

Семинары по линейным классификаторам

Евгений Соколов

30 января 2016 г.

3 Логистическая регрессия

Как известно, линейный классификатор можно настраивать, оптимизируя любую гладкую функцию потерь $L(y, z)$:

$$Q(w, X^\ell) = \sum_{i=1}^{\ell} L(y_i, \langle w, x_i \rangle) \rightarrow \min_w.$$

Выбирая разные функции потерь, можно получать классификаторы, обладающие различными свойствами. Так, логистическая функция потерь

$$L(y, z) = \log(1 + \exp(-yz))$$

позволяет обучить алгоритм $a(x)$, оценивающий вероятность принадлежности объекта к каждому из классов. Попробуем выяснить, откуда возникает это свойство.

§3.1 Предсказание вероятностей и квадратичные потери

Разберемся, каким требованиям должен удовлетворять классификатор, чтобы его выход можно было расценивать как оценку вероятности класса.

Пусть в каждой точке $x \in \mathbb{X}$ пространства объектов задана вероятность $p(y = 1 | x)$ того, что данный объект относится к первому классу. Например, объекты могут являться рекламными баннерами. Каждый объект в выборке соответствует показу одного из баннеров, и если был осуществлен клик, то объект относится к положительному классу ($y = 1$), иначе — к отрицательному ($y = -1$). Один и тот же баннер x может встретиться в выборке несколько раз, при этом часть из них будет относиться к положительному классу, а часть к отрицательному. Доля случаев, в которых баннер x относится к положительному классу, должна быть примерно равна $p(y = 1 | x)$.

Пусть алгоритм $b(x)$ возвращает числа из отрезка $[0, 1]$. Будем требовать, что эти предсказания пытались в каждой точке x приблизить вероятность положительного класса $p(y = 1 | x)$. Разумеется, выполнение этого требования зависит от функции потерь — минимум ее математического ожидания в каждой точке x должен достигаться на данной вероятности:

$$\arg \min_{b \in \mathbb{R}} \mathbb{E} [L(y, b) | x] = p(y = 1 | x).$$

Это требование можно воспринимать более просто. Пусть один и тот же объект встречается в выборке 1000 раз, из которых 100 раз он относится к классу 1, и 900 раз — к классу -1 . Поскольку это один и тот же объект, классификатор должен выдавать один ответ для каждого из тысячи случаев. Можно оценить матожидание функции потерь в данной точке по 1000 примеров при прогнозе b :

$$\mathbb{E}[L(y, b) | x] \approx \frac{100}{1000}L(1, b) + \frac{900}{1000}L(-1, b).$$

Наше требование, по сути, означает, что оптимальный ответ с точки зрения этой оценки должен быть равен $1/10$:

$$\arg \min_{b \in \mathbb{R}} \left(\frac{100}{1000}L(1, b) + \frac{900}{1000}L(-1, b) \right) = \frac{1}{10}.$$

Задача 3.1. Покажите, что квадратичная функция потерь $L(y, z) = (y - z)^2$ позволяет предсказывать корректные вероятности.

Решение. Заметим, что поскольку алгоритм возвращает числа от 0 до 1, то его ответ должен быть близок к единице, если объект относится к положительному классу, и к нулю — если объект относится к отрицательному классу.

Запишем матожидание функции потерь в точке x :

$$\mathbb{E}[L(y, b) | x] = p(y = 1 | x)(b - 1)^2 + (1 - p(y = 1 | x))(b - 0)^2.$$

Продифференцируем по b :

$$\begin{aligned} \frac{\partial}{\partial b} \mathbb{E}[L(y, b) | x] &= 2p(y = 1 | x)(b - 1) + 2(1 - p(y = 1 | x))b = \\ &= 2b - 2p(y = 1 | x) = 0. \end{aligned}$$

Легко видеть, что оптимальный ответ алгоритма действительно равен вероятности:

$$b = p(y = 1 | x).$$

■

§3.2 Логистические потери

Если алгоритм $b(x) \in [0, 1]$ действительно выдает вероятности, то они должны согласовываться с выборкой. С точки зрения алгоритма вероятность того, что в выборке встретится объект x_i с классом y_i , равна $b(x_i)^{[y_i=+1]}(1 - b(x_i))^{[y_i=-1]}$. Исходя из этого, можно записать правдоподобие выборки (т.е. вероятность получить такую выборку с точки зрения алгоритма):

$$\mathcal{L}(a, X^\ell) = \prod_{i=1}^{\ell} b(x_i)^{[y_i=+1]}(1 - b(x_i))^{[y_i=-1]}.$$

Данное правдоподобие можно использовать как функционал для настройки алгоритма — с той лишь оговоркой, что удобнее оптимизировать его логарифм:

$$-\sum_{i=1}^{\ell} ([y_i = +1] \log b(x_i) + [y_i = -1] \log(1 - b(x_i))) \rightarrow \min$$

Данная функция потерь называется логарифмической (log-loss). Покажем, что она также позволяет корректно предсказывать вероятности.

Задача 3.2. *Покажите, что логарифмическая функция потерь $L(y, z) = -[y = +1] \log z - [y = -1] \log(1 - z)$ достигает минимума на корректных вероятностях.*

Решение. Запишем матожидание функции потерь в точке x :

$$\mathbb{E}[L(y, b) | x] = -p(y = 1 | x) \log b - (1 - p(y = 1 | x)) \log(1 - b).$$

Продифференцируем по b :

$$\frac{\partial}{\partial b} \mathbb{E}[L(y, b) | x] = -\frac{p(y = 1 | x)}{b} + \frac{1 - p(y = 1 | x)}{1 - b} = 0.$$

Легко видеть, что оптимальный ответ алгоритма равен вероятности:

$$b = p(y = 1 | x).$$

■

§3.3 Логистическая регрессия

Везде выше мы требовали, чтобы алгоритм $b(x)$ возвращал числа из отрезка $[0, 1]$. Этого легко достичь, если положить $b(x) = \sigma(\langle w, x \rangle)$, где в качестве σ может выступать любая монотонно неубывающая функция с областью значений $[0, 1]$. Мы будем использовать сигмоидную функцию: $\sigma(z) = \frac{1}{1 + \exp(-z)}$. Таким образом, чем больше скалярное произведение $\langle w, x \rangle$, тем больше будет предсказанная вероятность. Как при этом можно интерпретировать данное скалярное произведение? Чтобы ответить на этот вопрос, преобразуем уравнение

$$p(y = 1 | x) = \frac{1}{1 + \exp(-\langle w, x \rangle)}.$$

Выражая из него скалярное произведение, получим

$$\langle w, x \rangle = \log \frac{p(y = +1 | x)}{p(y = -1 | x)}.$$

Получим, что скалярное произведение будет равно логарифму отношения вероятностей классов (log-odds).

Выше мы убедились, что при использовании квадратичной функции потерь алгоритм будет пытаться предсказывать вероятности. Однако, данная функция потерь

является далеко не самой лучшей, поскольку слабо штрафует за грубые ошибки — если алгоритм присвоит положительному объекту вероятность ноль, то штраф будет равен всего лишь единице. Логарифмическая функция потерь подходит гораздо лучше, поскольку не позволяет алгоритму сильно ошибаться в вероятностях.

Подставим трансформированный ответ линейной модели в логарифмическую функцию потерь:

$$\begin{aligned}
 & - \sum_{i=1}^{\ell} \left([y_i = +1] \log \frac{1}{1 + \exp(-\langle w, x_i \rangle)} + [y_i = -1] \log \frac{\exp(-\langle w, x_i \rangle)}{1 + \exp(-\langle w, x_i \rangle)} \right) = \\
 & = - \sum_{i=1}^{\ell} \left([y_i = +1] \log \frac{1}{1 + \exp(-\langle w, x_i \rangle)} + [y_i = -1] \log \frac{1}{1 + \exp(\langle w, x_i \rangle)} \right) = \\
 & = \sum_{i=1}^{\ell} \log (1 + \exp(-y_i \langle w, x_i \rangle)).
 \end{aligned}$$

Полученная функция в точности представляет собой логистические потери, упомянутые в начале. Линейная модель классификации, настроенная на данный функционал, называется логистической регрессией. Как видно из приведенных рассуждений, она оптимизирует правдоподобие выборки и дает корректные оценки вероятности принадлежности к положительному классу.