Sparse Multinomial Logistic Regression via Bayesian L1 Regularisation

Доклад делали: Тинчурин Максим, Павлов Кирилл,
Смирнов Алексей, Мария Торговкина

Авторы статьи: Gavin C. Cawley, Nicola L. C. Talbot, Mark Girolami

Мотивация и контекст

- Многоклассовая логистическая регрессия – модель для классификации,
- Зачем разрежённость: отбор признаков, компактность и скорость
- Проблема: обычно нужен подбор гиперпараметра регуляризации через кросс-валидацию (это затратно по вычислительным ресурсам)

Базовая постановка (MLR + L1)

$$p(t_i^n|\boldsymbol{x}^n) = y_i^n = \frac{\exp\{a_i^n\}}{\sum_{j=1}^{c}\exp\{a_j^n\}} \qquad \text{where} \qquad a_i^n = \sum_{j=1}^{d} w_{ij}x_j^n$$

$$E_{\mathcal{D}} = \sum_{n=1}^{\ell} E_{\mathcal{D}}^n = -\sum_{n=1}^{\ell}\sum_{i=1}^{c} t_i^n \log\{y_i^n\}$$

$$L = E_{\mathcal{D}} + \alpha E_{\mathcal{W}} \qquad \text{where} \qquad E_{\mathcal{W}} = \sum\sum |w_{ij}|$$

$$\left|\frac{\partial E_{\mathcal{D}}}{\partial w_{ij}}\right| = \alpha \quad \text{if } |w_{ij}| > 0 \qquad \text{and} \qquad \left|\frac{\partial E_{\mathcal{D}}}{\partial w_{ij}}\right| < \alpha \quad \text{if } |w_{ij}| = 0.$$

Байесовская идея: “интегрируем” штраф

Ключевой трюк — не подбирать гиперпараметр, а интегрировать его из априорного распределения.

$$p(\boldsymbol{w}|\mathcal{D}) \propto P(\mathcal{D}|\boldsymbol{w})P(\boldsymbol{w}).$$

$$P(\boldsymbol{w}) = \left(\frac{\alpha}{2}\right)^W \exp\{-\alpha E_W\} = \prod_{i=1}^{W} \frac{\alpha}{2} \exp\{-\alpha|w_i|\},$$

$$p(\boldsymbol{w}) = \int p(\boldsymbol{w}|\alpha)p(\alpha)d\alpha.$$

Байесовская идея: “интегрируем” штраф

$$p(\boldsymbol{w}) = \frac{1}{2^W} \int_0^\infty \alpha^{W-1} \exp\{-\alpha E_{\mathcal{W}}\} d\alpha.$$

$$p(\boldsymbol{w}) = \frac{1}{2^W} \frac{\Gamma(W)}{E_{\mathcal{W}}^W} \quad \Longrightarrow \quad -\log p(\boldsymbol{w}) \propto W \log E_{\mathcal{W}},$$

$$M = E_{\mathcal{D}} + W \log E_{\mathcal{W}},$$

Эквивалентность ручному подбиранию гиперпараметра

$$L(\boldsymbol{w}) = L(\boldsymbol{w}^{\mathrm{MP}}) + \frac{1}{2}(\boldsymbol{w} - \boldsymbol{w}^{\mathrm{MP}})^T \boldsymbol{A}(\boldsymbol{w} - \boldsymbol{w}^{\mathrm{MP}})$$

$$-\log P(\mathcal{D}) = E_{\mathcal{D}} + \alpha E_{\mathcal{W}} + \frac{1}{2}\log|\boldsymbol{A}| + \log Z_{\mathcal{W}} + \text{constant.}$$

$$-\log\{\mathcal{D}\} = E_{\mathcal{D}} + W\log E_{\mathcal{W}} - \log\left\{\frac{\Gamma(W)}{2^W}\right\} + \frac{1}{2}\log|\boldsymbol{A}| + \text{constant}$$

Имплементация алгоритма

$$\frac{\partial E_{\mathcal{D}}^n}{\partial a_j^n} = \sum_{i=1}^{c} \frac{\partial E_{\mathcal{D}}^n}{\partial y_i^n} \frac{\partial y_i^n}{\partial a_j^n} \qquad \text{where} \qquad \frac{\partial E_{\mathcal{D}}^n}{\partial y_i^n} = -\frac{t_i^n}{y_i^n}, \qquad \frac{\partial y_i^n}{\partial a_j^n} = y_i \delta_{ij} - y_i y_j$$

and $\delta_{ij} = 1$ if $i = j$ and otherwise $\delta_{ij} = 0$. Substituting, we obtain,

$$\frac{\partial E_{\mathcal{D}}}{\partial \boldsymbol{a}_i} = \sum_{n=1}^{\ell} \left[y_i^n - t_i^n \right] \quad \Longrightarrow \quad \frac{\partial E_{\mathcal{D}}}{\partial w_{ij}} = \sum_{n=1}^{\ell} \left[y_i^n - t_i^n \right] x_j^n = \sum_{n=1}^{\ell} y_i^n x_j^n - \sum_{n=1}^{\ell} t_i^n x_j^n.$$

$$\frac{\partial^2 E_{\mathcal{D}}}{\partial w_{ij}} = \sum_{n=1}^{\ell} x_j^n \frac{\partial y_i^n}{\partial w_{ij}} = \sum_{n=1}^{\ell} y_i^n \left(1 - y_i^n\right) \left[x_j^n\right]^2 .$$

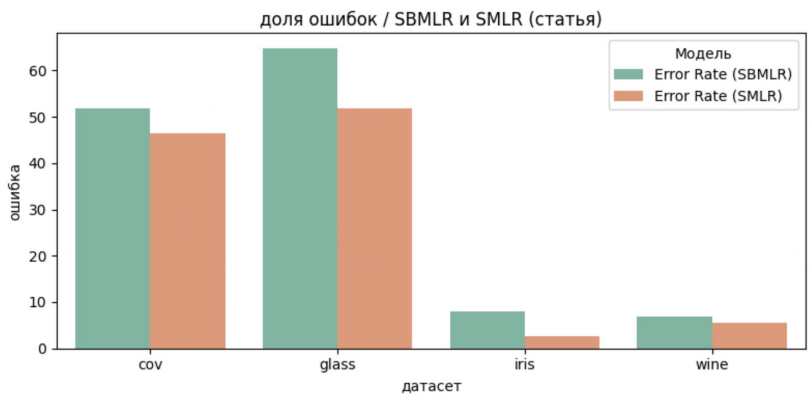
$$\frac{\partial L}{\partial w_{ij}} = \begin{cases} \dfrac{\partial E_{\mathcal{D}}}{\partial w_{ij}} + \alpha & \text{if } w_{ij} > 0 \\ \dfrac{\partial E_{\mathcal{D}}}{\partial w_{ij}} - \alpha & \text{if } w_{ij} < 0 \\ \dfrac{\partial E_{\mathcal{D}}}{\partial w_{ij}} + \alpha & \text{if } w_{ij} = 0 \text{ and } \dfrac{\partial E_{\mathcal{D}}}{\partial w_{ij}} + \alpha < 0 \\ \dfrac{\partial E_{\mathcal{D}}}{\partial w_{ij}} - \alpha & \text{if } w_{ij} = 0 \text{ and } \dfrac{\partial E_{\mathcal{D}}}{\partial w_{ij}} - \alpha > 0 \\ 0 & \text{otherwise} \end{cases}$$

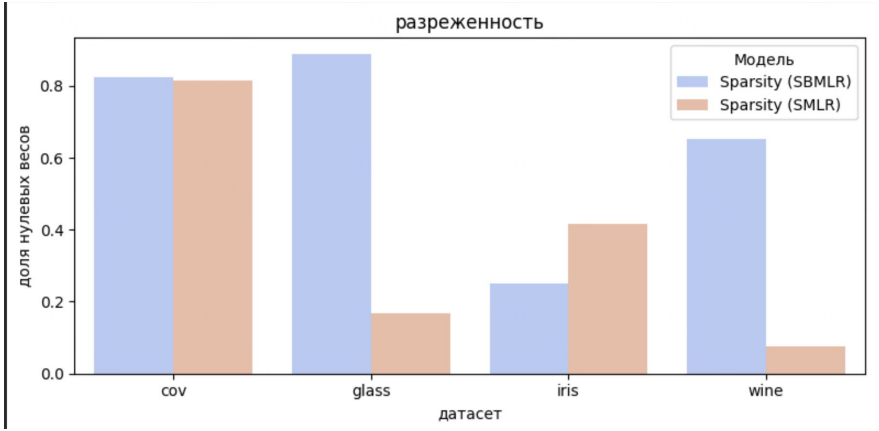
$$w_{ij} \leftarrow w_{ij} - \frac{\partial E_{\mathcal{D}}}{\partial w_{ij}} \left[\frac{\partial^2 E_{\mathcal{D}}}{\partial w_{ij}^2} \right]^{-1} .$$

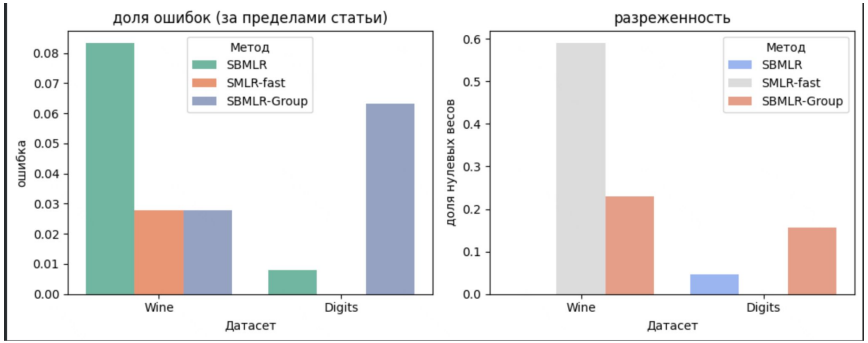
Результаты на бенчмарках

- 9 датасетов; сравнение SBMLR (без CV) vs SMLR (CV по гиперпараметру)
- Точность и кросс-энтропия — сопоставимы (нет значительного преимущества)
- Разреженность: SBMLR чаще чуть “разреженнее”
- Скорость: обычно ~100× быстрее (в худшем случае в 5 раз быстрее)

Визуализация







Когда применять: сильные стороны и ограничения

- Много признаков
- Важна скорость

Выводы

- Можно получить разрежённую многоклассовую логистическую регрессию без подбора силы L1-штрафа - за счёт байесовского интегрирования этого параметра.

Реализация кода