Сложность моделей глубокого обучения

Бахтеев Олег

МФТИ

02.11.2016

План

1. **Сложность модели**
2. **Вариационная нижняя оценка**
3. **Получение оценок для порождающих моделей**
4. **Получение оценок для разделяющих моделей**

Сложность модели

Мотивация

Принцип минимальной длины описания

$$\mathrm{MDL}(\mathbf{f}, \mathbf{X}) = L(\mathbf{f}) + L(\mathbf{X}|\mathbf{f}),$$

где $\mathbf{f}$ — модель, $\mathbf{X}$ — выборка, L — длина описания в битах.

$$\mathrm{MDL}(\mathbf{f}, \mathbf{X}) \sim L(\mathbf{f}) + L(\mathbf{W}^*|\mathbf{f}) + L(\mathbf{X}|\mathbf{W}^*, \mathbf{f}),$$

$\mathbf{w}^*$ — оптимальные параметры модели.

$\mathbf{f}_1 : L(\mathbf{f}_1) \quad L(\mathbf{W}_1^*|\mathbf{f}_1) \quad L(\mathbf{X}|\mathbf{W}_1^*, \mathbf{f}_1)$

$\mathbf{f}_2 : L(\mathbf{f}_2) \quad L(\mathbf{W}_2^*|\mathbf{f}_2) \quad L(\mathbf{X}|\mathbf{W}_2^*, \mathbf{f}_2)$

$\mathbf{f}_3 : L(\mathbf{f}_3) \quad L(\mathbf{W}_3^*|\mathbf{f}_3) \quad L(\mathbf{X}|\mathbf{W}_3^*, \mathbf{f}_3)$

MDL и Колмогоровская сложность

Колмогоровская сложность — длина минимального кода для выборки на предварительно заданном языке.

Теорема об инвариантности кодов

Для двух сводимых по Тьюрингу языков колмогоровской сложность отличается не более чем на константу, не зависящую от мощности выборки.

Отличия от MDL:

- Колмогоровская сложность невычислима.
- Длина кода может зависеть от выбранного языка. Для небольших выборок теорема об инвариантности кодов не дает адекватных результатов.

Оптимальная универсальная модель MDL

Пусть выборка $\mathbf{X}$ лежит в некотором конечном множестве $\mathbb{X} : \mathbf{X} \subset \mathbb{X}$.

$$\text{MDL}(\mathbf{f}, \mathbf{X}) = L(\mathbf{X}|\mathbf{W}^*(\mathbf{X}), \mathbf{f}) + \text{COMP}(\mathbf{f}),$$

$$L(\mathbf{X}|\mathbf{W}^*, \mathbf{f}) = -\log p(\mathbf{X}|\mathbf{W}^*(\mathbf{X}), \mathbf{f}), \quad \text{COMP} = log \sum_{\mathbf{X}' \in \mathbb{X}} P(\mathbf{X}'|\mathbf{W}^*(\mathbf{X}'), \mathbf{f}).$$

В случае (TODO: уточнить) оценка MDL совпадает с точностью до $o(1)$ с байесовской оценкой правдоподобия (“Evidence”):

$$p(\mathbf{X}|\mathbf{f}) = \int_{\mathbf{w}} p(\mathbf{X}|\mathbf{w}) p(\mathbf{w}) d\mathbf{w},$$

где $p(\mathbf{w})$ — априорное распределение специанльного вида:

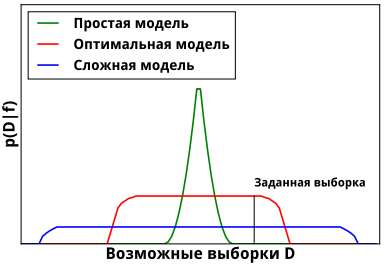
$$p(\mathbf{w}) = \frac{\sqrt{|J(\mathbf{w})|}}{\int_{\mathbf{w}'} \sqrt{|J(\mathbf{w}')|} d\mathbf{w}'},$$

$J(\mathbf{w})$ — информация Фишера.

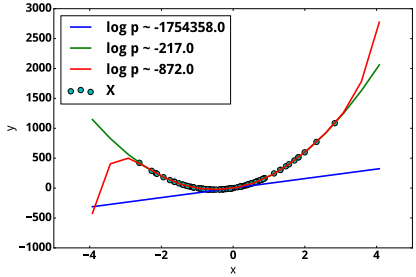
Байесовый подход к сложности

Правдоподобие модели (“Evidence”):

$$p(\mathbf{X}|\mathbf{f}) = \int_{\mathbf{w}} p(\mathbf{X}|\mathbf{w})p(\mathbf{w})d\mathbf{w}.$$



(a) Схема выбора модели по правдоподобию



(b) Пример: полиномы

Байесовый подход к сложности

Порождение vs описание
связь с Optimal MDL
Подбор априорных распределений

Кросс-валидация vs Evidence

Оценка Evidece:

$$log\ p(\mathbf{X}|\mathbf{f}) = log\ p(\mathbf{x}_1|\mathbf{f}) + log\ p(\mathbf{x}_2|\mathbf{x}_1, \mathbf{f}) + \cdots + log\ p(\mathbf{x}_n|\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}, \mathbf{f}).$$

Оценка leave-one-out:

$$LOU = \mathrm{E} log\ p(\mathbf{x}_n|\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}, \mathbf{f}).$$

Кросс-валидация оценивает сложность описания одной части выборки при условии другой части выборки.
Evidence оценивает **полную** сложность описания заданной выборки.

Методы получения оценок Evidence

MC, Laplace

Вариационная оценка

Зачем нужна, что такое

Пример: логистическая функция

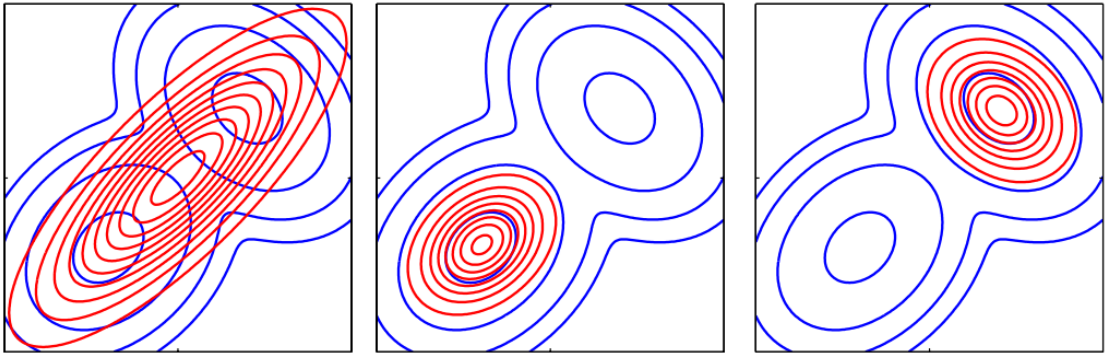
Копипсата работы Адуенко

Получение вариацонной оценки

Формула получения нижней оценки

$D_{\mathbf{KL}}$

Пример: нормальное распределение



Использование вариационной нижней оценки

Для чего используют variational inference?

- получение оценок Evidence;
- получение оценок распределений моделей со скрытыми переменными (тематическое моделирование, снижение размерности).

Зачем используют variational inference?

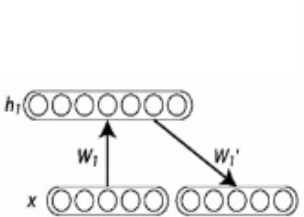
- сводит задачу нахождения апостериорной вероятности к методам оптимизации;
- проще масштабируется, чем аппроксимация Лапласа;
- проще в использовании, чем MCMC.

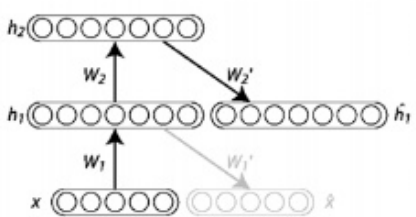
Пример: автокодировщик

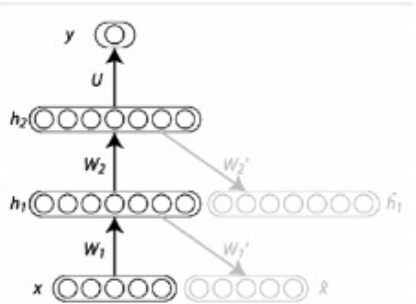
Автокодировщик — модель снижения размерности:

$$\mathbf{H} = \sigma(\mathbf{W}_e\mathbf{X}),$$

$$||\sigma(\mathbf{W}_d\mathbf{H}) - \mathbf{X}||_2^2 \to \min.$$







Автокодировщик как energy-based модель

Интегралы и картинки из Bengio

Вариационный автокодировщик

Формулы

Вариационный автокодировщик: правдоподобии модели

Полная формула

Вариационный автокодировщик: правдоподобии модели

Графики, примеры работы

Разделяющие модели: правдоподобие

аппроксимация нормальным распределением

Градиентный спуск для оценки правдоподобия

Иллюстрация

Переобучение

Иллюстрация

Динамика Ланжевина

иллюстрация

Результаты