

Автоматическое определение релевантности параметров нейросети

Грабовой Андрей Валериевич

Московский физико-технический институт

ИОИ-12, г. Гаэта, Италия, 2018

Исследуются

Методы автоматического прореживания нейросетей.

Требуется

Предложить алгоритм нахождения оптимальной структуры модели на основе алгоритмов прореживания.

Проблемы

- Вычислительная сложность оптимизации,
- Невозможность получения адекватной статистической оценки параметров.

- *LeCun Y., Denker J. , Solla S.*
Optimal Brain Damage // Advances in Neural Information Processing Systems, 1989. Vol. 2. P. 598–605.
- *Graves A.*
Practical Variational Inference for Neural Networks // Advances in Neural Information Processing Systems, 2011. P. 2348–2356.
- *Bishop C.*
Pattern Recognition and Machine Learning. — Berlin: Springer, 2006. 758 p.
- *Neychev R., Katrutsa A., Strijov V.*
Robust selection of multicollinear features in forecasting // Factory Laboratory, 2016. Vol. 82. No 2. P. 68–74.

$$\mathcal{D} = \{\mathbf{x}_i, y_i\}, \quad i = 1, \dots, N,$$

где $\mathbf{x}_i \in \mathbb{R}^m$, $y_i \in \{1, \dots, Y\}$, Y — число классов.

$$f(\mathbf{x}, \mathbf{w}) : \mathbb{R}^m \times \mathbb{R}^n \rightarrow \{1, \dots, Y\},$$

где $\mathbf{w} \in \mathbb{R}^n$ — пространство параметров модели

$$f(\mathbf{x}, \mathbf{w}) = \text{softmax}(f_1(f_2(\dots(f_l(\mathbf{x}, \mathbf{w}))),$$

где $f_i(\mathbf{x}, \mathbf{w}) = \sigma(\mathbf{w}^T \mathbf{x})$, $i \in \{1 \dots l\}$, l — число слоев нейросети, σ некоторая функция активации.

$$\mathbb{W}_{\mathcal{A}} = \{\mathbf{w} \in \mathbb{R}^n \mid w_j \neq 0, j \in \mathcal{A}\},$$

где $\mathcal{A} \subset \mathcal{J} = \{1, \dots, n\}$ — множество активных индексов

Правдоподобие выборки:

$$\mathcal{L}_{\mathcal{D}}(\mathcal{D}, \mathcal{A}, \mathbf{w}) = \log p(\mathcal{D}|\mathcal{A}, \mathbf{w}),$$

где $p(\mathcal{D}|\mathcal{A}, \mathbf{w})$ — апостериорная вероятность \mathcal{D} при заданных \mathbf{w}, \mathcal{A}

Правдоподобие модели:

$$\mathcal{L}_{\mathcal{A}}(\mathcal{D}, \mathcal{A}) = \log p(\mathcal{D}|\mathcal{A}) = \log \int_{\mathbf{w} \in \mathbb{W}_{\mathcal{A}}} p(\mathcal{D}|\mathbf{w})p(\mathbf{w}|\mathcal{A})d\mathbf{w},$$

где $p(\mathbf{w}|\mathcal{A})$ — априорная вероятность \mathbf{w} в пространстве $\mathbb{W}_{\mathcal{A}}$

$$\begin{aligned}\mathcal{L}_{\mathcal{A}}(\mathcal{D}, \mathcal{A}) &= \log p(\mathcal{D}|\mathcal{A}) = \log \int_{\mathbf{w} \in \mathbb{W}_{\mathcal{J}}} p(\mathcal{D}|\mathbf{w})p(\mathbf{w}|\mathcal{A})d\mathbf{w} = \\ &= \int_{\mathbf{w} \in \mathbb{W}_{\mathcal{J}}} q(\mathbf{w}) \log \frac{p(\mathcal{D}, \mathbf{w}|\mathcal{A})}{q(\mathbf{w})} d\mathbf{w} - \int_{\mathbf{w} \in \mathbb{W}_{\mathcal{J}}} q(\mathbf{w}) \log \frac{p(\mathbf{w}|\mathcal{D}, \mathcal{A})}{q(\mathbf{w})} d\mathbf{w} \approx \\ &\approx \int_{\mathbf{w} \in \mathbb{W}_{\mathcal{J}}} q(\mathbf{w}) \log \frac{p(\mathcal{D}, \mathbf{w}|\mathcal{A})}{q(\mathbf{w})} d\mathbf{w} = \\ &= \int_{\mathbf{w} \in \mathbb{W}_{\mathcal{J}}} q(\mathbf{w}) \log \frac{p(\mathbf{w}|\mathcal{A})}{q(\mathbf{w})} d\mathbf{w} + \int_{\mathbf{w} \in \mathbb{W}_{\mathcal{J}}} q(\mathbf{w}) \log p(\mathcal{D}|\mathcal{A}, \mathbf{w}) d\mathbf{w} = \\ &= \mathcal{L}_{\mathbf{w}}(\mathcal{D}, \mathcal{A}, \mathbf{w}) + \mathcal{L}_{E}(\mathcal{D}, \mathcal{A}),\end{aligned}$$

где $q(\mathbf{w})$ — распределение аппроксимирующее неизвестное апостериорное распределение $p(\mathbf{w}|\mathcal{D}, \mathcal{A})$

$$q(\mathbf{w}) \sim \mathcal{N}(\mathbf{m}, \mathbf{A}_{ps})$$

где $\mathbf{m}, \mathbf{A}_{ps}^{-1}$ — вектор средних и матрица ковариации.

$$p(\mathbf{w}|\mathcal{A}) \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{A}_{pr}^{-1})$$

где $\boldsymbol{\mu}, \mathbf{A}_{pr}$ — вектор средних и матрица ковариации.

Задача оптимизации:

$$\begin{aligned} \hat{\mathbf{w}} &= \arg \min_{\mathbf{w} \in \mathbb{W}_{\mathcal{A}, \mathbf{A}_{ps}, \mathbf{A}_{pr}}} -\mathcal{L}_{\mathcal{A}}(\mathcal{D}, \mathcal{A}, \mathbf{w}) = \\ &= \arg \min_{\mathbf{w} \in \mathbb{W}_{\mathcal{A}, \mathbf{A}_{ps}, \mathbf{A}_{pr}}} D_{KL}(q(\mathbf{w}) || p(\mathbf{w}|\mathcal{A})) - \mathcal{L}_{\mathcal{D}}(\mathcal{D}, \mathcal{A}, \mathbf{w}) \end{aligned}$$

Случайное удаление параметров:

$\xi \sim \mathcal{U}(\mathcal{A})$ — индекс наименее релевантного параметра.

Оптимальное прореживание:

$$\delta \mathcal{L} = \sum_{j \in \mathcal{A}} g_j \delta w_j + \frac{1}{2} \sum_{i, j \in \mathcal{A}} h_{ij} \delta w_i \delta w_j + O(\|\delta \mathbf{w}\|^3)$$

Релевантность параметров определяется как рост ошибки вызванной удалением w_j :

$$\xi = \arg \min_{j \in \mathcal{A}} h_{jj} \frac{w_j^2}{2} \text{ — индекс наименее релевантного параметра.}$$

Вариационная оценка:

$$\xi = \arg \max_{j \in \mathcal{A}} \frac{p_j(\mathbf{w}|\mathcal{A})(0)}{p_j(\mathbf{w}|\mathcal{A})(\mu_j)} \text{ — индекс наименее релевантного параметра.}$$

Рассмотрим:

$$\hat{\mathbf{w}} = \arg \min_{\mathcal{A} \subset \mathcal{J}, \mathbf{w} \in \mathbb{W}_{\mathcal{A}}} -\mathcal{L}_{\mathcal{D}}(\mathcal{D}, \mathcal{A}, \mathbf{w})$$

Пусть:

\mathbf{A}_{ps} — матрица ковариационная матрица вектора $\hat{\mathbf{w}}$

$$\mathbf{A}_{ps} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T \Rightarrow \eta_j = \frac{\lambda_j}{\max(\mathbf{\Lambda})}$$

$$\xi = \arg \max_{j \in \mathcal{A}} \eta_j \quad q_{ij} = \frac{u_{ij}^2 / \lambda_{jj}}{\sum_{j=1}^n u_{ij}^2 / \lambda_{jj}}$$

$q_{\xi j}$ — максимальные значения отвечают наиболее зависимым параметрам

Иллюстрация метода Белсли

$$\hat{\mathbf{w}} = \begin{bmatrix} \sin(x) \\ \cos(x) \\ 2+\cos(x) \\ 2+\sin(x) \\ \cos(x) + \sin(x) \\ x \end{bmatrix}, x \in [0.0, 0.02, \dots, 20.0]$$

η_0	η_1	η_2	η_3	η_4	η_5
1.0	1.5	3.3	$2 \cdot 10^{15}$	$8 \cdot 10^{15}$	$1 \cdot 10^{16}$

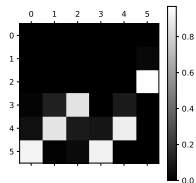
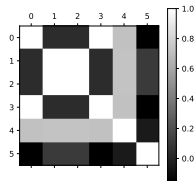


Рис.: Иллюстрация метода Белсли

Таблица: Описание выборок

Выборка	Тип задачи	Размер выборки	Число признаков
Wine	классификация	178	13
Boston Housing	регрессия	506	13
Synthetic data	регрессия	10000	100

Этап первый:

$$\mathbf{w}_{\text{synthetic}} \sim \mathcal{N}(\mathbf{m}_{\text{synthetic}}, \mathbf{A}_{\text{synthetic}})$$

$$\mathbf{m}_{\text{synthetic}} = \begin{bmatrix} 1.0 \\ 0.0025 \\ \dots \\ 0.0025 \end{bmatrix} \quad \mathbf{A}_{\text{synthetic}} = \begin{bmatrix} 1.0 & 10^{-3} & \dots & 10^{-3} & 10^{-3} \\ 10^{-3} & 1.0 & \dots & 0.95 & 0.95 \\ \dots & \dots & \dots & \dots & \dots \\ 10^{-3} & 0.95 & \dots & 0.95 & 1.0 \end{bmatrix}$$

Этап второй:

$$\mathcal{D}_{\text{synthetic}} = \{(\mathbf{x}_i, y_i) | \mathbf{x}_i \sim \mathcal{N}(\mathbf{1}, \mathbf{I}), y_i = x_{i0}, i = 1 \dots 10000\}$$

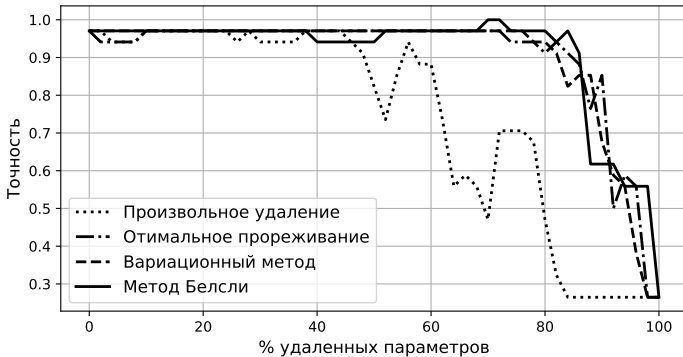


Рис.: Качество прогноза при удалении параметров на выборке Wine

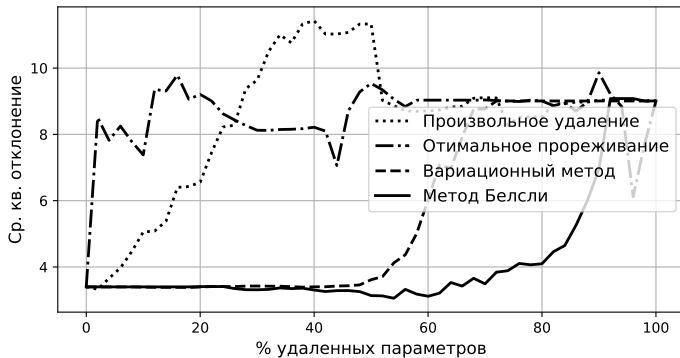


Рис.: Качество прогноза при удалении параметров на выборке Boston

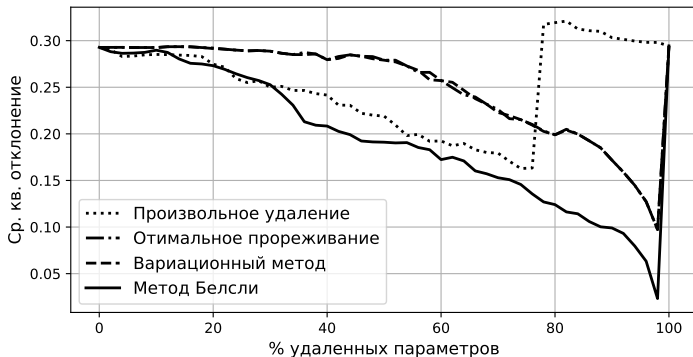
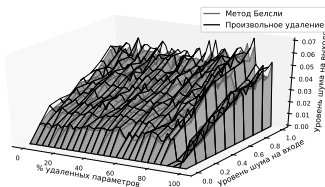
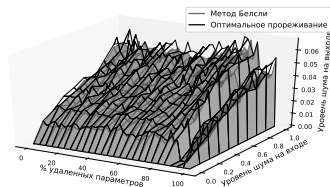


Рис.: Качество прогноза при удалении параметров на синтетических данных

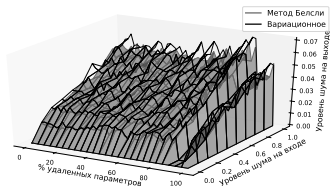
Результаты эксперимента



(a) Произвольное удаление параметров



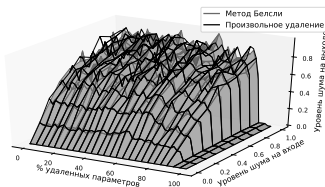
(b) Оптимальное прореживание



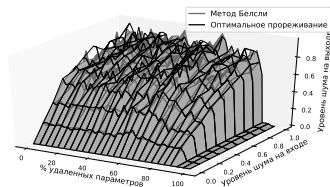
(c) Вариационный метод

Рис.: Влияние шума в начальных данных на шум выхода нейросети на выборке Wine

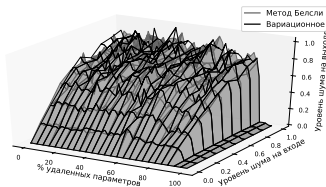
Результаты эксперимента



(a) Произвольное удаление параметров



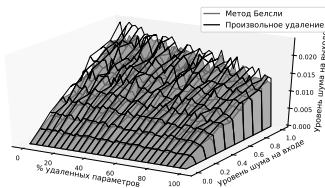
(b) Оптимальное прореживание



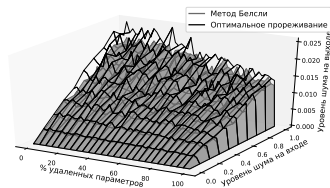
(c) Вариационный метод

Рис.: Влияние шума в начальных данных на шум выхода нейросети на выборке Boston

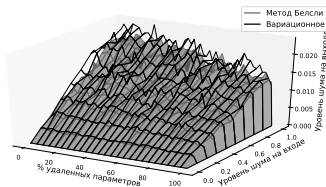
Результаты эксперимента



(a) Произвольное удаление параметров



(b) Оптимальное прореживание



(c) Вариационный метод

Рис.: Влияние шума в начальных данных на шум выхода нейросети на синтетических данных

- Исследовались методы прореживания нейросетей,
- Был предложен алгоритм прореживания параметров модели на основе метода Белсли.

Нерешенные проблемы

- Вычислительная сложность оптимизации,
- Невозможность получения адекватной статистической оценки параметров.