Автоматическое построение нейросети оптимальной сложности

Маркин Валерий, Забазнов Антон, Горян Николай, Сергей Губанов, Сергей Таранов, Товкес Артём, Улитин Александр, Криницкий Константин

Московский физико-технический институт

10 декабря, 2018г.

Цель работы

Иследуется
Задача выбора структуры нейронной сети.

Требуется
Найти нейросеть оптимальной сложности.

Проблемы

- Большое количество параметров,
- Высокая вычислительная сложность оптимизации,
- Невозможность использования эвристических и переборных алгоритмов выбора струкутры модели

Литература

- *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.

Постановка задачи

$$\mathfrak{D}^{\text{train}} = \{\mathbf{x}_i, y_i\}, \quad i = 1, \dots, m^{\text{train}},$$
$$\mathfrak{D}^{\text{valid}} = \{\mathbf{x}_i, y_i\}, \quad i = 1, \dots, m^{\text{valid}},$$

где $\mathbf{x}_i \in \mathbf{X} \subset \mathbb{R}^{\text{n}}, \quad y_i \in \mathbf{Y} \subset \mathbb{R}$.

$y \in \mathbf{Y} = \{1, \dots, Z\}$, где Z - количество классов.

Модель задаётся ориентированным графом $\mathbf{G} = (\mathbf{V}, \mathbf{E})$

$\mathbf{g}^{i,j}$— базовые функции ребра (i, j) с весами $\gamma^{i,j}$

Требуется построить такую модель $\mathbf{f}$ с параметрами $\mathbf{W} \in \mathbb{R}^{\text{n}}$:

$$\mathbf{f}(\mathbf{x}, \mathbf{W}) = \{\mathbf{f}_i(\mathbf{x}, \mathbf{w}_i)\}_{i=1}^{|\mathbf{V}|}$$

где $\mathbf{f_i}(\mathbf{x}, \mathbf{w_i})$ - подмодель с параметрами $\mathbf{w}_i$ задаётся как:

$$\mathbf{f}_i(\mathbf{x}, \mathbf{w}_i) \ = \sum_{j \in adj(i)} \left\langle \gamma^{i,j}, \mathbf{g}^{i,j} \right\rangle \mathbf{f}_j(\mathbf{x}, \mathbf{w}_j)$$

.

Постановка задачи

Правдоподобие выборки:

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

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

Правдоподобие модели:

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

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

Постановка задачи

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

$$= \int_{\mathbf{w}\in\mathbb{W}_{\mathcal{J}}} q(\mathbf{w}) \log \frac{p(\mathfrak{D}, \mathbf{w}|\mathcal{A})}{q(\mathbf{w})} d\mathbf{w} - \int_{\mathbf{w}\in\mathbb{W}_{\mathcal{J}}} q(\mathbf{w}) \log \frac{p(\mathbf{w}|\mathfrak{D}, \mathcal{A})}{q(\mathbf{w})} d\mathbf{w} \approx$$

$$\approx \int_{\mathbf{w}\in\mathbb{W}_{\mathcal{J}}} q(\mathbf{w}) \log \frac{p(\mathfrak{D}, \mathbf{w}|\mathcal{A})}{q(\mathbf{w})} d\mathbf{w} =$$

$$= \int_{\mathbf{w}\in\mathbb{W}_{\mathcal{J}}} q(\mathbf{w}) \log \frac{p(\mathbf{w}|\mathcal{A})}{q(\mathbf{w})} d\mathbf{w} + \int_{\mathbf{w}\in\mathbb{W}_{\mathcal{J}}} q(\mathbf{w}) \log p(\mathfrak{D}|\mathcal{A}, \mathbf{w}) d\mathbf{w} =$$

$$= \mathcal{L}_{\mathbf{w}}(\mathfrak{D}, \mathcal{A}, \mathbf{w}) + \mathcal{L}_{E}(\mathfrak{D}, \mathcal{A}),$$

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

Постановка задачи

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

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

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

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

Задача оптимизации:

$$\hat{\mathbf{w}} = \operatorname*{arg\,min}_{\mathbf{w}\in\mathbb{W}_{\mathcal{A}}, \mathbf{A}_{\text{ps}}, \mathbf{A}_{\text{pr}}} -\mathcal{L}_{\mathcal{A}}(\mathfrak{D}, \mathcal{A}, \mathbf{w}) =$$

$$= \operatorname*{arg\,min}_{\mathbf{w}\in\mathbb{W}_{\mathcal{A}}, \mathbf{A}_{\text{ps}}, \mathbf{A}_{\text{pr}}} D_{KL}\left(q(\mathbf{w})||p(\mathbf{w}|\mathcal{A})\right) - \mathcal{L}_{\mathfrak{D}}(\mathfrak{D}, \mathcal{A}, \mathbf{w})$$

Некоторые методы прореживания нейросетей

Случайное удаление параметров:

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

Оптимальное прореживание:

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

Релеватность параметров определяется как рост ошибки вызванной удалением w_j:

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

Вариационная оценка:

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

Рассмотрим:

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

Пусть:

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

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

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

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

Илюстрация метода Белсли

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

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

Экспериментальные данные

Таблица: Описание выборок

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

Синтетические данные

Этап первый:

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

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

Этап второй:

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

Вывод

- 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.