Выбор структуры модели глубокого обучения

Бахтеев Олег

МФТИ

20.11.2019

Резюме прошлых семинаров

Заданы:

- Вариационное распределение $q_{\mathbf{w}}(\mathbf{w}|\mathbf{\Gamma}, \boldsymbol{\theta}_{\mathbf{w}})$ с параметрами $\boldsymbol{\theta}$;
- Априорное распределение $p(\mathbf{w}|\mathbf{\Gamma}, \mathbf{h}, \boldsymbol{\lambda})$ с параметрами $\mathbf{h}$;
- Функция потерь L и функция валидации Q.

Требуется: предложить метод выбора структуры модели $\mathbf{\Gamma}$.

Вопросы:

- Как задать структуру модели?
- Как провести ее выбор?
- Какова вероятностная интерпретация структуры?

Automatic relevance determination

Идея: при оптимизации Evidence *априорное* распределение неинформативных параметров будет сконцентрировано в нуле:

$$\mathbf{w} \sim \mathcal{N}(0, \mathbf{A}), \quad \mathbf{A} = \text{diag}(\boldsymbol{\alpha}).$$

w_i — неинформативен $\rightarrow \alpha_i \approx 0$.

Параллель с вариационным выводом: прунинг параметра w_i определяется относительной плотностью:

$$\lambda = \frac{q(\mathbf{0})}{q(\boldsymbol{\mu}_{i,q})} = \exp(-\frac{\mu_i^2}{2\sigma_i^2}).$$

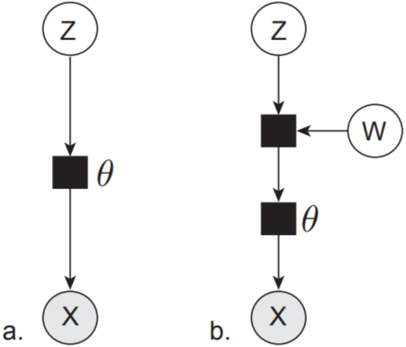
Пример: вариационный автокодировщик + ARD

VAE:

$$L = \int_{\mathbf{z}} p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z}.$$

VAE + ARD:

$$L = \iint_{\mathbf{z},\gamma} p(\mathbf{x}|\mathbf{z} \odot \gamma)p(\mathbf{z})p(\gamma)d\mathbf{z}d\gamma.$$

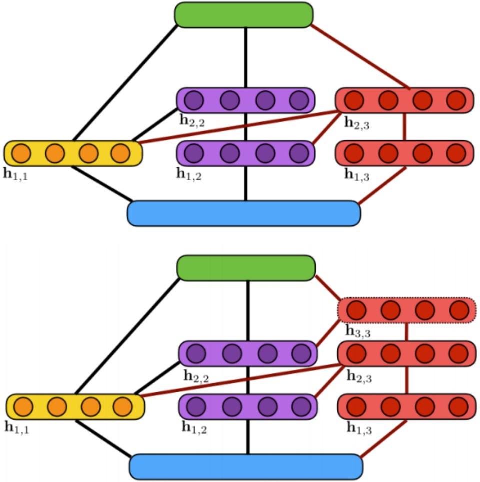


AdaNet

В качестве алгоритма выбора структуры модели выступает бустинговый алгоритм.
На каждом шаге бустинга рассматривается две альтернативы: добавить новую слабую модель той же глубины или более глубокую.
Оптимизируемый функционал:

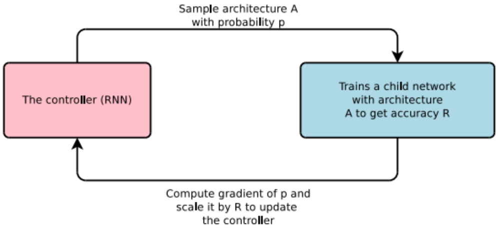
$$Q = \sum_i \Phi(1 - y_i f_{t-1}(\mathbf{x}_i) - y_i \gamma' f'(\mathbf{x})) + \mathcal{R},$$

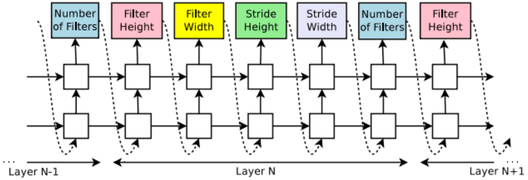
где $\mathcal{R}$ — оценка сложности по Радемахеру.



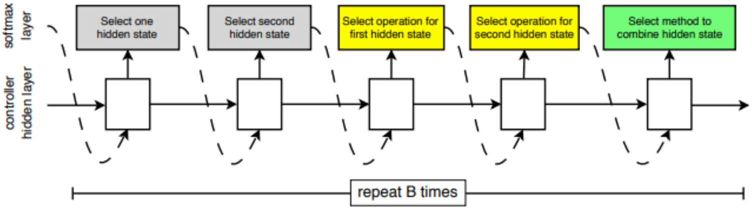
Neural Architecture Search

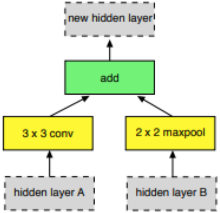
Идея алгоритма: модели порождаются с использованием обучения с подкреплением. Контроллер - рекуррентная нейронная сеть.





Neural Architecture Search





Для оптимизации использовалось 500 GPU.

Neural Architecture Search: результаты

Model	image size	# parameters	Mult-Adds	Top 1 Acc. (%)	Top 5 Acc. (%)
Inception V2 [29]	224×224	11.2 M	1.94 B	74.8	92.2
NASNet-A (5 @ 1538)	**299×299**	**10.9 M**	**2.35 B**	**78.6**	**94.2**
Inception V3 [59]	299×299	23.8 M	5.72 B	78.0	93.9
Xception [9]	299×299	22.8 M	8.38 B	79.0	94.5
Inception ResNet V2 [57]	299×299	55.8 M	13.2 B	80.4	95.3
NASNet-A (7 @ 1920)	**299×299**	**22.6 M**	**4.93 B**	**80.8**	**95.3**
ResNeXt-101 (64 x 4d) [67]	320×320	83.6 M	31.5 B	80.9	95.6
PolyNet [68]	331×331	92 M	34.7 B	81.3	95.8
DPN-131 [8]	320×320	79.5 M	32.0 B	81.5	95.8
SENet [25]	**320×320**	**145.8 M**	**42.3 B**	**82.7**	**96.2**
NASNet-A (6 @ 4032)	**331×331**	**88.9 M**	**23.8 B**	**82.7**	**96.2**

Zoph et al., 2017. Сложность моделей отличается почти в два раза при одинаковом качестве.

Neural Architecture Search: постановка задачи

$\mathbf{w}$ (или $q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma}, \theta_{\mathbf{w}})$) — параметры модели, оптимизируемые при заданной структуре.
$\boldsymbol{\Gamma}$ (или $q_{\boldsymbol{\Gamma}}(\boldsymbol{\Gamma}|\theta_{\boldsymbol{\Gamma}})$) — структура модели, задается контроллером, должна доставлять максимум валидации.

$$\boldsymbol{\Gamma}^* = \arg\max Q(\mathbf{w}^*, \boldsymbol{\Gamma}),$$

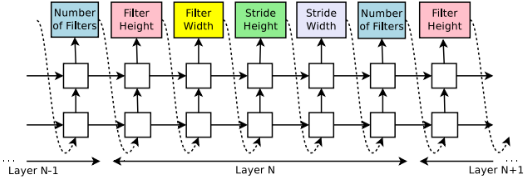
$$\mathbf{w}^* = \arg\max L(\mathbf{w}, \boldsymbol{\Gamma}).$$

Нужно ли здесь обучение с подкреплением?

DARTS

Модель — мультиграф, где ребра $[\mathbf{g}^e]$ соответствуют подмоделям, а вершины $\mathbf{f}_v(\mathbf{x})$ — результату действия подмоделей на выборку. Результат применения подмоделей:

$$\mathbf{f}_v = \langle \gamma, \mathbf{softmax}([\mathbf{g}^e(\mathbf{x})]) \rangle.$$



DARTS

Задача оптимизации:

$$\boldsymbol{\Gamma}^* = \arg\max Q(\mathbf{w}^*, \boldsymbol{\Gamma}),\ Q\text{— ошибка на валидации},$$

$$\mathbf{w}^* = \arg\max L(\mathbf{w}, \boldsymbol{\Gamma}),\ L\text{— ошибка на обучении}.$$

Оптимизация структуры производится жадным градиентным методом:

$$\nabla_{\boldsymbol{\Gamma}} Q(\mathbf{w}', \boldsymbol{\Gamma}) = \lambda_L \nabla_{\boldsymbol{\Gamma},\mathbf{w}} L(\mathbf{w}, \boldsymbol{\Gamma}) \nabla_{\mathbf{w}} Q(\boldsymbol{\Gamma}, \mathbf{w}').$$

Напоминание:
Численное приближение аналитической формулы:

$$\nabla_{\mathbf{h}} Q(\boldsymbol{\theta}^\eta, \mathbf{h}) - \nabla_{\mathbf{h}} \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}^\eta, \mathbf{h})^{\mathsf{T}} \mathbf{H}^{-1} \nabla_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}^\eta, \mathbf{h}).$$

Для быстрого вычисления множителя $\nabla_{\boldsymbol{\Gamma},\mathbf{w}} L(\mathbf{w}, \boldsymbol{\Gamma})$ используется метод конечных приращений.

Графовое представление модели глубокого обучения

Заданы:

1. ациклический граф (V, E);
2. для каждого ребра $(j, k) \in E$: вектор базовых дифференцируемых функций $\mathbf{g}^{j,k} = [\mathbf{g}^{j,k}_0, \dots, \mathbf{g}^{j,k}_{K^{j,k}}]$ мощности $K^{j,k}$;
3. для каждой вершины $v \in V$: дифференцируемая функция агрегации $\mathbf{agg}_v$.
4. Функция $\mathbf{f} = \mathbf{f}_{|V|-1}$, задаваемая по правилу

$$\mathbf{f}_v(\mathbf{w}, \mathbf{x}) = \mathbf{agg}_v \left(\{ \langle \boldsymbol{\gamma}^{j,k}, \mathbf{g}^{j,k} \rangle \circ \mathbf{f}_j(\mathbf{x}) | j \in \text{Adj}(v_k) \} \right), v \in \{1, \dots, |V| - 1\}, \quad \mathbf{f}_0(\mathbf{x}) = \mathbf{x} \qquad (1)$$

и являющаяся функцией из признакового пространства $\mathbb{X}$ в пространство меток $\mathbb{Y}$ при значениях векторов, $\boldsymbol{\gamma}^{j,k} \in [0, 1]^{K^{j,k}}$.

Определение

Граф (V, E) со множестом векторов базовых функций $\{\mathbf{g}^{j,k}, (j, k) \in E\}$ и функций агрегаций $\{\mathbf{agg}_v, v \in V\}$ назовем *параметрическим семейством моделей* $\mathfrak{F}$.

Утверждение

Для любого значения $\boldsymbol{\gamma}^{j,k} \in [0, 1]^{K^{j,k}}$ функция $\mathbf{f} \in \mathfrak{F}$ является моделью.

Выбор структуры: двуслойная нейросеть

Модель $\mathbf{f}$ задана **структурой** $\boldsymbol{\Gamma} = [\boldsymbol{\gamma}^{0,1}, \boldsymbol{\gamma}^{1,2}]$.

$$\text{Модель: } \mathbf{f}(\mathbf{x}) = \mathbf{softmax}\left((\mathbf{w}_0^{1,2})^{\mathsf{T}}\mathbf{f}_1(\mathbf{x})\right), \quad \mathbf{f}(\mathbf{x}) : \mathbb{R}^n \to [0,1]^{|\mathbb{Y}|}, \quad \mathbf{x} \in \mathbb{R}^n.$$

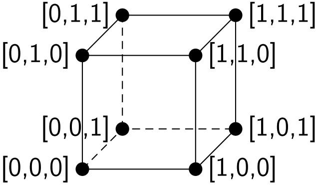
$$\mathbf{f}_1(\mathbf{x}) = \gamma_0^{0,1}\mathbf{g}_0^{0,1}(\mathbf{x}) + \gamma_1^{0,1}\mathbf{g}_1^{0,1}(\mathbf{x}),$$

где $\mathbf{w} = [\mathbf{w}_0^{0,1}, \mathbf{w}_1^{0,1}, \mathbf{w}_0^{1,2}]^{\mathsf{T}}$ — матрицы параметров, $\{\mathbf{g}_{0,1}^{0}, \mathbf{g}_{0,1}^{1}, \mathbf{g}_{1,2}^{0}\}$ — обобщенно-линейные функции скрытых слоев нейросети.

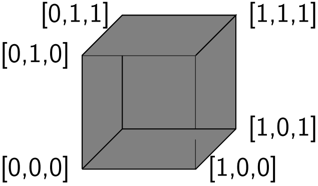
$$\begin{tikzcd}
\mathbf{f}_0(\mathbf{x}) = \mathbf{x} \arrow[r, bend left, "\gamma_0^{0,1}\mathbf{g}_0^{0,1}(\mathbf{x}) = \gamma_0^{0,1}\boldsymbol{\sigma}\left((\mathbf{w}_0^{0,1})^{\mathsf{T}}\mathbf{x}\right)"] \arrow[r, bend right, "\gamma_1^{0,1}\mathbf{g}_1^{0,1}(\mathbf{x}) = \gamma_1^{0,1}\boldsymbol{\sigma}\left((\mathbf{w}_1^{0,1})^{\mathsf{T}}\mathbf{x}\right)"'] & \mathbf{f}_1(\mathbf{x}) \arrow[r, "\gamma_0^{1,2}\mathbf{g}_0^{1,2}(\mathbf{x}) = \gamma_0^{1,2}\mathbf{softmax}\left((\mathbf{w}_0^{1,2})^{\mathsf{T}}\mathbf{x}\right)"] & \mathbf{f}_2(\mathbf{x})
\end{tikzcd}$$

Ограничения на структурные параметры

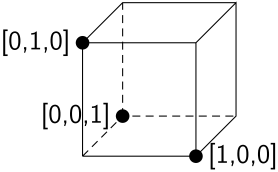
Примеры ограничений для одного структурного параметра $\gamma, |\gamma| = 3$.



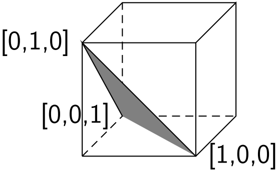
На вершинах куба



Внутри куба



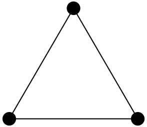
На вершинах симплекса



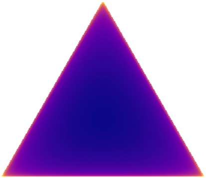
Внутри симплекса

Распределение Дирихле

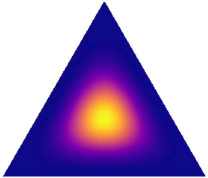
Каждая точка на симплексе задает модель.



$\lambda_{\text{temp}} \to 0$



$\lambda_{\text{temp}} = 0.995$



$\lambda_{\text{temp}} = 5.0$

Репараметризация

Определение

Случайную величину ψ с распределением q с параметрами $\boldsymbol{\theta}_\psi$ назовем репараметризованной через случайную величину ε, чье распределение не зависит от параметров $\boldsymbol{\theta}_\psi$, если:

$$\psi = g(\varepsilon, \boldsymbol{\theta}_\psi)$$

где g — некоторая непрерывная функция.

Пример

$$\mathsf{E}_{q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma},\boldsymbol{\theta}_{\mathbf{w}})} \log\ p(\mathbf{y}|\mathbf{X},\mathbf{w},\boldsymbol{\Gamma}) = \int_{\mathbf{w}} \log\ p(\mathbf{y}|\mathbf{X},\mathbf{w},\boldsymbol{\Gamma}) q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma},\boldsymbol{\theta}_{\mathbf{w}}) d\mathbf{w}.$$

Продифференцируем по параметрам $\boldsymbol{\theta}_{\mathbf{w}}$:

$$\nabla_{\boldsymbol{\theta}_{\mathbf{w}}} \mathsf{E}_{q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma},\boldsymbol{\theta}_{\mathbf{w}})} \log p(\mathbf{y}|\mathbf{X},\mathbf{w},\boldsymbol{\Gamma}) = \int_{\mathbf{w}} \log p(\mathbf{y}|\mathbf{X},\mathbf{w},\boldsymbol{\Gamma}) \nabla_{\boldsymbol{\theta}_{\mathbf{w}}} q_{\mathbf{w}}(\mathbf{w}|\boldsymbol{\Gamma},\boldsymbol{\theta}_{\mathbf{w}}) d\mathbf{w}.$$

Пусть возможна репараметризация: $\mathbf{w} = \mathbf{g}(\boldsymbol{\varepsilon}, \boldsymbol{\theta}_{\mathbf{w}})$. Тогда:

$$\nabla_{\boldsymbol{\theta}_{\mathbf{w}}} \mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}|\boldsymbol{\theta})} \log p(\mathbf{y}|\mathbf{X},\mathbf{w},\boldsymbol{\Gamma}) = \nabla_{\boldsymbol{\theta}_{\mathbf{w}}} \mathsf{E}_{\boldsymbol{\varepsilon}} \log p(\mathbf{y}|\mathbf{X},\mathbf{g}(\boldsymbol{\varepsilon}),\boldsymbol{\Gamma}) =$$

$$= \int_{\boldsymbol{\varepsilon}} \nabla_{\boldsymbol{\theta}_{\mathbf{w}}} \log p(\mathbf{y}|\mathbf{X},\mathbf{g}(\boldsymbol{\varepsilon}),\boldsymbol{\Gamma}) p(\boldsymbol{\varepsilon}) d\boldsymbol{\varepsilon} = \mathsf{E}_{\boldsymbol{\varepsilon}} \nabla_{\boldsymbol{\theta}_{\mathbf{w}}} \log p(\mathbf{y}|\mathbf{X},\mathbf{g}(\boldsymbol{\varepsilon}),\boldsymbol{\Gamma}).$$

Проблема: не всегда просто найти $\mathbf{g}$.

Implicit Reparameterization Gradients

Пусть $\mathbf{g}^{-1}$ — обратная функция к функции $\mathbf{g}$. **Формула полной производной:**

$$\nabla^{\text{total}} f(x_1, \dots, x_n) = \sum \frac{\partial f}{\partial x_i}.$$

Применяем к равенству:

$$\mathbf{g}^{-1}(\mathbf{w}) = \varepsilon.$$

Получаем равенстнво:

$$\nabla_{\boldsymbol{\theta}_{\mathbf{w}}} \mathbf{g}(\varepsilon, \boldsymbol{\theta}_{\mathbf{w}}) = -\left(\nabla_{\mathbf{w}} \mathbf{g}^{-1}(\mathbf{w})\right)^{-1} \nabla_{\boldsymbol{\theta}_{\mathbf{w}}} \mathbf{g}^{-1}(\mathbf{w}).$$

Универсальная функция стандартизации:

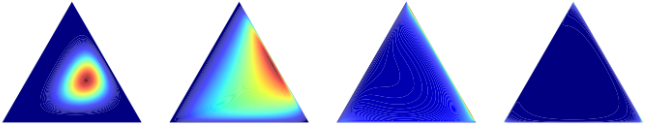
$$\mathbf{g}^{-1}(\mathbf{w}) = F(\mathbf{w}) \sim \mathcal{U}(0, 1),$$

можно использовать методы сэмплирования типа MCMC.

Logit-Normal

$$Z \sim P(\mathcal{N}(\mu, \Sigma))$$

$$Z_k = \frac{\exp(X_k)}{\sum_{i=1}^{K} \exp(X_i)}, X \sim \mathcal{N}(\mu, \Sigma)$$



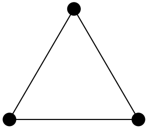
Probability density of $P(\mathcal{N}(0, c \cdot \sigma I))$ for respectively $c = 1, c = 2., c = 3., c = 4.$, and $\sigma = [1, 0.5, 0.7]$

[Источник: Deep Generative Models, http://stat.columbia.edu]

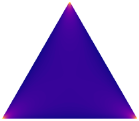
Априорное распределение на структуре модели

Каждая точка на симплексе задает модель.

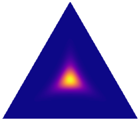
Распределение Гумбель-софтмакс: $\boldsymbol{\Gamma} \sim \text{GS}(\mathbf{s}, \lambda_{\text{temp}})$



$\lambda_{\text{temp}} \to 0$

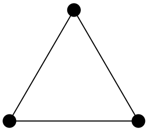


$\lambda_{\text{temp}} = 0.995$

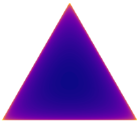


$\lambda_{\text{temp}} = 5.0$

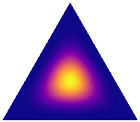
Распределение Дирихле: $\boldsymbol{\Gamma} \sim \text{Dir}(\mathbf{s}, \lambda_{\text{temp}})$



$\lambda_{\text{temp}} \to 0$



$\lambda_{\text{temp}} = 0.995$



$\lambda_{\text{temp}} = 5.0$

Обобщающая задача оптимизации

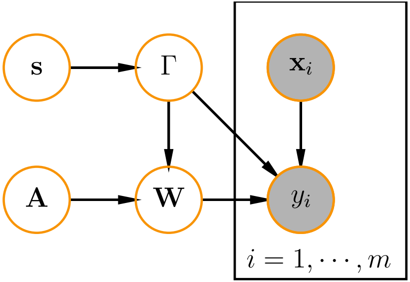
Какие требования можно выдвинуть к “хорошей” функции оптимизации?

1. При некоторых значениях метапараметров функция должна приближать метод максимального правдоподобия.
2. При некоторых значениях метапараметров функция должна штрафовать излишне сложные модели.
3. При некоторых значениях метапараметров функция должна приближать обоснованность модели.
4. При некоторых значениях метапараметров функция должна позволять переходить между оптимальными структурами модели.
5. Функции потерь и валидации должны быть непрерывны по метапараметрам.
6. Область определения функции должна быть нетривиальна.

Вероятностная модель

Базовая модель:

- **параметры** модели
  $\mathbf{w} \sim \mathcal{N}(0, \alpha^{-1})$,
- **гиперпараметры** модели $\mathbf{h} = [\alpha]$.



Предлагаемая модель:

- **параметры** модели
  $\mathbf{w}_r^{j,k} \sim \mathcal{N}(0, (\gamma_r^{j,k})^2(\mathbf{A}_r^{j,k})^{-1})$, $\mathbf{A}_r^{j,k}$ — диагональная матрица параметров, соответствующих базовых функций $\mathbf{g}_r^{j,k}$,
  $(\mathbf{A}_r^{j,k})^{-1} \sim \text{inv-gamma}(\lambda_1, \lambda_2)$,
- **структурные параметры** модели
  $\mathbf{\Gamma} = \{\gamma^{j,k}, (j,k) \in E\}$,
  $\gamma^{j,k} \sim \text{GS}(\mathbf{s}^{j,k}, \lambda_{\text{temp}})$,
- **гиперпараметры** модели $\mathbf{h} = [\text{diag}(\mathbf{A}), \mathbf{s}]$,
- **метапараметры** $\lambda_1, \lambda_2, \lambda_{\text{temp}}$.

Предлагаемая задача оптимизации

Теорема [Бахтеев, 2018]

Пусть функции потерь и валидации L, Q являются непрерывно-дифференцируемыми на компакте U. Тогда следующая задача является обобщающей на U.

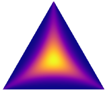
$$\mathbf{h}^* = \arg\max_{\mathbf{h}} Q =$$

$$= \lambda^{\mathrm{Q}}_{\mathrm{likelihood}} \mathsf{E}_{q^*} \log\, p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}, \mathbf{h}, \lambda_{\mathrm{temp}}, \mathbf{f}) -$$
$$-^{\mathrm{prior}}_{\mathrm{Q}} \mathsf{D}_{KL}\bigl(q^*(\mathbf{w}, \mathbf{\Gamma}) || p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}, \lambda_{\mathrm{temp}}, \mathbf{f})\bigr) -$$
$$- \sum_{p' \in \mathfrak{P}, \lambda \in \boldsymbol{\lambda}^{\mathrm{struct}}_{\mathbf{Q}}} \lambda \mathsf{D}_{KL}(\mathbf{\Gamma}|p') + \log p(\mathbf{h}|\mathbf{f}),$$

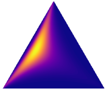
где

$$q^* = \arg\max_{q} L = \mathsf{E}_q \log\, p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}, \mathbf{h}, \lambda_{\mathrm{temp}}, \mathbf{f}) \qquad (L^*)$$
$$-^{\mathrm{prior}}_{\mathrm{L}} \mathsf{D}_{KL}\bigl(q^*(\mathbf{w}, \mathbf{\Gamma}) || p(\mathbf{w}, \mathbf{\Gamma}|\mathbf{h}, \lambda_{\mathrm{temp}}, \mathbf{f})\bigr).$$

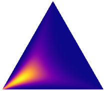
Оптимизационная задача обобщает алгоритмы оптимизации: оптимизация правдоподобия и обоснованности, последовательное увеличение и снижение сложности модели, полный перебор структуры.



$\boldsymbol{\lambda}^{Q}_{\mathrm{struct}} = [0; 0; 0].$



$\boldsymbol{\lambda}^{Q}_{\mathrm{struct}} = [1; 0; 0].$



$\boldsymbol{\lambda}^{Q}_{\mathrm{struct}} = [1; 1; 0].$

Свойства задачи оптимизации

- Коэффициенты при D_{KL} контролируют эффективный размер выборки.
- При $\lambda^{\mathrm{Q}}_{\mathrm{prior}} = \lambda^{\mathrm{L}}_{\mathrm{prior}} = \lambda^{\mathrm{Q}}_{\mathrm{likelihood}} = 1$ — вариационная оценка.
- При $\frac{\lambda^{\mathbf{Q}}_{\mathbf{prior}}}{\lambda^{\mathbf{Q}}_{\mathbf{likelihood}}} = \lambda^{\mathrm{L}}_{\mathrm{prior}}$ — сводится к одноуровневой оптимизации.
- При $\lambda^{\mathrm{Q}}_{\mathrm{prior}} = 0, \lambda^{\mathrm{L}}_{\mathrm{prior}} = 1$ — вариационная оценка обоснованности, при гиперпараметрах, доставляющих максимум правдоподобия.
- При $\lambda^{\mathrm{L}}_{\mathrm{prior}} = 0$ — метод максимального правдоподобия.

Утверждение

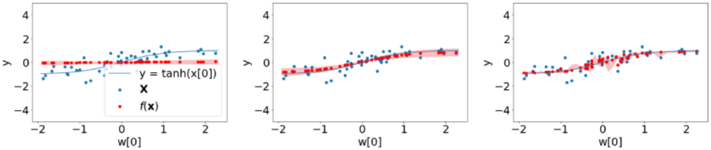
Пусть $\boldsymbol{\lambda}^{Q}_{\mathrm{struct}} = \mathbf{0}$. Пусть $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \mathbf{h}_1, \mathbf{h}_2$ — результаты оптимизации при разных значениях гиперпараметров $\lambda^{\mathrm{Q}}_{\mathrm{prior}_1}, \lambda^{\mathrm{Q}}_{\mathrm{prior}_2}, \lambda^{\mathrm{Q}}_{\mathrm{prior}_1} > \lambda^{\mathrm{Q}}_{\mathrm{prior}_2}$ на компакте U. Пусть функция $Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\lambda})$ является вогнутой на U при $\lambda^{\mathrm{Q}}_{\mathrm{prior}_2}$. Тогда:

$$C_p(\boldsymbol{\theta}_1|U_{\mathbf{h}}, \boldsymbol{\lambda}_1) - C_p(\boldsymbol{\theta}_2|U_{\mathbf{h}}, \boldsymbol{\lambda}_2) < \frac{\lambda^{\mathrm{L}}_{\mathrm{prior}}}{\lambda^{\mathrm{Q}}_{\mathrm{prior}_2}}(\lambda^{\mathrm{Q}}_{\mathrm{prior}_2} - \lambda^{\mathrm{L}}_{\mathrm{prior}})C,$$

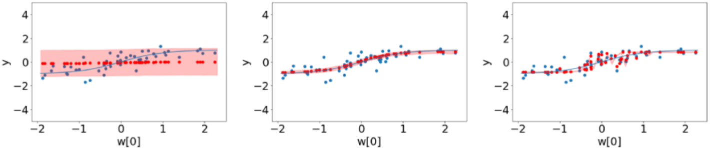
где C — некоторая константа.

Пример

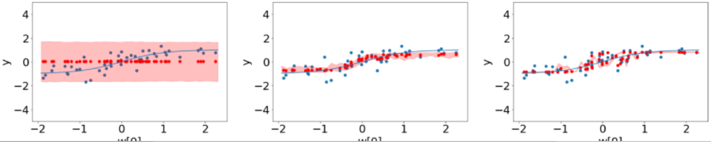
Подмодели: обобщено-линейная с одним признаком, с 11 признаками, константа.



(а) (б) (в)



(г) (д) (е)

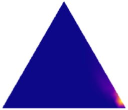


Пример

Подмодели: обобщено-линейная с одним признаком, с 11 признаками, константа.

(а)

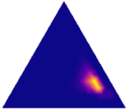
(б)



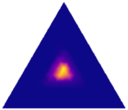
(в)

(г)

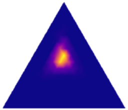
(д)



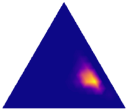
(е)



(ж)



(з)



(и)

Список источников

- MacKay, David JC. "Bayesian nonlinear modeling for the prediction competition." ASHRAE transactions 100.2 (1994): 1053-1062.
- Graves, Alex. "Practical variational inference for neural networks." Advances in neural information processing systems. 2011.
- Karaletsos, Theofanis, and Gunnar Rätsch. "Automatic relevance determination for deep generative models." arXiv preprint arXiv:1505.07765 (2015).
- Cortes, Corinna, et al. "Adanet: Adaptive structural learning of artificial neural networks." Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 2017.
- Zoph, Barret, and Quoc V. Le. "Neural architecture search with reinforcement learning." arXiv preprint arXiv:1611.01578 (2016).
- Zoph, B., Vasudevan, V., Shlens, J. and Le, Q.V., 2018. Learning transferable architectures for scalable image recognition
- Liu, Hanxiao, Karen Simonyan, and Yiming Yang. "Darts: Differentiable architecture search." arXiv preprint arXiv:1806.09055 (2018).

Список источников

- Figurnov M., Mohamed S., Mnih A. Implicit reparameterization gradients //Advances in Neural Information Processing Systems. – 2018. – С. 441-452.
- http://stat.columbia.edu/~cunningham/teaching/GR8201/STAT_GR8201_2019_SPRG_slides_lec03.pdf
- Jang, Eric, Shixiang Gu, and Ben Poole. "Categorical reparameterization with gumbel-softmax." arXiv preprint arXiv:1611.01144 (2016).
- Maddison, Chris J., Andriy Mnih, and Yee Whye Teh. "The concrete distribution: A continuous relaxation of discrete random variables." arXiv preprint arXiv:1611.00712 (2016).

ДЗ: выбор задания

Дедлайн: 27 ноября, 0 часов.

```
from zlib import crc32

theory = crc32('фамилия кириллицей'.lower().encode('utf-8'))%2+1

practice = 1
```

Задания заливаются на github:
https://github.com/Intelligent-Systems-Phystech/model_selection/фамилия латиницей

ДЗ: теория

Формат: tex + pdf.
Задание 1:
Доказать, что при устремлении параметра температуры к бесконечности, плотность Gumbel-Softmax концентрируется в центре симплекса.
(за формулами Gumbel-Softmax обращаться к оригинальным статьям, Jang et al., Maddison et al.).

ДЗ: теория

Формат: tex + pdf.
Задание 2: Доказать утверждение:

утверждение

Пусть задан компакт $U = U_{\mathbf{h}} \times U_{\boldsymbol{\theta}}$ и $\boldsymbol{\lambda}^{\mathrm{Q}}_{\mathrm{struct}} = \mathbf{0}$. Пусть решение задачи

$$\min_{\mathbf{h} \in U_{\mathbf{h}}} D_{\mathrm{KL}}\bigl(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_2)||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h})\bigr)$$

является единственным для некоторых $\lambda^{\mathrm{Q}}_{\mathrm{prior}_1}, \lambda^{\mathrm{Q}}_{\mathrm{prior}_2}, \lambda^{\mathrm{Q}}_{\mathrm{prior}_1} > \lambda^{\mathrm{Q}}_{\mathrm{prior}_2}$ на U при некоторых фиксированных $\lambda^{\mathrm{Q}}_{\mathrm{likelihood}}, \lambda^{\mathrm{L}}_{\mathrm{prior}}, \lambda_{\mathrm{temp}}$. Пусть также решения задач из слайда 20 являются единственными на U при $\lambda^{\mathrm{Q}}_{\mathrm{prior}_1}, \lambda^{\mathrm{Q}}_{\mathrm{prior}_2}$ и $\lambda^{\mathrm{Q}}_{\mathrm{likelihood}}, \lambda^{\mathrm{L}}_{\mathrm{prior}}, \lambda_{\mathrm{temp}}$. Тогда справедливо следующее неравенство:

$$D_{\mathrm{KL}}\bigl(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_1)||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}_1)\bigr) < D_{\mathrm{KL}}\bigl(q(\mathbf{w}, \boldsymbol{\Gamma}|\boldsymbol{\theta}_2)||p(\mathbf{w}, \boldsymbol{\Gamma}|\mathbf{h}_2)\bigr),$$

где $\mathbf{h}_1, \boldsymbol{\theta}_1$, $\mathbf{h}_2, \boldsymbol{\theta}_2$ — решения задачи при $\lambda^{\mathrm{Q}}_{\mathrm{prior}_1}, \lambda^{\mathrm{Q}}_{\mathrm{prior}_2}$,

$$\boldsymbol{\theta}_1 = \boldsymbol{\theta}^*(\mathbf{h}_1), \quad \boldsymbol{\theta}_2 = \boldsymbol{\theta}^*(\mathbf{h}_2).$$

ДЗ: практика

Формат: ipynb. Реализовать визуализация зависимосcти распределения Gumbel-Softmax от температуры.
Пример для распределения Дирихле:
http://blog.bogatron.net/blog/2014/02/02/visualizing-dirichlet-distributions/
При оценивании будут учитываться аккуратность кода ноутбуков и наглядность примера.

Пример должен быть выполнен на **простых** игрушечных синтетических данных.