Последовательное порождение моделей глубокого обучения оптимальной сложности

О. Ю. Бахтеев

Научный руководитель: д.ф.-м.н. В.В. Стрижов
Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

16 июня 2016 г.

Цели работы

Требуется

Предложить алгоритм автоматического построения сетей глубокого обучения субоптимальной сложности.

Проблемы выбора структуры сети:

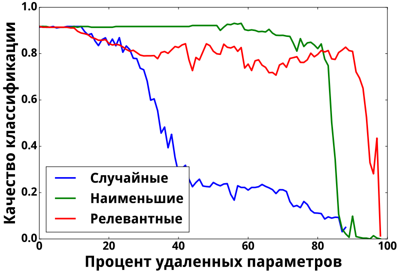
- многоэкстремальность задачи оптимизации,
- избыточность множества параметров модели,
- неустойчивость параметров модели относительно возмущений,
- зависимость качества модели от начального приближения параметров.

Методы

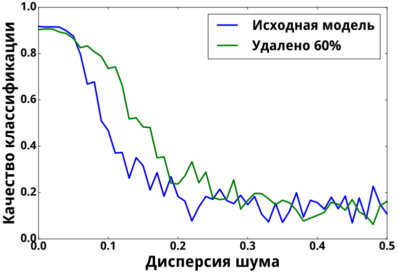
Предлагается декомпозировать модель на порождающую и разделяющую. В качестве функционалов качества для каждой из них выступает нижняя вариационная оценка интеграла правдоподобия модели.

Проблемы обучения сетей

Качество моделей с избыточным количеством параметров не меняется при удалении параметров.



Избыточность параметров модели



Неустойчивость модели

Исследование основывается на следующих работах:

- Kuznetsov M.P., Tokmakova A.A., Strijov V.V. Analytic and stochastic methods of structure parameter estimation // Informatica, 2016.
- D. P. Kingma, M. Welling. Auto-Encoding Variational Bayes // Proceedings of the 2nd International Conference on Learning Representations, 2014.
- M. Welling, Y. Teh. Bayesian learning via stochastic gradient Langevin dynamics // International Conference on Machine Learning, 2011.
- D. Duvenaud, D. Maclaurin, R. P. Adams. Early Stopping as Nonparametric Variational Inference // Artificial Intelligence and Statistics, 2016.
- Бахтеев О.Ю., Попова М.С., Стрижов В.В. Системы и средства глубокого обучения в задачах классификации // Системы и средства информатики, 2016, 2.

Формальная постановка задачи

Задана выборка

$$\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}, i = 1, \dots, N,$$

состоящая из множества пар объект-класс.
Каждый объект $\mathbf{x}_i \in \mathbf{X}$ принадлежит одному из Z классов с меткой $y_i \in \mathbf{Y}$.

Сетью глубокого обучения $\mathbf{f}$ назовем суперпозицию функций

$$\mathbf{f}(\mathbf{w}, \mathbf{x}) = \mathbf{f}_1(\mathbf{f}_2(\dots \mathbf{f}_K(\mathbf{x}))) : \mathbb{R}^n \to [0, 1]^Z,$$

где $\mathbf{f}_k, k \in \{1, \dots, K\}$, — модели; $\mathbf{w}$ — вектор параметров; r-я компонента вектора $\mathbf{f}(\mathbf{x}, \mathbf{w})$ — вероятность отнесения объекта $\mathbf{x}_i$ к классу с меткой r.

Исследуемые модели

Автокодировщик

Скрытое представление:

$$\mathbf{z} = \sigma(\mathbf{W}\mathbf{x} + \mathbf{b}).$$

Реконструкция $\mathbf{x}$:

$$\mathbf{r}(\mathbf{x}) = \sigma(\mathbf{W}^{\mathsf{T}}\mathbf{z} + \mathbf{b}_r).$$

Оптимизируемый функционал — ошибка реконструкции:

$$||\mathbf{r}(\mathbf{x}) - \mathbf{x}||_2^2 \to \min.$$

Нейросеть с одним скрытым слоем

$$\mathbf{a}(\mathbf{x}) = \mathbf{W}_2^{\mathsf{T}}\mathbf{tanh}(\mathbf{W}_1^{\mathsf{T}}\mathbf{x}),$$

$$\mathbf{f}_{\mathrm{SM}}(\mathbf{x}) = \frac{\mathbf{exp}\big(\mathbf{a}(\mathbf{x})\big)}{\sum_{j=1}^{Z} \exp\big(a_j(\mathbf{x})\big)},$$

Оптимизируемый функционал — правдоподобие:

$$\sum_{\mathbf{x},y\in\mathfrak{D}} \log p(y|\mathbf{x},\mathbf{w}) \to \max.$$

Эксплуатационные критерии качества модели

Точность S модели $\mathbf{f}(\mathbf{x}, \mathbf{w})$ — величина ошибки на контрольной выборке.

Устойчивость модели $\mathbf{f}(\mathbf{x}, \mathbf{w})$ — число обусловленности матрицы $\mathbf{A}$:

$$\eta(\mathbf{w}) = \frac{\lambda_{\text{max}}}{\lambda_{\text{min}}} \quad \text{при гипотезе } \mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{A}^{-1}),$$

λ_{max} — максимальное, а λ_{min} — минимальное собственные числа матрицы $\mathbf{A}$.

Структурная сложность C модели $\mathbf{f}(\mathbf{x}, \mathbf{w})$ — количество параметров $\mathbf{w}$ модели $\mathbf{f}$.

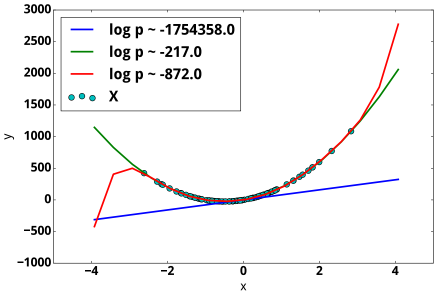
Статистическая сложность модели $\mathbf{f}$:

$$\mathrm{MDL}(\mathfrak{D}, \mathbf{f}) = -\log p(\mathbf{f}) - \log \left(p(\mathfrak{D}|\mathbf{f})\delta\mathfrak{D})\right),$$

где $\delta\mathfrak{D}$ — допустимая точность передачи информации о выборке $\mathfrak{D}$.

Модель $\mathbf{f} \in \mathfrak{F}$ оптимальна, если достигается максимум **правдоподобия модели**:

$$p(\mathfrak{D}|\mathbf{f}) = \int_{\mathbf{w}} p(\mathfrak{D}|\mathbf{w})p(\mathbf{w}|\mathbf{f})d\mathbf{w}.$$



Вариационная оценка интегральной функции правдоподобия

Проблема: вычисление оценки правдоподобия модели имеет высокую вычислительную сложность.

Утверждение [Bishop, 2006]. Справедливы нижние оценки интегральной функции правдоподобия:

$$\log p(\mathfrak{D}|\mathbf{f}) \geq \int q(\mathbf{w}) \log \frac{p(\mathfrak{D}, \mathbf{w}|\mathbf{f})}{q(\mathbf{w})} d\mathbf{w} =$$

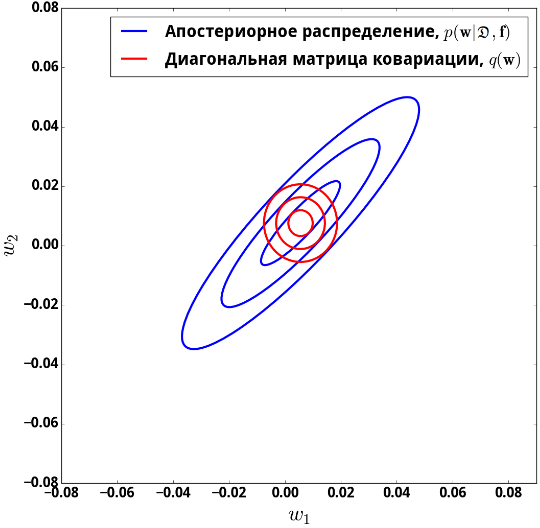
$$= -\mathrm{D}_{\mathrm{KL}}(q(\mathbf{w})||p(\mathbf{w}|\mathbf{f})) + \int q(\mathbf{w}) \log p(\mathfrak{D}|\mathbf{w}, \mathbf{f}) d\mathbf{w},$$

где

$$\mathrm{D}_{\mathrm{KL}}(q(\mathbf{w})||p(\mathbf{w}|\mathbf{f})) = -\int q(\mathbf{w}) \log \frac{p(\mathbf{w}|\mathbf{f})}{q(\mathbf{w})} d\mathbf{w},$$

$q \in Q$ — параметрическое семейство распределений.

Максимизация нижней оценки интегральной функции правдоподобия эквивалентна минимизации $\mathsf{D}_{\mathsf{KL}}(q(\mathbf{w})||p(\mathbf{w}|\mathfrak{D},\mathbf{f}))$.



Суперпозиция субоптимальных моделей: $\mathbf{f} = \mathbf{f}_D\bigl(\mathbf{f}_G(\mathbf{x})\bigr)$

Порождающую модель $\mathbf{f}_G$ назовем **субоптимальной** на множестве порождающих моделей $\mathfrak{F}_G$ по семейству распределений Q, если модель доставляет максимум нижней вариационной оценке интеграла:

$$\max_{q \in Q} \int_{\mathbf{w}} q(\mathbf{w}) \log \frac{p(\mathbf{X}, \mathbf{w}|\mathbf{f}_G)}{q(\mathbf{w})} d\mathbf{w}.$$

Разделяющую модель $\mathbf{f}_D$ назовем **субоптимальной** для модели $\mathbf{f}_G$ на множестве разделяющих моделей $\mathfrak{F}_D$, если модель доставляет максимум нижней вариационной оценке интеграла:

$$\max_{q \in Q} \int_{\mathbf{w}} q(\mathbf{w}) \log \frac{p(\mathbf{Y}, \mathbf{w}|\hat{\mathbf{Z}}, \mathbf{f}_D)}{q(\mathbf{w})} d\mathbf{w},$$

где $\hat{\mathbf{Z}}$ — скрытое представление выборки $\mathbf{X}$:
$\hat{\mathbf{Z}} = \mathrm{argmax}_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X})$.

Вариационный автокодировщик $\mathbf{f}_G$

Пусть объекты $\mathbf{X}$ порождены при условии скрытого представления $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$:

$$\mathbf{x} \sim p(\mathbf{x}|\mathbf{z}, \mathbf{w}).$$

Распределение $p(\mathbf{x}|\mathbf{z}, \mathbf{w})$ — неизвестно.
Правдоподобие выборки:

$$\log p(\mathbf{x}|\mathbf{w}) \geq \mathsf{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \log p(\mathbf{x}|\mathbf{z}, \mathbf{w}) - \mathsf{D}_{\mathsf{KL}}\left(q_\phi(\mathbf{z}|\mathbf{x}) || p(\mathbf{z})\right) \to \max .$$

$q_\phi(\mathbf{z}|\mathbf{x}), p(\mathbf{x}|\mathbf{z}, \mathbf{w})$ — распределения, задаваемые нейросетью:

$$q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_\phi(\mathbf{x}), \boldsymbol{\sigma}^2_\phi(\mathbf{x})),$$

$$p(\mathbf{x}|\mathbf{z}, \mathbf{w}) = \mathcal{N}(\boldsymbol{\mu}_\mathbf{w}(\mathbf{z}), \boldsymbol{\sigma}^2_\mathbf{w}(\mathbf{z})).$$

Правдоподобие модели:

$$\log p(\mathbf{x}|\mathbf{f}) \geq \int_\mathbf{w} q_\mathbf{w}(\mathbf{w}) \left(\log p(\mathbf{x}|\mathbf{w}) + \log p(\mathbf{w}|\mathbf{f}) - \log q_\mathbf{w}(\mathbf{w})\right) d\mathbf{w}.$$

Вариационная оценка: разделяющая модель $\mathbf{f}_D$

$$\log p(\mathbf{Y}|\mathbf{X},\mathbf{f}) \geq \mathsf{E}_{q(\mathbf{w})}[\log p(\mathbf{Y},\mathbf{w}|\mathbf{X},\mathbf{f})] - \mathsf{S}(q(\mathbf{w})),$$

S — энтропия, q — вспомогательное распределение из семейства Q:

$$q^\tau = T(q^{\tau-1}).$$

Теорема [Бахтеев, 2016]. Пусть L — функция потерь, градиент которой — непрерывно-дифференцируемая функция с константой Липшица K. Пусть $\mathbf{w}^1, \dots, \mathbf{w}^r$ — начальные приближения оптимизации модели. Пусть α — шаг градиентного спуска, такой что:

- $\alpha < \frac{1}{K}$,
- $\alpha^{(-1)} > \max_{\gamma \in \{1,\dots,r\}} \lambda_{\max}(\mathbf{H}(\mathbf{w}^\gamma))$.

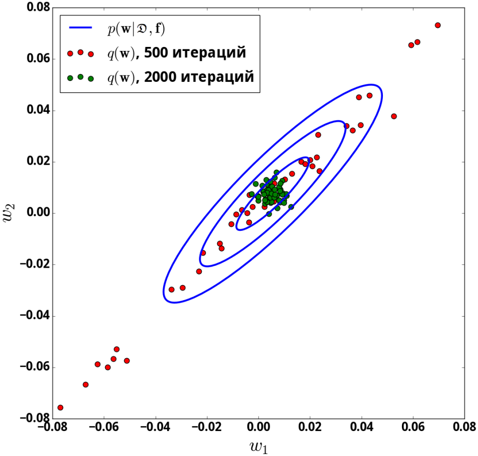
Тогда

$$\mathsf{S}(q^\tau(\mathbf{w})) - \mathsf{S}(q^{\tau-1}(\mathbf{w})) \sim \frac{1}{r}\sum_{\gamma=1}^{r}\left(\alpha Tr[\mathbf{H}(\mathbf{w}^\gamma)] - \alpha^2 Tr[\mathbf{H}(\mathbf{w}^\gamma)\mathbf{H}(\mathbf{w}^\gamma)]\right) + o_{\alpha\to 0}(1),$$

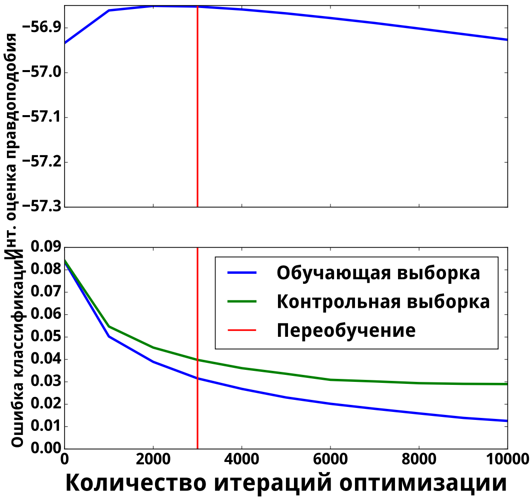
где $\mathbf{H}$ — гессиан функции потерь L.

Вариационная оценка с использованием градиентного спуска

Градиентный спуск не минимизирует $D_{KL}(q(\mathbf{w})||p(\mathbf{w}|\mathfrak{D},\mathbf{f}))$.



Контоль переобучения $\mathbf{f(x)}$



Стохастическая динамика Ланжевина

Модификация стохастического градиентного спуска:

$$\Delta \mathbf{w} = \alpha \nabla (\log p(\mathbf{w}) + \frac{m}{\hat{m}} \log p(\hat{\mathfrak{D}}|\mathbf{w})) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \frac{\alpha}{2})$$

где $\hat{m}$ — размер подвыборки, $\hat{\mathfrak{D}} \subset \mathfrak{D}$ — подвыборка, шаг оптимизации α изменяется с количеством итераций:

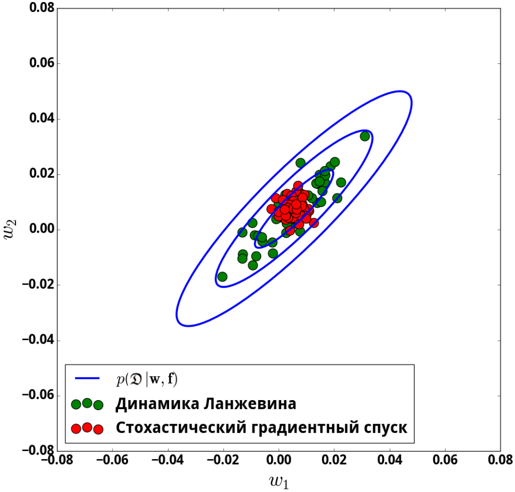
$$\sum_{\tau=1}^{\infty} \alpha_\tau = \infty, \quad \sum_{\tau=1}^{\infty} \alpha_\tau^2 < \infty.$$

Утверждение [Welling, 2011]. Распределине $q^\tau(\mathbf{w})$ сходится к апостериорному распределению $p(\mathbf{w}|\mathfrak{D}, \mathbf{f})$.
Изменение энтропии с учетом добавленного шума:

$$\hat{\mathsf{S}}(q^\tau(\mathbf{w})) \geq \frac{1}{2}|\mathbf{w}| \log(\exp(\frac{2\mathsf{S}(q^\tau(\mathbf{w}))}{|\mathbf{w}|}) + \exp(\frac{2\mathsf{S}(\epsilon)}{|\mathbf{w}|})).$$

Вариационная оценка: разделяющая модель

Распределения параметров после 2000 итераций:



Алгоритм выбора модели субоптимальной сложности

Найти:

- субоптимальную модель порождения $\mathbf{f}_G \in \mathfrak{F}_G$;
- Оптимальные параметры $\mathbf{w}_G$ модели порождения $\mathbf{f}_G$:

$$\mathbf{w}_G = \operatorname{argmax} \, p(\mathbf{X}, \mathbf{w}_G | \mathbf{f}_G).$$

- субоптимальную модель разделения $\mathbf{f}_D \in \mathfrak{F}_D$;
- оптимальные параметры $\mathbf{w}_D$ модели порождения $\mathbf{f}_D$:

$$\mathbf{w}_D = \operatorname{argmax} \, p(\mathbf{Y}, \mathbf{w}_D | \hat{\mathbf{Z}}, \mathbf{f}_D), \quad \hat{\mathbf{Z}} = \operatorname{argmax}_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}).$$

Дообучить сеть: $p(\mathbf{Y}, \mathbf{w}|\mathbf{X}, \mathbf{f}) \to \max$.

Вычислительный эксперимент

Цель эксперимента: анализ качества субоптимальных моделей до и после дообучения.

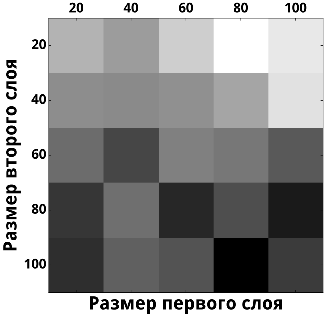
Данные. выборка MNIST:

- Набор изображений рукописных цифр размером 28×28 пикселей (784 признака).
- Мощность выборки: 60000 векторов (50000 для обучения и 10000 для контроля).
- Ряд экспериментов проводился на подвыборке MNIST с меньшим количеством признаков.

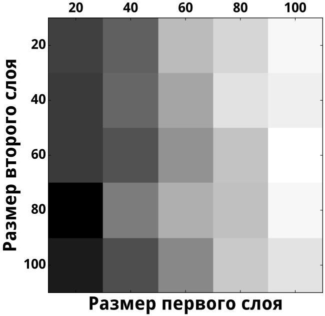
Рассматриваемые модели: нейросети с тремя скрытыми слоями (2 слоя — автокодировщик, 1 слой - softmax-сеть).

Результат: вариационный автокодировщик

Размер слоев субоптимальной модели: (80,20).
Размер слоев модели, полученной по оценке максимальной апостериорной вероятности: (100,60).



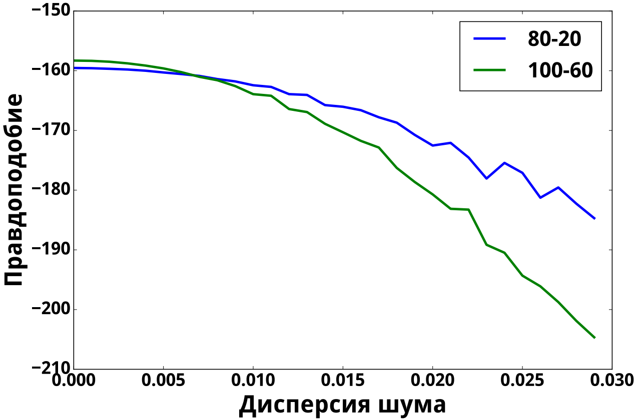
Правдоподобие моделей



Оценка максимальной апостериорной вероятности

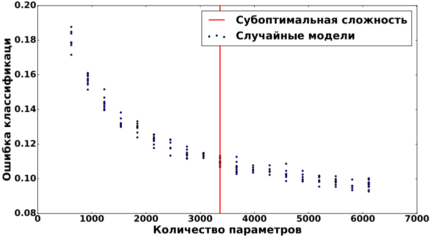
Результат: вариационный автокодировщик

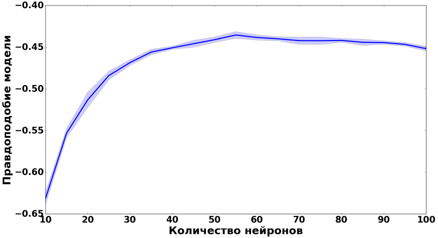
Зависимость правдоподобия выборки от дисперсии шума параметров

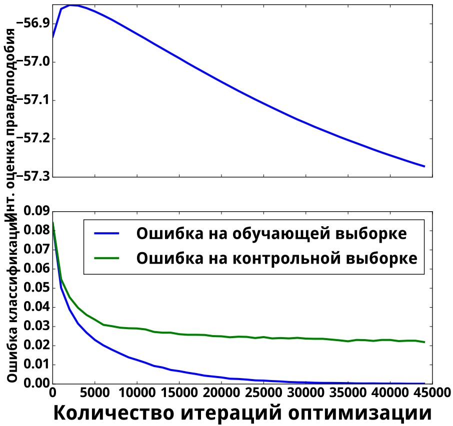


Результат: разделяющая модель $\mathbf{f}_D$

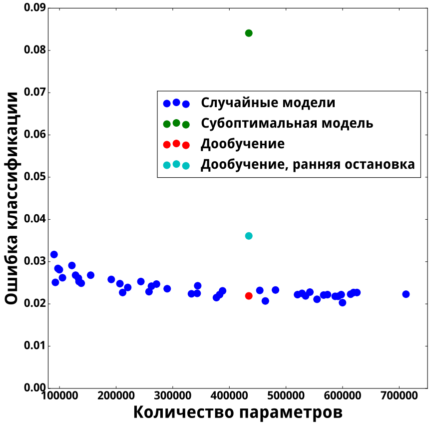
Зависимость правдоподобия модели от количества нейронов







Зависимость ошибки и интегральной оценки от количества итераций



Полученные модели

Заключение

- Предложены критерии субоптимальной сложности модели классификации
- Исследована зависимость интегральной оценки правдоподобия от устойчивости модели и возможности переобучения
- Предложен алгоритм выбора субоптимальной модели классификации без использования кросс-валидации
- Доказана теорема об энтропии распределения под действием градиентного спуска

Публикации:

- Бахтеев О. Ю. Восстановление панельной матрицы и ранжирующей модели по метризованной выборке в разнородных шкалах // Машинное обучение и анализ данных, 2015. Т. 1, 14.
- Бахтеев О.И. Восстановление пропущенных значений в разнородных шкалах с большим числом пропусков // Машинное обучение и анализ данных. 2015. Т. 1, 11.
- Бахтеев О.Ю., Попова М.С., Стрижов В.В. Системы и средства глубокого обучения в задачах классификации // Системы и средства информатики, 2016, 2.
- Бахтеев О.Ю., Стрижов В.В. Последовательное порождение моделей глубокого обучения оптимальной сложности // Заводская лаборатория, диагностика материалов — готовится к печати.