Задание №2.5. Регуляризация по Байесу

В байесовской статистике почти любая величина является случайной, при этом мы либо ее наблюдаем, либо нет. Например, параметры θ обычно являются скрытыми случайными величинами, в то время как входные данные x и y – наблюдаемыми. Совместное распределение всех случайных величин называется моделью (т.е. $p(x, y, \theta)$). Каждую неизвестную величину можно оценить путем *обусловливания* модели всеми наблюдаемыми величинами. Подобное условное распределение вероятностей скрытых случайных величин при заданных известных называется *апостериорным распределением*. Например, $p(\theta|x, y)$ – апостериорное распределение. Последствием подобного подхода является то, что мы должны наделить параметры нашей модели, т.е. $p(\theta)$, *априорным распределением*. При этом априорные вероятности должны быть оценены до того, как мы начали работать с данными – они отражают наши предварительные представления о том, каким является распределение наших параметров.

В чистой байесовской интерпретации требуется, чтобы модель хранила все апостериорное распределение по параметрам (которое будет называться апостериорным предиктивным распределением), а итоговое предсказание модели будет являться его ожидаемым значением. Однако в большинстве случаев это очень дорого с вычислительной точки зрения, поэтому мы прибегаем к компромиссному решению.

Компромисс заключается в том, чтобы оценивать распределение для конкретных значений параметров (вместо всего распределения), а именно моду[1] апостериорного распределения. Оценка моды аспостериорного распределения еще называется *оценкой апостериорного максимума* (MAP – maximum a posteriori estimation), то есть:

$$\theta_{MAP} = \mathrm{argmax}_\theta\, p(\theta|x, y)\,.$$

Сравните это с оценкой максимального правдоподобия, которую мы уже неоднократно использовали:

$$\theta_{MLE} = \mathrm{argmax}_\theta\, p(y|x, \theta).$$

В этом задании мы исследуем взаимосвязь между MAP оценкой и стандартными техниками регуляризации, которые используются в MLE. В частности, вы покажете, как выбор конкретного априорного распределения θ (например, нормального или Лапласа) эквивалентно выбору различных методов регуляризации (например, L_2 или L_1 регуляризации).

Вопрос №1 [1 балл]

Покажите, что $\theta_{MAP} = \mathrm{argmax}_\theta\, p(y|x, \theta)p(\theta)$ в предположении $p(\theta) = p(\theta|x)$. Указанное предположение будет корректным для таких моделей как линейная регрессия, где входные данные x не моделируются явным образом с помощью θ. (Обратите внимание, что это означает, что x и θ маргинально независимы, но не являются условно независимыми при заданном y.)

Вопрос №2 [1.5 баллов]

Вспомним, что L_2 регуляризация минимизирует L_2 норму параметров модели в функции потерь (то есть в отрицательном лог-правдоподобии в случае вероятностных моделей). Теперь мы покажем, что MAP оценка с нормальным априорным распределением θ с нулевым средним, то есть

[1] Мода – точка локального максимума плотности распределения (или функции вероятности для дискретного случая).

$\theta \sim \mathcal{N}(0, \eta^2 I)$, эквивалентна применению L_2 регуляризации в MLE оценке. Более конкретно, покажите, что

$$\theta_{MAP} = \mathrm{argmin}_\theta(-\log p(y|x, \theta) + \lambda\|\theta\|_2^2).$$

Чему равно λ?

Вопрос №3 [3 балла]

Рассмотрим конкретный пример – модель линейной регрессии, задаваемую как $y = \theta^T x + \epsilon$, где $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Предположим, что случайный шум $\epsilon^{(i)}$ является независимым для каждого входного примера $x^{(i)}$. Как и ранее предполагаем гауссовское априорное распределение параметров модели $\theta \sim \mathcal{N}(0, \eta^2 I)$. Для определенности в обозначениях будем считать, что X – это матрица всей обучающей выборки, в которой каждая строка представляет собой отдельный пример, а y – это вектор-столбец всех меток. Выведите выражение для θ_{MAP} в закрытой форме.

Вопрос №4 [1.5 балла]

Теперь возьмем распределение Лапласа, обладающее следующей функцией плотности:

$$f_{\mathcal{L}}(z|\mu, b) = \frac{1}{2b}\exp\left(-\frac{|z-\mu|}{b}\right).$$

Опять возьмем модель линейной регрессии, задаваемую выражением $y = \theta^T x + \epsilon$, где $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Однако теперь рассмотрим в качестве априорного распределения параметров распределение Лапласа, при этом каждый параметр θ_i является маргинально независимым и распределенным как $\theta_i \sim \mathcal{L}(0, b)$.

Покажите, что в этом случае θ_{MAP} эквивалентно решению задачи линейной регрессии с L_1 регуляризацией, когда функция потерь задается формулой:

$$J(\theta) = \|X\theta - y\|_2^2 + \gamma\|\theta\|_1.$$

Чему равно значение γ?

Замечание. Решения в закрытой форме для задачи линейной регрессии с L_1 регуляризацией не существует. Для решения оптимизационной задачи мы используем градиентный спуск со случайной инициализацией параметров.

Линейная регрессия с L_2 регуляризацией часто называется гребневой (Ridge regression), а при L_1 регуляризации – регрессией лассо (Lasso regression). Эти методы регуляризации могут быть использованы для любой обобщенной линейной модели (при замене $\log(p(y|x, \theta)$ подходящей функцией правдоподобия). Вышеуказанные регуляризационные техники называются затуханием весов или стягиванием. Гауссовское и распределение Лапласа в качестве априорных заставляют значения весов стягиваться к их среднему значению (т.е. к нулю).

Замечание. Регрессия Лассо (т.е. с L_1 регуляризацией) известна тем, что в результате ее применения получается разреженный вектор параметров, т.е. в котором большая часть элементов равна нулю.