1 Введение

Задана выборка:

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

где $\mathbf{x}_i \in \mathbb{R}^m$, $\mathbf{y}_i \in \{1, \ldots, Y\}$, Y — число классов. Рассмотрим модель

$$\mathbf{f}(\mathbf{x}, \mathbf{w}) : \mathbb{R}^m \times \mathbb{R}^n \longrightarrow \mathbb{R}^Y,$$

где $\mathbf{w} \in \mathbb{R}^n$ — пространство параметров модели. Пусть задано априорное распределение вектора параметров в пространстве $\mathbb{R}^n$:

$$p(\mathbf{w}) \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{A}_{\text{pr}}^{-1}),$$

где $\boldsymbol{\mu}, \mathbf{A}_{\text{pr}}^{-1}$ — вектор средних и матрица ковариации априорного распределения. Пусть также задано вариационное распределение

$$q(\mathbf{w}) \sim \mathcal{N}(\mathbf{m}, \mathbf{A}_{\text{ps}}^{-1}).$$

Здесь $\mathbf{m}, \mathbf{A}_{\text{ps}}^{-1}$ — вектор средних и матрица ковариации, аппроксимирующее неизвестное апостериорное распределение $p(\mathbf{w}|\mathfrak{D})$.

Рассматривается задача оптимизации параметров модели по обобщенной функции обоснованности L:

$$L(\lambda) = \lambda \mathcal{L}_{\mathbf{w}}(\mathfrak{D}, \mathbf{w}) + \mathcal{L}_E(\mathfrak{D}), \tag{1}$$

где параметр сложности λ контролирует влияние априорного распределения на выбор итоговой модели.

Первое слагаемое формулы— это различие между апостериорным и априорным распределением параметров. Оно определяется расстоянием Кульбака–Лейблера, то есть расстоянием между вариационным распределением $q(\mathbf{w})$ и априорным распределением $p(\mathbf{w})$:

$$\mathcal{L}_{\mathbf{w}}(\mathfrak{D}, \mathbf{w}) = -D_{\text{KL}}\big(q(\mathbf{w})||p(\mathbf{w})\big).$$

Второе слагаемое формулы представляет собой математическое ожидание правдоподобия выборки $\mathcal{L}_{\mathfrak{D}}(\mathfrak{D}|\mathbf{w})$:

$$\mathcal{L}_E = \mathbb{E}_{q(\mathbf{w})} \mathcal{L}_{\mathfrak{D}}(\mathfrak{D}|\mathbf{w}).$$

Пусть задано множество параметров Λ, контролирующих сложность модели. Гиперсеть — это параметрическое отображение из множества Λ во множество параметров модели:

$$\mathbf{G} : \Lambda \times \mathbb{R}^u \to \mathbb{R}^n,$$

где $\mathbb{R}^u$ — множество допустимых параметров гиперсети.

Предлагаемый метод основан на предположении, что гиперсети $\mathbf{G}$ аппроксимирующие обычные модели $\mathbf{f}$, полученные с разными значениями сложности λ, приближают их не только по качеству решения задачи, но и по статистическим свойствам. Это позволяет нам получать модели из гиперсетей, настраивать их и сокращать их параметры аналогично обычным моделям. Ниже приводится теорема, подтверждающая это предположение для простого случая компактной области, содержащей минимумы модели для всех значений сложности λ.

2 Теорема

Пусть выполнены следующие условия:

1. существует компакт $\mathbb{U} \in \mathbb{R}^n$, который содержит единственный минимум $\mathbf{w}^*(\lambda)$ для каждого $\lambda \in \Lambda$;

2. существует последовательность моделей $\mathbf{w}_n(\lambda)$ такая, что $\mathbb{E}L(\mathbf{w}_n(\lambda)) \underset{n\to\infty}{\to} \max$.

Тогда $g(\mathbf{w}_n(\lambda)) \overset{L_1}{\to} g(\mathbf{w}^*(\lambda))$, где g это непрервыный критерий для удаления параметров.

Доказательство

Из второго условия

$$\mathbb{E}L(\mathbf{w}_n(\lambda)) \underset{n\to\infty}{\to} \mathbb{E}(L(\mathbf{w}^*(\lambda)).$$

Тогда $\mathbb{E}|L(\mathbf{w}^*(\lambda)) - L(\mathbf{w}_n(\lambda))| \underset{n\to\infty}{\to} 0$, откуда следует $L(\mathbf{w}_n(\lambda)) \overset{L^1}{\to} L(\mathbf{w}^*(\lambda))$.

Мы можем показать, что аругмент функции $\mathbf{w}_n(\lambda)$ сходится к $\mathbf{w}^*(\lambda)$ в среднем. Предположим, что это не так, тогда

$$\exists\varepsilon : \forall i \quad \exists j > i : \mathbb{E}|\mathbf{w}_j - \mathbf{w}^*| > \varepsilon.$$

Пусть δ максимальное значение функции L для $\mathbf{w}_j$ из компакта $\mathbb{U}$ таких, что $\mathbb{E}|\mathbf{w}_j - \mathbf{w}^*| > \varepsilon$. Заметим что $\delta < \mathbb{E}L(\mathbf{w}^*)$. Тогда имеем бесконечную подпоследовательность моделей таких, что

$$L(\mathbf{w}_j) \le \delta < L(\mathbf{w}^*(\lambda)).$$

Но $\mathbb{E}|L(\mathbf{w}^*(\lambda)) - L(\mathbf{w}_n(\lambda))| \overset{n\to\infty}{\to} 0$. Получили противоречие с условием. Таким образом:

$$\mathbf{w}_n(\lambda) \overset{L_1}{\to} \mathbf{w}^*(\lambda).$$

Используем теорему Манна — Вальда, которая гласит: рассмотрим последовательномть X_n, X — случайные элементы, определённые на метрическом пространстве S. Пусть функция $g : S \to S'$ разрывна в некоторых точках из множества D_g таких, что $P[X \in D_g] = 0$, тогда:

$$X_n \overset{L_1}{\to} X \quad \Rightarrow \quad g(X_n) \overset{L_1}{\to} g(X).$$

Используя теорему и те факты, что наша область ограничена и функция $g(\mathbf{w}_n)$ удовлетворяет всем условиям теоремы получаем

$$g(\mathbf{w}_n(\lambda)) \overset{L_1}{\to} g(\mathbf{w}^*(\lambda)).$$