Машинное обучение-2 (Судный день)

Ппилиф Ульянкин

7 сентября 2022 г.

Листочек 1: потери потерь

Блин блинский! Это потеря потерь!

Кузя из Универа

Потери и правдоподобие

Задача 1 Рассмотрим линейную регрессию. Будем считать, что задан некоторый вектор весов w, и метка объекта $y(x)$ генерируется следующим образом: вычисляется линейная функция $\langle w, x\rangle$, и к результату прибавляется шум:

$$y(x) = \langle w, x\rangle + \varepsilon$$

1. Пусть $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. Покажите, что метод максимального правдоподобия в таком случае эквивалентен минимизации MSE (методу наименьших квадратов).
2. Будем считать, что ошибка имеет распределение Лапласа. Ошибка в таком случае обладает плотностью распределения

$$f_{\varepsilon}(t) = \frac{1}{2\sigma} e^{-\frac{|t|}{\sigma}}$$

Покажите, что метод максимального правдоподобия в данном случае эквивалентен минимизации MAE.

Задача 2 Пусть переменная y_i принимает только целочисленные значения. Например, это лайки на странице Маши в Instagram. Она получает их с какой-то интенсивностью λ, зависящей от характеристик её постов x_i. Например, может быть, что $\lambda = \lambda(x_i) = \langle w, x_i\rangle$. Такая модель называется *пуассоновской регрессией.* Какую функцию потерь нужно минимизировать, чтобы получить оценку w, исходя из принципа максимизации правдоподобия?

Задача 3 Выведите логистические потери, logloss, для задачи бинарной классификации, руководствуясь принципом максимизации правдоподобия.

Что мы прогнозируем?

Задача 4 Рассмотрим линейную регрессию. Будем считать, что задан некоторый вектор весов w, и метка объекта $y(x)$ генерируется следующим образом: вычисляется линейная функция $\langle w, x\rangle$, и к результату прибавляется шум:

$$y(x) = \langle w, x\rangle + \varepsilon$$

1. Пусть для оптимизаии мы используем MSE. Покажите, что оптимальным прогнозом в таком случае будет условное математическое ожидание $E(y \mid X)$.
2. Пусть для оптимизации мы используем MAE. Покажите, что оптимальным прогнозом в таком случае будет условная медиана $\mathrm{Med}(y \mid X)$.
3. Пусть для оптимизации мы используем квантильную ошибку.

$$L(y_i, \hat{y}_i) = \begin{cases} (1-\alpha)\cdot(\hat{y}_i - y_i), & \hat{y}_i > y_i \\ \alpha\cdot(y_i - \hat{y}_i), & \hat{y}_i \leqslant y_i \end{cases}$$

   Покажите, что оптимальным прогнозом в таком случае будет условный квантиль уровня α.

Задача 5 еред Винни-Пухом стоит задача классифицировать пчёл на правильных и неправильных. В его распоряжении есть выборка (y_i, x_i). Переменная y_i принимает значение 1, если пчела правильная и значение 0, если пчела неправильная. Переменная x_i — это густота мёда пчелы.

Выборка собрана, исследовательский энтузиазм зашкаливает. Есть только одна беда. Непонятно какую именно функцию потерь лучше использовать. Однако есть варианты:

1. $L(y, b(x)) = (y - b(x))^2$
2. $L(y, b(x)) = |y - b(x)|$
3. $L(y, b(x)) = y\cdot b(x) + (1-y)\cdot(1-b(x))$

Винни очень бы хотелось на выходе обязательно получить оценку вероятности принадлежности пчелы к определённому классу. Какую из функций лучше использовать исследователю?

Регуляризация

Задача 6 Скоро первая самостоятельная работа. Чтобы подготовиться к ней, Эконом ест конфеты и решает задачи. Число решённых задач y зависит от числа съеденных конфет x. Если студент не съел ни одной конфеты, то он не хочет решать задачи. Поэтому для описания зависимости числа решённых задач от числа съеденных конфет используется линейная модель с одним признаком без константы $y_i = w \cdot x_i$. В аналитическом виде найдите оценки параметра w, минимизируя следующие функции потерь:

1. Линейная регрессия без штрафа: $Q(w) = \frac{1}{\ell}\sum_{i=1}^{\ell}(y_i - wx_i)^2$;
2. Ridge-регрессия: $Q(w) = \frac{1}{\ell}\sum_{i=1}^{\ell}(y_i - wx_i)^2 + \lambda w^2$;
3. LASSO-регрессия: $Q(w) = \frac{1}{\ell}\sum_{i=1}^{\ell}(y_i - wx_i)^2 + \lambda |w|$;
4. Пусть решения этих задач равны $\hat{w}$, $\hat{w}_R$ и $\hat{w}_L$ соответственно. Найдите пределы

$$\lim_{\lambda \to 0} \hat{w}_R, \quad \lim_{\lambda \to \infty} \hat{w}_R, \quad \lim_{\lambda \to 0} \hat{w}_L, \quad \lim_{\lambda \to \infty} \hat{w}_L.$$

5. Как можно проинтерпретировать гиперпараметр λ?

Hint: в случае Lasso-регрессии придётся повозиться с модулем. Обратите внимание на то, что $Q(w)$ парабола, это поможет корректно найти аналитическое решение. Подумайте, с чем возникнут проблемы, если у нас будет не один параметр, а сотня.

Задача 7 ася измерил вес трёх покемонов, $y_1 = 6$, $y_2 = 6$, $y_3 = 10$. Вася хочет спрогнозировать вес следующего покемона с помощью константной модели $y_i = w$. Для оценки параметра w Вася использует целевую функцию

$$\frac{1}{\ell}\sum_{i=1}^{\ell}(y_i - w)^2 + \lambda w^2.$$

1. Найдите оптимальное w при произвольном λ.
2. Подберите оптимальное λ с помощью кросс-валидации leave one out («выкинь одного»). На первом шаге мы оцениваем модель на всей выборке без первого наблюдения, а на первом тестируем её. На втором шаге мы оцениваем модель на всей выборке без второго наблюдения, а на втором тестируем её. И так далее ℓ раз. Чтобы найти λ_{CV} мы минимизируем среднюю ошибку, допущенную на тестовых выборках.
3. Найдите оптимальное значение w при λ_{CV}, подобранном на предыдущем шаге.

Задача 8 Рассмотрим линейную регрессию. Будем считать, что задан некоторый вектор весов w, и метка объекта $y(x)$ генерируется следующим образом: вычисляется линейная функция $\langle w, x\rangle$, и к результату прибавляется шум:

$$y(x) = \langle w, x\rangle + \varepsilon$$

1. Введем априорное распределение на векторе весов:

$$p(w_j) = \mathcal{N}(0, \alpha^2), \quad j = 1, \dots, d.$$

   Иными словами, мы предполагаем, что веса концентрируются вокруг нуля. Покажите, что максимизация апостериорной вероятности $p(w \mid y, x)$ для модели линейной регрессии с нормальным априорным распределением эквивалентна решению задачи гребневой регрессии.

2. Какое априорное распределение надо наложить на вектор коэффициентов, если мы хотим получить LASSO-регрессию?

Потери потерь

Задача 9 Использование MSE в качестве функции потерь очень распространено. Эта функция сильнее штрафует за большие ошибки и дифференцируема. Более того, её широкой применение можно обосновать с помощью разложения в ряд Тэйлора.

Пусть выполняются следующие условия:

1. Функция потерь представима в виде $L(y, a(x)) = g(y - a(x))$,
2. Если ответ верный, тогда ошибка нулевая, $g(0) = 0$,
3. Чем больше отклонение, тем выше ошибка $|z_1| \leqslant |z_2| \Rightarrow g(z_1) \leqslant g(z_2)$
4. У функции $g(z)$ существуют первые две производные.

Покажите, что если разложить $L(y, a(x))$ до второго члена в ряд Тэйлора, то получится MSE.

Задача 10 Бандерлог утверждает, что открыл новую дифференциируемую верхнюю границу для пороговой функции потерь,

$$\tilde{L}(M_i) = \frac{9}{10} - \frac{1}{\pi} \cdot \arctan(M_i),$$

где $M_i = y_i \cdot \langle w, x_i \rangle$. Прав ли бандерлог[1]?

Задача 11 ассмотрим целевую функцию логистической регрессии

$$Q(w) = \frac{1}{\ell} \log(1 + \exp(-y \langle w, x \rangle)),$$

[1]Взял из задачника Бориса Демешева https://github.com/bdemeshev/mlearn_pro/blob/master/mlearn_pro.pdf

1. Найдите градиент ∇Q_w и упростите итоговое выражение таким образом, чтобы в нём участвовала сигмоидная функция

   $$\sigma(z) = \frac{1}{1 + \exp(-z)}.$$

   При решении данной задачи вам может понадобиться следующий факт (убедитесь, что он действительно выполняется):

   $$\sigma'(z) = \sigma(z)(1 - \sigma(z)).$$

2. Выпишите, как будет выглядеть шаг градиентного спуска.
3. Найдите вторую производную целевой функции по w.
4. Выпишите квадратичную аппроксимацию для $Q(w)$ в окрестности $w = 0$. Для этого разложите функцию потерь в ряд Тейлора до второго члена в окрестности точки $w = 0$. С какой задачей совпадает задача минимизации квадратичной аппроксимации?

Последний результат немного пригодится нам для градиентного бустинга.