2010). Lehmann family of ROC curves. *Medical Decision Making, 30*, 509–17.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874.

Jennings, A. (2015). Expanding the credit eligible population in the USA: A case study. In *Credit Scoring and Credit Control XIV Conference - Conference Papers*. Retrieved April 27, 2018, from https://www.business-school.ed.ac.uk/crc/category/conference-papers/2015/

Kochański, B. (2017), Fractal ROC curves – a simple model for the impact of the Gini coefficient's improvement on credit losses, Manuscript submitted for publication.

D. Mossman and H. Peng, "Using dual beta distributions to create "proper" ROC curves based on rating category data," Medical Decision Making (2015).

Powell, M. J. D. (2009), The BOBYQA algorithm for bound constrained optimization without derivatives. Technical Report DAMTP 2009/NA06, Cambridge: Centre for Mathematical Sciences, University of Cambridge. Retrieved April 27, 2018, from http://www.damtp.cam.ac.uk/user/na/NA_papers/NA2009_06.pdf

Řezáč, M., & Řezáč, F. (2011). How to Measure the Quality of Credit Scoring Models. *Czech Journal of Economics and Finance (Finance a Uver), 61*(5), 486-507.

Tobback, E., & Martens, D. (2017). Retail credit scoring using fine-grained payment data. In *Credit Scoring and Credit Control XV Conference - Presented Papers*. Retrieved April 27, 2018, from https://www.business-school.ed.ac.uk/crc-conference/accepted-papers/

Wójcicki, B., & Migut, G. (2010). Wykorzystanie skoringu do przewidywania wyludzen w Invest Banku. In *Skoring w Zarządzaniu Ryzykiem* (pp. 47-57). Kraków: Statsoft.