

1 Постановка задачи

Задана выборка

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

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 $\langle i, j \rangle \in E$ определено множество функций $\mathbf{o}(i, j)$. Граф V, E с множеством функций \mathbf{O} называется моделью, если функция, задаваемая рекурсивно как

$$f_i(\mathbf{x}) = \sum_{j \in \text{Adj}(v_i)} o(i, j)(f_j(\mathbf{x})),$$

является непрерывной дифференцируемой функцией из \mathbb{R}^n во множество \mathbb{Y} при любом $\mathbf{o}(i, j)$, являющемся линейной комбинацией функций из множества $\mathbf{o}(i, j)$.

Пусть \mathbf{w} — множество всех параметров функций из $\mathbf{o}(i, j), \langle i, j \rangle \in E$. Положим распределение параметров \mathbf{w} нормальным с нулевым средним и диагональной ковариационной матрицей:

$$\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{A}^{-1}).$$

Пусть для каждого ребра i, j задан нормированный положительный вектор $\gamma_{i,j} \in \mathbb{R}^{+|\mathbf{o}(i,j)|}$, определяющий веса функций из множества $\mathbf{o}(i, j)$. Будем считать, что бинарный вектор $\gamma_{i,j}$ распределен по распределению Дирихле:

$$\gamma_{i,j} \sim \text{Dir}(c, \mathbf{m}_{i,j}).$$

где $\mathbf{c}_{i,j}$ — вектор концентрации распределения, $\mathbf{m}_{i,j}$ — вектор средних. Обозначим за структуру модели Γ множество всех векторов γ .

Пусть также определено правдоподобие выборки $p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \gamma)$. **Определение** Правдоподобием модели \mathbf{f} назовем следующее выражение:

$$p(\mathbf{y}|\mathbf{X}, \mathbf{A}, \mathbf{m}, c) = \int_{\mathbf{w}, \Gamma} p(\mathbf{y}|\mathbf{X}, \mathbf{W}, \Gamma) p(\mathbf{w}|\mathbf{A}) p(\gamma|\mathbf{m}, c) d\mathbf{W} d\Gamma. \quad (2)$$

Пусть задано значение концентрации c . Требуется найти гиперпараметры модели \mathbf{A}, \mathbf{m} доставляющие максимум правдоподобия модели:

$$\arg \max_{\mathbf{A}, \mathbf{m}} \in \log p(\mathbf{y}|\mathbf{X}, \mathbf{A}, \mathbf{m}, c).$$

Хочется: чтобы c находилось автоматически, см. spike-and-slab.

Утверждение (предварительно). При $c \ll 0$ оптимизация (2) эквивалентна оптимизации при ограничениях $\gamma_{i,j} \in 2^{|\mathbf{o}(i,j)|}$.

2 Вариационная постановка задачи

Пусть заданы распределения q_w, q_γ , аппроксимирующие апостериорные распределения $p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \mathbf{A}, \mathbf{m}, c), p(\Gamma|\mathbf{y}, \mathbf{X}, \mathbf{A}, \mathbf{m}, c)$.

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

Пусть L — вариационная оценка правдоподобия:

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

Пусть Q — валидационная функция:

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

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

Сформулируем задачу поиска оптимальной модели как двухуровневую задачу.

$$\hat{\mathbf{h}} = \arg \max_{\mathbf{h} \in \mathbb{R}^h} Q(T^\eta(\theta_0, \mathbf{h})), \quad (3)$$

где T — оператор оптимизации, решающий задачу оптимизации:

$$L(T^\eta(\theta_0, \mathbf{h})) \rightarrow \min.$$

Вопрос: в последнем слагаемом априорные или вариационные распределения.

Утверждение. Пусть $D_{KL}(q_w||p(\mathbf{w})) = 0, D_{KL}(q_\gamma||p(\Gamma)) = 0$, пусть $c_1 = 1, c_2 = 1, c_3 = 0$. Тогда оптимизация (3) эквивалентна оптимизации (2).

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

$$C_p(\delta, \mathbf{w}) = \mathbb{E} \sum_{w \in \mathbf{w}} I(|w| > \delta).$$

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

$$C_s(\delta, \Gamma) = \mathbb{E} \sum_{\gamma \in \Gamma} \sum_{\gamma_i \in \gamma} I(\gamma_i > \delta).$$

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 $C_p(\delta, \mathbf{w}') \leq C_p(\delta, \mathbf{w})$.

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

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

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 $q_w \sim \mathcal{N}(\mathbf{0}, \mathbf{A}_q^{-1}), q_\gamma \sim \text{Gumbel-Softmax}(\mathbf{g}, \tau)$. **Проблема:** Gumbel-Softmax близко к распределению Дирихле. В случае оптимизации Evidence ($c_1 = c_2 = 1, c_3 = 0$) оптимизация вырождается к $q_\gamma = p_\gamma$.