

Выбор структуры модели глубокого обучения

Бахтеев Олег

МФТИ

20.11.2019

Резюме прошлых семинаров

Заданы:

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

Требуется: предложить метод выбора структуры модели Γ .

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:

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

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

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

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

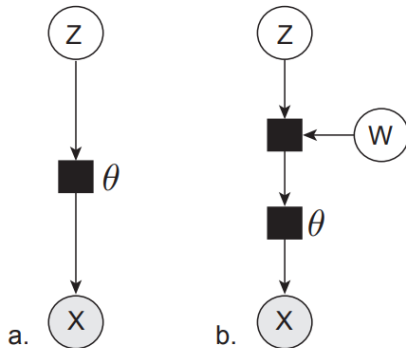
Пример: вариационный автокодировщик + ARD

VAE:

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

VAE + ARD:

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



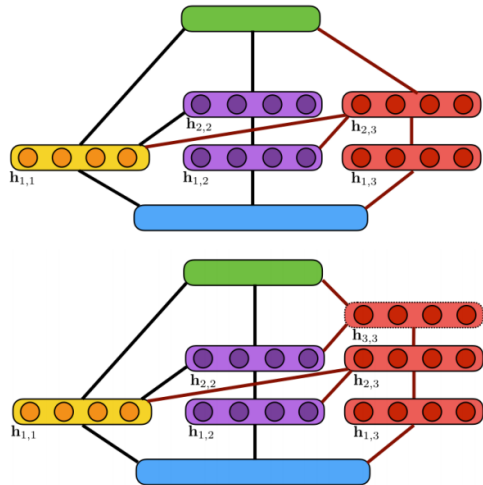
AdaNet

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:

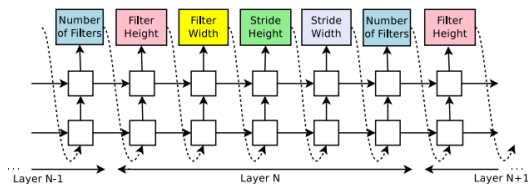
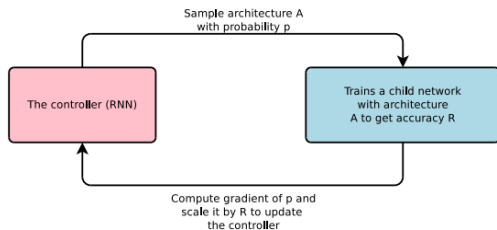
$$Q = \sum_i \Phi(1 - y_i f_{t-1}(\mathbf{x}_i) - y_i \gamma' f'(\mathbf{x})) + \mathcal{R},$$

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

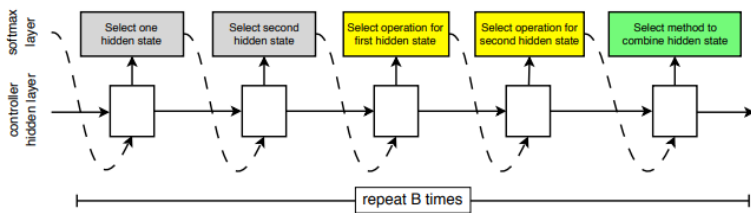


Neural Architecture Search

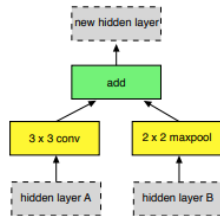
Идея алгоритма: модели порождаются с использованием обучения с подкреплением. Контроллер - рекуррентная нейронная сеть.



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 $q_{\mathbf{w}}(\mathbf{w}|\Gamma, \theta_{\mathbf{w}})$) — параметры модели, оптимизируемые при заданной структуре.

Γ (или $q_{\Gamma}(\Gamma|\theta_{\Gamma})$) — структура модели, задается контроллером, должна доставлять максимум валидации.

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

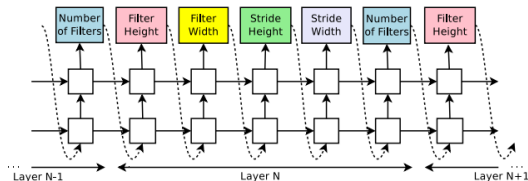
$$\mathbf{w}^* = \arg \max L(\mathbf{w}, \Gamma).$$

Нужно ли здесь обучение с подкреплением?

DARTS

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

$$f_v = \langle \gamma, \text{softmax}([g^e(x)]) \rangle.$$



DARTS

Задача оптимизации:

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

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

Оптимизация структуры производится жадным градиентным методом:

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

Напоминание:

Численное приближение аналитической формулы:

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

Для быстрого вычисления множителя $\nabla_{\Gamma, \mathbf{w}} L(\mathbf{w}, \Gamma)$ 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 $\mathbf{g}^{j,k} = [\mathbf{g}_0^{j,k}, \dots, \mathbf{g}_{K^{j,k}}^{j,k}]$ 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

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

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

Определение

Граф (V, E) со множеством 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 $\gamma^{j,k} \in [0, 1]^{K^{j,k}}$ функция $\mathbf{f} \in \mathfrak{F}$ является моделью.

Выбор структуры: двуслойная нейросеть

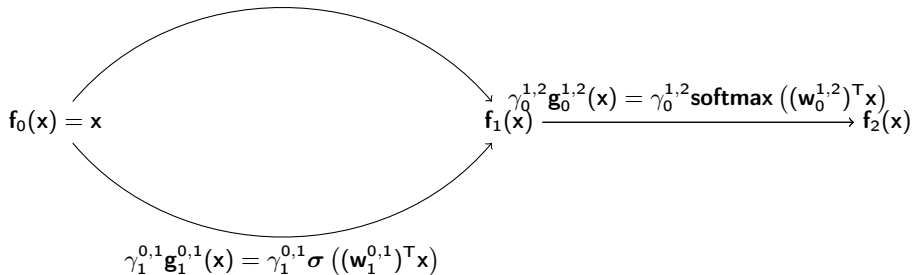
Модель f задана структурой $\Gamma = [\gamma^{0,1}, \gamma^{1,2}]$.

$$\text{Модель: } f(x) = \text{softmax} \left((w_0^{1,2})^T f_1(x) \right), \quad f(x) : \mathbb{R}^n \rightarrow [0, 1]^{|Y|}, \quad x \in \mathbb{R}^n.$$

$$f_1(x) = \gamma_0^{0,1} g_0^{0,1}(x) + \gamma_1^{0,1} g_1^{0,1}(x),$$

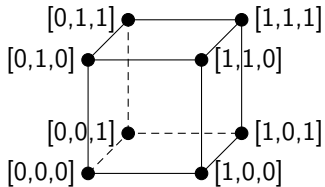
где $w = [w_0^{0,1}, w_1^{0,1}, w_0^{1,2}]^T$ — матрицы параметров, $\{g_0^0, g_0^1, g_1^0, g_1^1\}$ — обобщенно-линейные функции скрытых слоев нейросети.

$$\gamma_0^{0,1} g_0^{0,1}(x) = \gamma_0^{0,1} \sigma \left((w_0^{0,1})^T x \right)$$

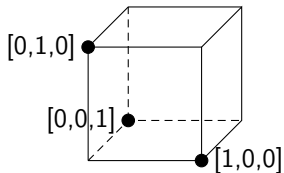


Ограничения на структурные параметры

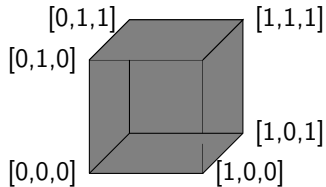
Примеры ограничений для одного структурного параметра γ , $|\gamma| = 3$.



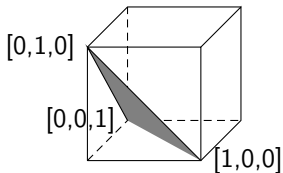
На вершинах куба



На вершинах симплекса



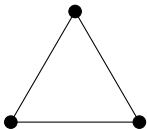
Внутри куба



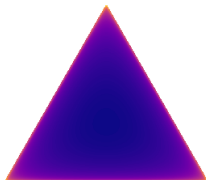
Внутри симплекса

Распределение Дирихле

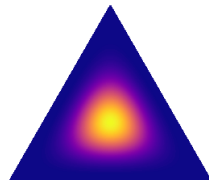
Каждая точка на симплексе задает модель.



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



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



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

Репараметризация

Определение

Случайную величину ψ с распределением q с параметрами θ_ψ назовем репараметризованной через случайную величину ε , чье распределение не зависит от параметров θ_ψ , если:

$$\psi = g(\varepsilon, \theta_\psi)$$

где g — некоторая непрерывная функция.

Пример

$$E_{q_w(w|\Gamma, \theta_w)} \log p(y|X, w, \Gamma) = \int_w \log p(y|X, w, \Gamma) q_w(w|\Gamma, \theta_w) dw.$$

Продифференцируем по параметрам θ_w :

$$\nabla_{\theta_w} E_{q_w(w|\Gamma, \theta_w)} \log p(y|X, w, \Gamma) = \int_w \log p(y|X, w, \Gamma) \nabla_{\theta_w} q_w(w|\Gamma, \theta_w) dw.$$

Пусть возможна репараметризация: $w = g(\varepsilon, \theta_w)$. Тогда:

$$\begin{aligned} \nabla_{\theta_w} E_{q(w, \Gamma|\theta)} \log p(y|X, w, \Gamma) &= \nabla_{\theta_w} E_\varepsilon \log p(y|X, g(\varepsilon), \Gamma) = \\ &= \int_\varepsilon \nabla_{\theta_w} \log p(y|X, g(\varepsilon), \Gamma) p(\varepsilon) d\varepsilon = E_\varepsilon \nabla_{\theta_w} \log p(y|X, g(\varepsilon), \Gamma). \end{aligned}$$

Проблема: не всегда просто найти 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 равенство:

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

Универсальная функция стандартизации:

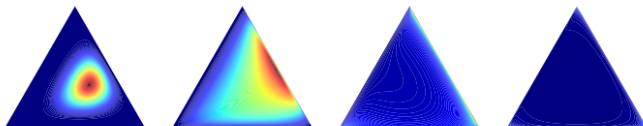
$$\mathbf{g}^{-1}(\mathbf{w}) = F(\mathbf{w}) \sim \mathcal{U}(0, 1),$$

можно использовать методы сэмплирования типа MCMC.

Logit-Normal

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

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



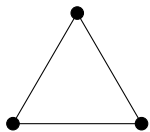
Probability density of $P(\mathcal{N}(0, c \cdot \sigma I))$ for respectively $c = 1, c = 2., c = 3., c = 4.,$ and
 $\sigma = [1, 0.5, 0.7]$

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

Априорное распределение на структуре модели

Каждая точка на симплексе задает модель.

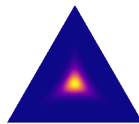
Распределение Гумбель-софтмакс: $\Gamma \sim \text{GS}(\mathbf{s}, \lambda_{\text{temp}})$



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

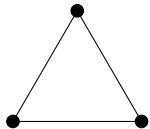


$$\lambda_{\text{temp}} = 0.995$$



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

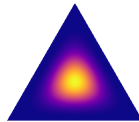
Распределение Дирихле: $\Gamma \sim \text{Dir}(\mathbf{s}, \lambda_{\text{temp}})$



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



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



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

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 .

$$\begin{aligned} \mathbf{h}^* &= \arg \max_{\mathbf{h}} Q = \\ &= \lambda_{\text{likelihood}}^Q E_{q^*} \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \Gamma, \mathbf{h}, \lambda_{\text{temp}}, \mathbf{f}) - \\ &\quad - \lambda_Q^{\text{prior}} D_{KL}(q^*(\mathbf{w}, \Gamma) || p(\mathbf{w}, \Gamma | \mathbf{h}, \lambda_{\text{temp}}, \mathbf{f})) - \\ &\quad - \sum_{p' \in \mathfrak{P}, \lambda \in \lambda_Q^{\text{struct}}} \lambda D_{KL}(\Gamma | p'), \end{aligned}$$

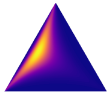
где

$$\begin{aligned} q^* &= \arg \max L = E_q \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \Gamma, \mathbf{h}, \lambda_{\text{temp}}, \mathbf{f}) \\ &\quad - \lambda_L^{\text{prior}^q} D_{KL}(q^*(\mathbf{w}, \Gamma) || p(\mathbf{w}, \Gamma | \mathbf{h}, \lambda_{\text{temp}}, \mathbf{f})). \end{aligned} \quad (L^*)$$

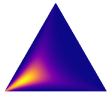
Оптимизационная задача обобщает алгоритмы оптимизации: оптимизация правдоподобия и обоснованности, последовательное увеличение и снижение сложности модели, полный перебор структуры.



$$\lambda_{\text{struct}}^Q = [0; 0; 0].$$



$$\lambda_{\text{struct}}^Q = [1; 0; 0].$$



$$\lambda_{\text{struct}}^Q = [1; 1; 0].$$

Свойства задачи оптимизации

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

Утверждение

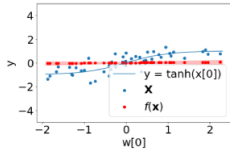
Пусть $\lambda_{struct}^Q = 0$. Пусть $\theta_1, \theta_2, \mathbf{h}_1, \mathbf{h}_2$ — результаты оптимизации при разных значениях гиперпараметров $\lambda_{prior_1}^Q, \lambda_{prior_2}^Q, \lambda_{prior_1}^Q > \lambda_{prior_2}^Q$ на компакте U . Пусть функция $Q(\mathbf{h}|\mathbf{y}, \mathbf{X}, \theta, \lambda)$ является вогнутой на U при $\lambda_{prior_2}^Q$. Тогда:

$$C_p(\theta_1|U_h, \lambda_1) - C_p(\theta_2|U_h, \lambda_2) < \frac{\lambda_{prior}^L}{\lambda_{prior_2}^Q} (\lambda_{prior_2}^Q - \lambda_{prior}^L) C,$$

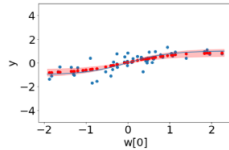
где C — некоторая константа.

Пример

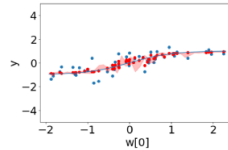
Подмодели: обобщено-линейная с одним признаком, с 11 признаками, константа.



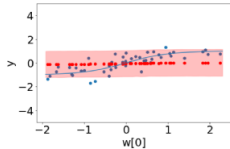
(a)



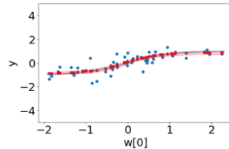
(б)



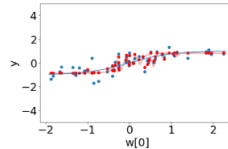
(в)



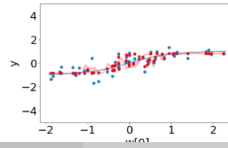
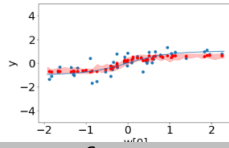
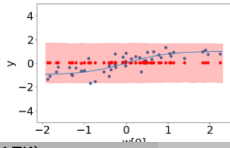
(г)



(д)

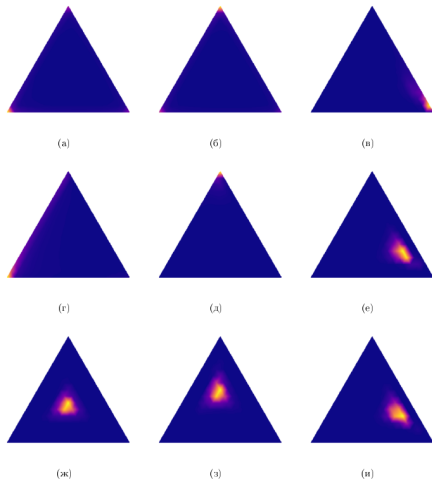


(е)



Пример

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



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 $U = U_h \times U_\theta$ и $\lambda_{\text{struct}}^Q = \mathbf{0}$. Пусть решение задачи

$$\min_{\mathbf{h} \in U_h} D_{\text{KL}}(q(\mathbf{w}, \Gamma | \theta_2) || p(\mathbf{w}, \Gamma | \mathbf{h}))$$

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

$$D_{\text{KL}}(q(\mathbf{w}, \Gamma | \theta_1) || p(\mathbf{w}, \Gamma | \mathbf{h}_1)) < D_{\text{KL}}(q(\mathbf{w}, \Gamma | \theta_2) || p(\mathbf{w}, \Gamma | \mathbf{h}_2)),$$

где $\mathbf{h}_1, \theta_1, \mathbf{h}_2, \theta_2$ — решения задачи при $\lambda_{\text{prior}_1}^Q, \lambda_{\text{prior}_2}^Q$,

$$\theta_1 = \theta^*(\mathbf{h}_1), \quad \theta_2 = \theta^*(\mathbf{h}_2).$$

ДЗ: практика

Формат: ірупв. Реализовать визуализация зависимости 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.