Python for Data Science - Midterm

Nhóm: **Nguyễn Phú Thành**: 18110014 - **Lê Hoàng Đức**: 18110075

13 - 12 - 2020

1 Thuật toán Logistic Regression

Cho bộ dữ liệu $\{\boldsymbol{x}_n, \boldsymbol{t}_n\}$, trong đó $\boldsymbol{t}_n \in \{0, 1\}$ với $n = 1, 2, \cdots, N$
Ta gọi C_1 là sự kiện biến ngẫu nhiên $\boldsymbol{t} = 1$, có xác suất:

$$p(C_1) = \mathbb{P}(\boldsymbol{t} = 1)$$

và C_2 là sự kiện biến ngẫu nhiên $\boldsymbol{t} = 0$, có xác suất:

$$p(C_2) = \mathbb{P}(\boldsymbol{t} = 0)$$

Hơn nữa giả sử vecto đầu vào $\boldsymbol{x}$ có phân phối chuẩn phụ thuộc vào lớp C_k. Giả sử thêm là các lớp có cùng ma trận hiệp phương sai $\boldsymbol{\Sigma}$, tức:

$$p(\boldsymbol{x}|C_k) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{\frac{-1}{2}(\boldsymbol{x} - \boldsymbol{\mu_k})^{\boldsymbol{T}}\boldsymbol{\Sigma^{-1}}(\boldsymbol{x} - \boldsymbol{\mu_k})\right\}$$

Khi đó phân phối hậu nghiệm của lớp C_1 dưới dạng:

$$\begin{aligned} p(C_1|\boldsymbol{x}) &= \frac{p(C_1)p(\boldsymbol{x}|C_1)}{p(\boldsymbol{x})} \\ &= \frac{p(C_1)p(\boldsymbol{x}|C_1)}{p(C_1)p(\boldsymbol{x}|C_1) + p(C_2)p(\boldsymbol{x}|C_2)} \\ &= \frac{1}{1 + \left(\dfrac{p(C_2)p(\boldsymbol{x}|C_2)}{p(C_1)p(\boldsymbol{x}|C_1)}\right)} \\ &= \frac{1}{1 + e^{-a}} = \sigma(a) \end{aligned}$$

trong đó:

$$\begin{aligned} a &= \ln\left(\frac{p(C_1)p(\boldsymbol{x}|C_1)}{p(C_2)p(\boldsymbol{x}|C_2)}\right) \\ &= \ln\left(\frac{p(C_1)}{p(C_2)}\right) - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu_1})^{\boldsymbol{T}}\boldsymbol{\Sigma^{-1}}(\boldsymbol{x} - \boldsymbol{\mu_1}) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu_2})^{\boldsymbol{T}}\boldsymbol{\Sigma^{-1}}(\boldsymbol{x} - \boldsymbol{\mu_2}) \\ &= \ln\left(\frac{p(C_1)}{p(C_2)}\right) + \frac{1}{2}\left(\boldsymbol{\mu_2^T\Sigma^{-1}\mu_2} - \boldsymbol{\mu_1^T\Sigma^{-1}\mu_1}\right) + (\boldsymbol{\mu_1} - \boldsymbol{\mu_2})^{\boldsymbol{T}}\boldsymbol{\Sigma^{-1}}\boldsymbol{x} \\ &= \boldsymbol{w^T x} + w_0 \end{aligned}$$

với

$$\begin{aligned} \boldsymbol{w} &= \boldsymbol{\Sigma^{-1}}(\boldsymbol{\mu_1} - \boldsymbol{\mu_2}) \\ w_0 &= \ln\left(\frac{p(C_1)}{p(C_2)}\right) + \frac{1}{2}\left(\boldsymbol{\mu_2^T\Sigma^{-1}\mu_2} - \boldsymbol{\mu_1^T\Sigma^{-1}\mu_1}\right) \end{aligned}$$

Đặt $\widetilde{\boldsymbol{w}} = (w_0, \boldsymbol{w})^T, \widetilde{\boldsymbol{x}} = (1, \boldsymbol{x})^T, \boldsymbol{y(x)} = \sigma(\widetilde{\boldsymbol{w}}^T\widetilde{\boldsymbol{x}}) = \sigma(\boldsymbol{w}^T\boldsymbol{x} + w_0)$
Vậy:

$$p(C_1|\boldsymbol{x}) = \boldsymbol{y}(\boldsymbol{x})$$
$$p(C_2|\boldsymbol{x}) = 1 - \boldsymbol{y}(\boldsymbol{x})$$

với $\sigma(x) = \dfrac{1}{1+e^{-x}}$ là hàm sigmoid
Ta có hàm hợp lý:

$$\begin{aligned} p(\boldsymbol{t}|\boldsymbol{x}, \widetilde{\boldsymbol{w}}) &= \prod_{n=1}^{N} p(\boldsymbol{t}_n|\boldsymbol{x_n}, \widetilde{\boldsymbol{w}}) \\ &= \prod_{n=1}^{N} \boldsymbol{y_n^{t_n}}(1-\boldsymbol{y}_n)^{1-\boldsymbol{t}_n} \end{aligned}$$

trong đó: $\boldsymbol{y_n} = y(\boldsymbol{x}_n) = \sigma(\widetilde{\boldsymbol{w}}^T\widetilde{\boldsymbol{x_n}})$ và $\boldsymbol{t} = (\boldsymbol{t}_1, \boldsymbol{t}_2, \cdots \boldsymbol{t}_N)^T$
Lấy đối của hàm ln của hàm hợp lý, ta được:

$$E_D(\widetilde{\boldsymbol{w}}) = -\ln p(\boldsymbol{t}|\boldsymbol{x}, \widetilde{\boldsymbol{w}}) = -\sum_{n=1}^{N} [\boldsymbol{t}_n \ln \boldsymbol{y}_n + (1-\boldsymbol{t}_n)\ln(1-\boldsymbol{y}_n)]$$

Ta gọi hàm E_D là hàm sai số cross-entropy
Để tìm ước lượng hợp lý cực đại, ta tìm giá trị nhỏ nhất của hàm sai số cross - entropy này
Vì hàm sai số E_D là hàm lồi nên để tìm giá trị nhỏ nhất, ta đưa bài toán về việc tìm nghiệm của phương trình $\nabla E_D(\widetilde{\boldsymbol{w}}) = 0$
Ta có:

$$\begin{aligned} d\boldsymbol{y}_n &= d\sigma(\widetilde{\boldsymbol{w}}^T\widetilde{\boldsymbol{x}}_n) \\ &= \sigma(\widetilde{\boldsymbol{w}}^T\widetilde{\boldsymbol{x}}_n)(1-\sigma(\widetilde{\boldsymbol{w}}^T\widetilde{\boldsymbol{x}}_n))d(\widetilde{\boldsymbol{w}}^T\widetilde{\boldsymbol{x}}_n) \\ &= \boldsymbol{y}_n(1-\boldsymbol{y}_n)\widetilde{\boldsymbol{x}}_n d\widetilde{\boldsymbol{w}} \end{aligned}$$

Khi đó:

$$\begin{aligned} dE_D &= -\sum_{n=1}^{N} \boldsymbol{t}_n \frac{1}{\boldsymbol{y}_n} d\boldsymbol{y}_n + (1-\boldsymbol{t}_n)\frac{-1}{1-\boldsymbol{y}_n} d\boldsymbol{y}_n \\ &= -\sum_{n=1}^{N} \boldsymbol{t}_n \frac{1}{\boldsymbol{y}_n} \boldsymbol{y}_n(1-\boldsymbol{y}_n)\widetilde{\boldsymbol{x}}_n d\widetilde{\boldsymbol{w}} + (1-\boldsymbol{t}_n)\frac{-1}{1-\boldsymbol{y}_n}\boldsymbol{y}_n(1-\boldsymbol{y}_n)\widetilde{\boldsymbol{x}}_n d\widetilde{\boldsymbol{w}} \\ &= -\sum_{n=1}^{N} [\boldsymbol{t}_n(1-\boldsymbol{y}_n)\widetilde{\boldsymbol{x}}_n - (1-\boldsymbol{t}_n)\boldsymbol{y}_n\widetilde{\boldsymbol{x}}_n] d\widetilde{\boldsymbol{w}} \\ &= \sum_{n=1}^{N} (\boldsymbol{y}_n - \boldsymbol{t}_n)\widetilde{\boldsymbol{x}}_n d\widetilde{\boldsymbol{w}} \end{aligned}$$

Suy ra:

$$\nabla E_D(\widetilde{\boldsymbol{w}}) = \sum_{n=1}^{N} (\boldsymbol{y}_n - \boldsymbol{t}_n)\widetilde{\boldsymbol{x}}_n = \sum_{n=1}^{N} \left[\sigma(\widetilde{\boldsymbol{w}}^T\widetilde{\boldsymbol{x}_n}) - \boldsymbol{t}_n\right]\widetilde{\boldsymbol{x}}_n$$

Sau đây, ta sẽ trình bày một trong nhiều cách để giải phương trình $\dfrac{\partial E_D}{\partial \widetilde{\boldsymbol{w}}} = 0$

2 Iterative reweighted least squares (IRLS) method

Cách cập nhật $\widetilde{\boldsymbol{w}}$ để hàm E_D đạt giá trị nhỏ nhất, được sử dụng theo phương pháp Newton có dạng:

$$\widetilde{\boldsymbol{w}}^{(\tau+1)} = \widetilde{\boldsymbol{w}}^{(\tau)} - \eta\nabla^2 E_D(\widetilde{\boldsymbol{w}}^{(\tau)})^{-1}\nabla E_D(\widetilde{\boldsymbol{w}}^{(\tau)})$$

trong đó η là tốc độ học
Ta có:

$$\begin{aligned}\nabla E_D(\widetilde{\boldsymbol{w}}) &= \sum_{n=1}^{N}(\boldsymbol{y}_n - \boldsymbol{t}_n)\widetilde{\boldsymbol{x}}_n = \boldsymbol{X}^T(\boldsymbol{y}-\boldsymbol{t}) \\ \nabla^2 E_D(\widetilde{\boldsymbol{w}}) &= \nabla\sum_{n=1}^{N}(\boldsymbol{y}_n - \boldsymbol{t}_n)\widetilde{\boldsymbol{x}}_n \\ &= \sum_{n=1}^{N}\boldsymbol{y}_n(1-\boldsymbol{y}_n)\widetilde{\boldsymbol{x}}_n\widetilde{\boldsymbol{x}}_n^T = \boldsymbol{X}^T\boldsymbol{R}\boldsymbol{X}\end{aligned}$$

trong đó $\boldsymbol{R}$ là ma trận đường chéo có hệ số trên đường chéo chính là:

$$R_{nn} = \boldsymbol{y}_n(1-\boldsymbol{y}_n)$$

Thay vào ta được:

$$\begin{aligned}\widetilde{\boldsymbol{w}}^{(\mathbf{new})} &= \widetilde{\boldsymbol{w}}^{(\mathbf{old})} - \eta(\boldsymbol{X}^T\boldsymbol{R}\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{y}-\boldsymbol{t}) \\ &= (\boldsymbol{X}^T\boldsymbol{R}\boldsymbol{X})^{-1}\{\boldsymbol{X}^T\boldsymbol{R}\boldsymbol{X}\widetilde{\boldsymbol{w}}^{(\mathbf{old})} - \eta\boldsymbol{X}^T(\boldsymbol{y}-\boldsymbol{t})\} \\ &= (\boldsymbol{X}^T\boldsymbol{R}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{R}\boldsymbol{z}\end{aligned}$$

trong đó: $\boldsymbol{z} = \boldsymbol{X}\widetilde{\boldsymbol{w}}^{(\mathbf{old})} - \eta\boldsymbol{R}^{-1}(\boldsymbol{y}-\boldsymbol{t})$

3 Stochastic gradient descent (SGD) method

Tại mỗi vòng lặp, $\widetilde{\boldsymbol{w}}$ được cập nhật dựa trên một điểm dữ liệu ngẫu nhiên. Hàm mất mát E_D của hồi quy logistic với một điểm dữ liệu $(\boldsymbol{x_n}, \boldsymbol{t_n})$ và gradient của nó ta thu được từ trình bày phía trên

$$\nabla_{\widetilde{\boldsymbol{w}}} E_D = (\boldsymbol{y}_n - \boldsymbol{t}_n)\widetilde{\boldsymbol{x}}_n = (\sigma(\widetilde{\boldsymbol{w}}^T\widetilde{\boldsymbol{x_n}}) - \boldsymbol{t}_n)\widetilde{\boldsymbol{x}}_n$$

Từ đó, công thức cập nhật nghiệm cho hồi quy logistic sử dụng SGD là

$$\widetilde{\boldsymbol{w}} \longleftarrow \widetilde{\boldsymbol{w}} - \eta(\boldsymbol{y}_n - \boldsymbol{t}_n)\widetilde{\boldsymbol{x}}_n = \widetilde{\boldsymbol{w}} - \eta(\sigma(\widetilde{\boldsymbol{w}}^T\widetilde{\boldsymbol{x_n}}) - \boldsymbol{t}_n)\widetilde{\boldsymbol{x}}_n$$

với η là tốc độ học.

4 Hồi quy logistic với chính quy hóa

Một trong các kỹ thuật phổ biến giúp vectơ trọng số $\widetilde{\boldsymbol{w}}$ không "quá lớn" (dưới một chuẩn nào đó) là sử dụng kĩ thuật chính quy hóa (Regulization).

Đây là một kỹ thuật kiểm soát, trong đó một đại lượng tỉ lệ với bình phương chuẩn l_2 của vector trọng số $\widetilde{\boldsymbol{w}}$ được cộng vào hàm mất mát để kiểm soát độ lớn của các hệ số.

Khi đó tối ưu hàm tổng mất mát được viết dưới dạng

$$E(\widetilde{\boldsymbol{w}}) = E_D(\widetilde{\boldsymbol{w}}) + \lambda E_{\widetilde{W}}(\widetilde{\boldsymbol{w}})$$

trong đó, λ là hệ số chuẩn hóa để kiểm soát mối tương quan giữa hàm mất mát phụ thuộc vào dữ liệu huấn luyện $E_D(\widetilde{\boldsymbol{w}})$ và số hạng kiểm soát $E_{\widetilde{W}}(\widetilde{\boldsymbol{w}})$.
Hai hàm kiểm soát phổ biến nhất là l_1 norm và l_2 norm. Ví dụ, khi chọn $E_{\widetilde{W}}(\widetilde{\boldsymbol{w}}) = \frac{1}{2}\|\widetilde{\boldsymbol{w}}\|_2^2$

$$E_{\widetilde{W}}(\widetilde{\boldsymbol{w}}) = \frac{1}{2}\widetilde{\boldsymbol{w}}^T\widetilde{\boldsymbol{w}}$$

thì hàm mất mát trở thành

$$E(\widetilde{\boldsymbol{w}}) = -\sum_{n=1}^{N} [\boldsymbol{t}_n \ln \boldsymbol{y}_n + (1 - \boldsymbol{t}_n) \ln(1 - \boldsymbol{y}_n)] + \frac{\lambda}{2} \widetilde{\boldsymbol{w}}^T \widetilde{\boldsymbol{w}}$$

Tính toán tương tự trên ta sẽ được:

$$\begin{aligned} \nabla E_D(\widetilde{\boldsymbol{w}}) &= \sum_{n=1}^{N} (\boldsymbol{y}_n - \boldsymbol{t}_n) \widetilde{\boldsymbol{x}}_n + \lambda \widetilde{\boldsymbol{w}} = \sum_{n=1}^{N} \left[\sigma(\widetilde{\boldsymbol{w}}^T \widetilde{\boldsymbol{x}_n}) - \boldsymbol{t}_n \right] \widetilde{\boldsymbol{x}}_n + \lambda \widetilde{\boldsymbol{w}} \\ \nabla^2 E_D(\widetilde{\boldsymbol{w}}) &= \boldsymbol{X}^T \boldsymbol{R} \boldsymbol{X} + \lambda \end{aligned}$$

Thay các giá trị này vào các công thức trên ta được cách cập nhật cho tham số $\widetilde{\boldsymbol{w}}$