1 Постановка задачи

Задана выборка
$$\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}, i = 1, \dots, m, \tag{1}$$
состоящая из множества пар «объект-метка»
$$\mathbf{x}_i \in \mathbf{X} \subset \mathbb{R}^n, \quad y_i \in \mathbf{y} \subset \mathbb{Y}.$$

Метка y объекта $\mathbf{x}$ принадлежит либо множеству: $y \in \mathbb{Y} = \{1, \dots, Z\}$ в случае задачи классификации, где Z — число классов, либо некоторому подмножеству вещественных чисел $y \in \mathbb{Y} \subseteq \mathbb{R}$ в случае задачи регрессии. Определим множество архитектур моделей глубокого обучения для дальнейшего выбора оптимальной.

Пусть задан граф V, E. Пусть для каждого ребра $< i, j > \in E$ определено множество функций $\mathbf{o}(i, j)$. Граф V, E с множеством функций $\mathbf{O}$ называется моделью, если функция, задаваемая рекурсивно как
$$f_i(\mathbf{x}) = \sum_{j \in \mathrm{Adj}(v_i)} o(i, j)(f_j(\mathbf{x})),$$
является непрерывной дифференцируемой функцией из $\mathbb{R}^n$ во множество $\mathbb{Y}$ при любом $o(i, j)$, являющемся линейной комбинацией функций из множества $\mathbf{o}(i, j)$.

Пусть $\mathbf{w}$ — множество всех параметров функций из $\mathbf{o}(i, j), < i, j > \in E$. Положим распределение параметров $\mathbf{w}$ нормальным с нулевым средним и диагональной квариационной матрицей:
$$\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{A}^{-1}).$$

Пусть для каждого ребра i, j задан нормированный положительный вектор $\boldsymbol{\gamma}_{i,j} \in \mathbb{R}+^{|\mathbf{o}(i,j)|}$, определяющий веса функций из множества $\mathbf{o}(i, j)$. Будем считать, что вектор $\boldsymbol{\gamma}_{i,j}$ распределен по распределению Дирихле:
$$\boldsymbol{\gamma}_{i,j} \sim \mathrm{Dir}(c, \mathbf{m}_{i,j}).$$
где c — вектор концентрации распределения, $\mathbf{m}_{i,j}$ — вектор средних. Обозначим за структуру модели $\boldsymbol{\Gamma}$ множество всех векторов $\boldsymbol{\gamma}$.

Пусть также определено правдоподобие выборки $p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma})$.

Определение Правдоподобием модели $\mathbf{f}$ назовем следующее выражение:
$$p(\mathbf{y}|\mathbf{X}, \mathbf{A}, \mathbf{m}, c) = \int_{\mathbf{w}, \boldsymbol{\Gamma}} p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \boldsymbol{\Gamma}) p(\mathbf{w}|\mathbf{A}) p(\boldsymbol{\Gamma}|\mathbf{m}, c) d\mathbf{w} d\boldsymbol{\Gamma}. \tag{2}$$

Пусть задано значение концентрации c. Требуется найти гиперпараметры модели $\mathbf{A}, \mathbf{m}$ доставляющие максимум правдоподобия модели:
$$\arg\max_{\mathbf{A}, \mathbf{m}} p(\mathbf{y}|\mathbf{X}, \mathbf{A}, \mathbf{m}, c).$$

Утверждение (предварительно). При $c << 0$ оптимизация (2) эквивалентна оптимизации дискретной оптимизации: $\boldsymbol{\gamma}_{i,j} \in 2^{|\mathbf{o}(i,j)|}$.

2 Вариационная постановка задачи

В общем виде вычисление значения интеграла (2) является вычислительно сложной процедурой. В качестве приближенного значения интеграла будем использовать вариационную верхнюю оценку правдоподобия модели. Пусть заданы непрерывные параметрические распределения q_w, q_γ, аппроксимирующие апостериорные распределения $p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \mathbf{A}, \mathbf{m}, c)$, $p(\mathbf{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{A}, \mathbf{m}, c)$. Тогда верно следующее выражение:

$$\text{log}p(\mathbf{y}|\mathbf{X}, \mathbf{A}, \mathbf{m}, c) \leq \mathsf{E}_{q_w, q_\gamma}\text{log}p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{\Gamma}, \mathbf{A}, \mathbf{m}, c) - D_{KL}(q_\gamma||p(\mathbf{\Gamma})) - D_{KL}(q_w||p(\mathbf{w})). \quad (3)$$

Разница между верхней оценкой (3) и правдоподобием модели (2) определяется дивергенцией между апостериорными распределениями $p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \mathbf{A}, \mathbf{m}, c)$, $p(\mathbf{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{A}, \mathbf{m}, c)$. и вариационными распределениями q_w, q_γ.

Сформулируем основные требования к оптимизационной задаче и оптимизируемым функционалам:

1. Оптимизируемые функции должны быть дифференцируемы.
2. Оптимизация должна приводить к максимуму вариационной оценки.
3. Оптимизация должна позволять калибровать количество эффективных параметров
4. Оптимизация должна позволять калибровать количество эффекитвных ребер.
5. Оптимизация должна позволять проводить полный перебор структуры.

Положим $\boldsymbol{\theta}$ равным параметрам распределений q_w, q_γ. Положим $\mathbf{h} = [\mathbf{A}, \mathbf{m}]$.

Пусть L — приближенное значение вариационной оценки правдоподобия:

$$L = \text{log}p(\mathbf{y}|\hat{\mathbf{w}}, \hat{\mathbf{\Gamma}}) - D_{KL}(q_\gamma||p(\mathbf{\Gamma})) - D_{KL}(q_w||p(\mathbf{w})),$$

где $\hat{\mathbf{w}} \sim q_w, \quad \hat{\mathbf{\Gamma}} \sim q_\gamma$.

Пусть Q — валидационная функция:

$$Q(c, c_1, c_2, c_3, \mathbf{p}) = c_1\text{log}p(\mathbf{y}||\hat{\mathbf{w}}, \hat{\mathbf{\Gamma}}) + c_2[-D_{KL}(q_\gamma||p(\mathbf{\Gamma})) - D_{KL}(q_w||p(\mathbf{w}))] +$$

$$+ c_3 \sum_{p_k \in \mathbf{p}} D_{KL}(q_\gamma||p_k),$$

где $\mathbf{p}$ — заданные распределения на структурах,c_1, c_2, c_3 — коэффициенты.

Сформулируем задачу поиска оптимальной модели как двухуровневую задачу.

$$\hat{\mathbf{h}} = \arg\max_{\mathbf{h} \in \mathbb{R}^h} Q(T^\eta(\boldsymbol{\theta}_0, \mathbf{h})), \quad (4)$$

где T — оператор оптимизации, решающий задачу оптимизации:

$$L(T^\eta(\boldsymbol{\theta}_0, \mathbf{h})) \to \max .$$

Вопрос: в последнем слагаемом априорные или вариационные распределения.

Утверждение. Пусть $D_{KL}(q_w|p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \mathbf{A}, \mathbf{m}, c)) = 0, D_{KL}(q_\gamma|p(\mathbf{\Gamma}|\mathbf{y}, \mathbf{X}, \mathbf{A}, \mathbf{m}, c)) = 0$, пусть $c_1 = 1, c_2 = 1, c_3 = 0$. Тогда оптимизация (4) эквивалентна оптимизации (2).

Определение (предварительно)Параметрической δ-сложностью модели назовем матожидание следующей величины:

$$C_{\mathrm{p}}(\delta, \mathbf{w}) = \mathsf{E} \sum_{w \in \mathbf{w}} I(|w| > \delta).$$

Определение (предварительно) Структурной δ-сложностью модели назовем матожидание следующей величины:

$$C_{\mathrm{s}}(\delta, \mathbf{\Gamma}) = \mathsf{E} \sum_{\boldsymbol{\gamma} \in \mathbf{\Gamma}} \sum_{\gamma_i \in \boldsymbol{\gamma}} I(\gamma_i > \delta).$$

Утверждение (предварительно). Пусть $c_1 = 1, c_3 = 0, c_2 > 0, c_2' < c_2$. Пусть $\mathbf{w}, \mathbf{w}'$ — параметры, полученные в результате соответствующих оптимизаций. Тогда $C_{\mathrm{p}}(\delta, \mathbf{w}') \leq C_{\mathrm{p}}(\delta, \mathbf{w})$.

Идея доказательства: для примера: пусть вар. распределение — нормальное. При снижении c_2 до нуля получаем $\mathbf{A}_q \to \infty$.

Утверждение (предварительно). Пусть $c' < c$. Пусть $\mathbf{\Gamma}, \mathbf{\Gamma}'$ — параметры, полученные в результате соответствующих оптимизаций. Тогда $C_{\mathrm{s}}(\delta, \mathbf{w}') \leq C_{\mathrm{s}}(\delta, \mathbf{w})$.

Утверждение (предварительно, нужно развить). Пусть $c_3 > 0, c << 0$ и все $p_k \in \mathbf{p}$ отражают распределения на вершинах симплекса. Тогда оптимизация приведет к q_γ, сконцентрированному на одной из остальных вершин симплекса.

Утверждение (очень предварительно). Изменение c позволяет избежать ухода в локальный минимум.

Утверждение (очень предварительно). Изменение c_2 позволяет избежать ухода в локальный минимум.

Утверждение (предварительно). Пусть $c_1 = c_2 = c_3 = 0$. Пусть $q_w \sim \mathcal{N}(\mathbf{0}, \sigma), \sigma \sim 0$. Тогда оптимизация эквивалентна обычной оптимизации параметров с l_2 - регуляризацией.

Далее будем рассматривать $q_w \sim \mathcal{N}(\mathbf{0}, \mathbf{A}_q^{-1}), \quad q_\gamma \sim \text{Gumbel-Softmax}(\mathbf{g}, \tau)$.