Машинное обучение, ФКН ВШЭ
Семинар №4

1 Вероятностный взгляд на линейную регрессию

Ранее мы рассматривали линейную регрессию с точки зрения минимизации функции потерь. Вероятностный подход дает красивую статистическую интерпретацию различных функций потерь и регуляризации.

§1.1 Вероятностный взгляд на функции потерь

В вероятностном подходе к машинному обучению все сущности (объекты, целевые переменные, параметры моделей) рассматриваются как случайные величины. Задача ставится следующим образом: надо найти распределение на эти случайные величины, которое лучше всего описывает данные (множество пар объект − целевой признак). Это распределение будет описывать процесс порождения наших данных (рассмотрим далее на примерах). На этом семинаре мы будем рассматривать дискриминативные модели, в которых предсказывается распределение на целевую переменную при заданном объекте: $p(y_i|x_i, \theta)$, θ — параметры модели. Другой подход — это моделировать совместное распределение $p(y_i, x_i|\theta)$ (генеративная модель).

Для настройки параметров в вероятностном подходе часто применяют метод максимального правдоподобия, известный вам из курса математической статистики. В нем правдоподобие вероятностной модели:

$$p(y|X, \theta) = \prod_i p(y_i|x_i, \theta)$$

логарифмируется и оптимизируется по параметрам модели:

$$\log p(y|X, \theta) = \sum_i \log p(y_i|x_i, \theta) \to \max_\theta .$$

До этого момента мы рассматривали только точечные предсказания: $y_i = w^T x_i$. Когда мы начинаем рассматривать y_i как случайную величину, мы как бы признаем неточность этого предсказания и допускаем, что можем ошибиться. Логично взять распределение $p(y_i|x_i, \theta)$, имеющее моду в $y = w^T x$ и монотонно убывающее слева и справа от нее. Для начала возьмем нормальное распределение:

$$p(y_i|x_i, \theta) = \mathcal{N}(y_i|x_i^T w, \sigma^2).$$

Здесь мы обозначили $\theta = \{w, \sigma\}$ — множество параметров вероятностной модели. Процесс порождения данных в этом случае предельно простой: мы предполагаем, что y_i случайно сгенерировано из $\mathcal{N}(y_i|x_i^T w, \sigma^2)$.

Задача 1.1. *Найдите, какой функции потерь соответствует метод максимального правдоподобия для данной модели.*

Решение. Преобразуем логарифм правдоподобия:

$$\log p(y|X,\theta) = \sum_i \log \mathcal{N}(y_i|x_i^T w, \sigma^2) =$$

$$= \sum_i \log\left(\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y_i - w^T x_i)^2}{2\sigma^2}\right)\right) =$$

$$= \sum_i \left(-0.5\log(2\pi) - \log\sigma - \frac{(y_i - w^T x_i)^2}{2\sigma^2}\right) =$$

$$= C - \frac{1}{2\sigma^2}\sum_i (y_i - w^T x_i)^2 \to \max_w, \quad C = -\frac{\ell}{2}\log(2\pi)$$

Константы C, σ не влияют на точку оптимума, а знак минус можно удалить, заменив задачу макисимзации на задачу минимизации, поэтому в итоге мы получаем слудеющую оптимизационную задачу:

$$\sum_i (y_i - w^T x_i)^2 \to \min_w$$

Мы видим, что применение метода максимального правдоподобия к нашей задаче равносильно оптимизации MSE, что соответствует квадратичной функции потерь.

■

Итак, вероятностный подход дает следующую интерпретацию квадратичной функции потерь для линейных моделей: при ее применении мы предполагаем, что (а) зависимость целевой переменной от признаков линейна и (б) ошибки предсказаний линейной модели распределены нормально. На практике последнее условие можно попробовать проверить графически, а именно: построить гистограмму ошибок и оценить ее нормальность. Конечно, в статистике существуют специальные тесты для проверки нормальности, но мы не будем их затрагивать.

Аналогично можно провести анализ, если вместо нормального распределения взять, к примеру, распределение Лапласа:

$$p(y_i|x_i, \theta = \{w, \alpha\}) = \frac{\alpha}{2}\exp\left(-\alpha\left|y_i - w^T x_i\right|\right).$$

Для этой модели метод максимального правдоподобия соответствует оптимизации MAE.

§1.2 Вероятностный взгляд на регуляризацию

Во всех предыдущих рассуждениях мы рассматривали только целевую переменную как случайную величину, а признаки объекта и веса считали фиксированным векторамию. Мы можем несколько усложнить нашу вероятностную модель и считать, что веса модели — тоже случайная величина. Здесь мы придем к понятию априорного распределения $p(w)$: это распределение на веса линейной модели,

отображающее наше представление о них до того, как к нам поступили данные. Модель порождения данных тогда будет выглядеть так: 1. Сгенерировать вектор весов $w \sim p(w)$; 1. Для всех объектов сгенерировать $y_i \sim p(y_i|x_i, \theta = \{w, \dots\})$. Вероятностная модель:

$$p(y, w|X) = [\prod_i p(y_i|x_i, \theta)]p(w).$$

Когда мы получили данные, мы можем искать наиболее вероятное значение весов — моду апостериорного распределения $p(w|X, y)$.

Задача 1.2. *Какому критерию качества соответствует поиск моды апостериорного распределения, если $p(w) = \mathcal{N}(w|0, I)$ — стандартное многомерное нормальное распределение?*

Решение. Найдем моду распределения. По формуле Байеса:

$$p(w|X, y) = \frac{p(y|X, w)p(w)}{p(y|X)} = \frac{\prod_i p(y_i|x_i, w)p(w)}{p(y|X)}$$

Знаменатель не зависит от w и не повлияет на точку моды распределения, поэтому не будем далее его рассматривать. Прологарифмируем числитель:

$$\sum_i \log p(y_i|x_i, w) + \log p(w) \to \max_w$$

Первое слагаемое мы уже вычисляли ранее, оно соответствует среднеквадратичному отклонению. Займемся вторым слагаемым:

$$\log p(w) = \log \mathcal{N}(w|0, I) = \log\left(\frac{1}{\sqrt{2\pi}^d} \exp\big(-\frac{w^T w}{2}\big)\right) =$$

$$-\frac{d}{2}\log\pi - \frac{w^T w}{2}.$$

Первое слагаемое — константа, не влияющая на точку минимума, второе слагаемое — евклидова норма весов. Итак, мы получили следующий оптимизируемый критерий:

$$-\frac{1}{2\sigma^2}\sum_i (y_i - w^T x_i)^2 + \frac{1}{2}w^T w \to \min_w$$

Это в точности соответствует L_2-регуляризованной линейной регрессии. Обратите внимание, что здесь мы уже не можем опустить множитель $\frac{1}{2\sigma^2}$. Семантически он соответствует коэффициенту регуляризации и настраивает баланс между оптимизацией качества решения задачи и регуляризатором.

■

Аналогично можно рассмотреть, к примеру, априорное распределение Лапласа и получить, что оно соответствует L_1-регуляризации в линейной регрессии.

Напоследок (в этой секции) отметим, что все приведенные рассуждения можно повторить для любых алгоритмов машинного обучения, заменив $w^T x$ на другую модель зависимости.

2 Квантильная регрессия

В некоторых задачах цены занижения и завышения прогнозов могут отличаться друг от друга. Например, при прогнозировании спроса на товары интернет-магазина гораздо опаснее заниженные предсказания, поскольку они могут привести к потере клиентов. Завышенные же прогнозы приводят лишь к издержкам на хранение товара на складе. Функционал в этом случае можно записать как

$$Q(a, X^\ell) = \sum_{i=1}^{\ell} \rho_\tau(y_i - a(x_i)),$$

где

$$\rho_\tau(z) = (\tau - 1)[z < 0]z + \tau[z \geqslant 0]z = (\tau - \frac{1}{2})z + \frac{1}{2}|z|,$$

а параметр τ лежит на отрезке $[0, 1]$ и определяет соотношение важности занижения и завышения прогноза. Чем больше здесь τ, тем выше штраф за занижение прогноза.

Обсудим вероятностный смысл данного функционала. Будем считать, что в каждой точке $x \in \mathbb{X}$ пространства объектов задано вероятностное распределение $p(y \mid x)$ на возможных ответах для данного объекта. Такое распределение может возникать, например, в задаче предсказания кликов по рекламным баннерам: один и тот же пользователь может много раз заходить на один и тот же сайт и видеть данный баннер; при этом некоторые посещения закончатся кликом, а некоторые — нет.

Известно, что при оптимизации квадратичного функционала алгоритм $a(x)$ будет приближать условное матожидание ответа в каждой точке пространства объектов: $a(x) \approx \mathbb{E}[y \mid x]$; если же оптимизировать среднее абсолютное отклонение, то итоговый алгоритм будет приближать медиану распределения: $a(x) \approx \text{median}[p(y \mid x)]$. Рассмотрим теперь некоторый объект x и условное распределение $p(y \mid x)$. Найдем число q, которое будет оптимальным с точки зрения нашего функционала:

$$Q = \int_{\mathbb{Y}} \rho_\tau(y - q)p(y \mid x)dy.$$

Продифференцируем его (при этом необходимо воспользоваться правилами дифференцирования интегралов, зависящих от параметра):

$$\frac{\partial Q}{\partial q} = (1 - \tau) \int_{-\infty}^{q} p(y \mid x)dy - \tau \int_{q}^{\infty} p(y \mid x)dy = 0.$$

Получаем, что

$$\frac{\tau}{1 - \tau} = \frac{\int_{-\infty}^{q} p(y \mid x)dy}{\int_{q}^{\infty} p(y \mid x)dy}.$$

Данное уравнение будет верно, если q будет равно τ-квантили распределения $p(y \mid x)$. Таким образом, использование функции потерь $\rho_\tau(z)$ приводит к тому, что алгоритм $a(x)$ будет приближать τ-квантиль распределения ответов в каждой точке пространства объектов.

3 Предобработка данных

§3.1 Пропущенные значения

В реальных задачах значения некоторых признаков у некоторых объектов отсутствуют. Это может происходить по разным причинам: ошибки при записи данных, отказ респондента отвечать на вопрос, невозможность описать конкретное свойство у конкретного объекта (в таблице с данными об автомобилях будут пропуски у электромобилей в графе «объём топливного бака»). Многие алгоритмы машинного обучения (в частности линейная регрессия) не могут работать с пропущенными данными, поэтому эти пропуски необходимо заполнить.

Заполнять пропуски у объектов можно различными способами:

1. Константным уникальным значением – неудачный вариант для линейных методов (модель начнёт считать пропуск близким к некоторому другому значению выборки), но быстрый и популярный способ с другими алгоритмами в машинном обучении.

2. Средним арифметическим, медианой, модой – сохранение статистик выборки, но потеря информации о наличии пропуска в данных.

3. Предсказаниями другого алгоритма – затратно по времени (однако всё равно не приносит новой информации в датасет, хотя и может положительно сказаться на общем качестве).

Заметим, что в некоторых случаях наличие пропуска в данных несёт определённую информацию об объекте (например, отказ в ответе на вопрос о доходах клиента банка), поэтому полезно добавлять новые признаки – индикаторы пропусков. Иногда признаки содержат слишком много пропусков и их выгоднее удалить.

§3.2 Выбросы

На практике могут встречаться объекты, сильно отличающиеся от остальных. Их называют выбросами. Отличия могут выражаться как в значениях признаков, так и в целевой величине. Причины бывают различными: ошибки в заполнении данных (добавили лишний ноль), «исключительность» отдельных объектов (низкие цены на дома могут быть связаны с попыткой обхода налогов, а не с их характеристиками). Выбросы могут сильно сказываться на решении: например, квадратичная функция ошибок «реагирует» на выбросы и линейная регрессия отклоняется в их сторону, в отличие от средней абсолютной ошибки (рис. 1).

Искать выбросы можно следующим образом. Выбросы в признаках можно обнаружить, исследуя распределение признаков и в особенности хвосты распределений. Выбросы в целевой величине можно искать, считая ошибку предсказания модели на объектах обучающей выборки (вспомогательная модель не должна наблюдать при обучении проверяемый объект). Если ошибка велика (алгоритм с уверенностью предсказывает отрицательный класс, хотя метка у объекта положительная), то объект можно считать выбросом (если, конечно, дело не в плохой модели). Объекты-выбросы чаще всего не корректируют, а удаляют из выборки.

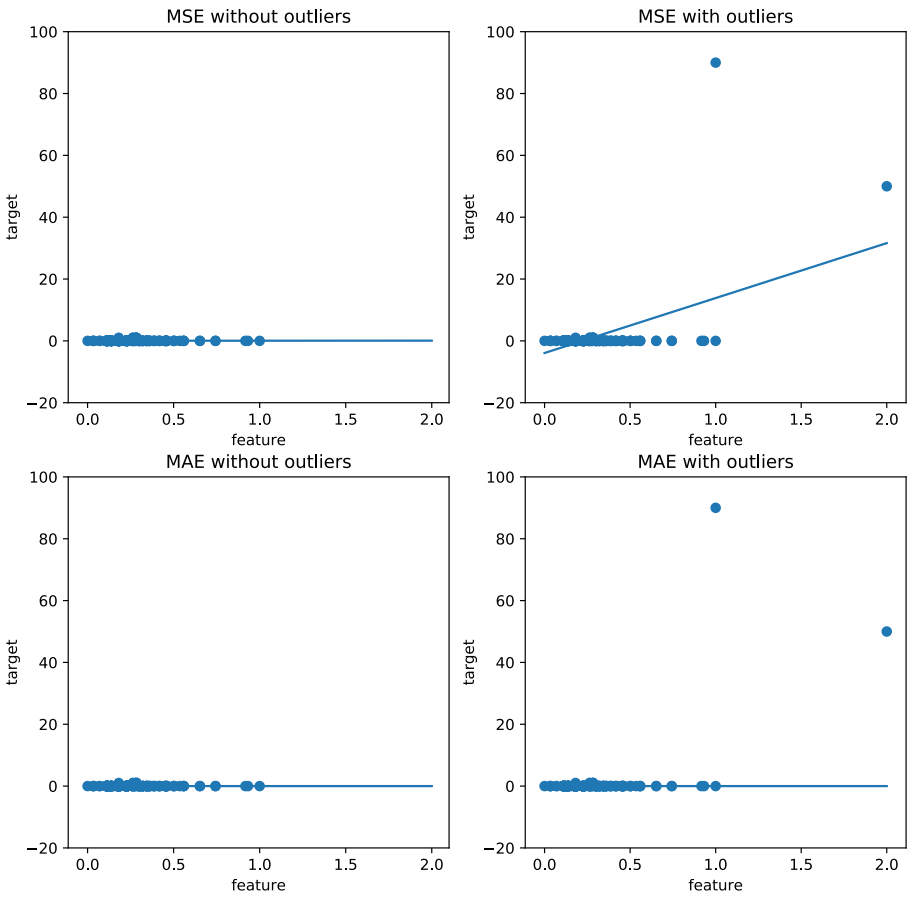

Рис. 1. Влияние выбросов на обученную линейную регрессию для MSE и MAE в качестве функции потерь.

Заметим, что не всегда необходимо удалять объекты-выбросы из выборки. В одном конкурсе помог следующий подход: оставить выбросы, чтобы не изменилось среднее предсказание алгоритма, при этом качество модели считать только по «нормальным» объетам, чтобы исключить шум от объектов-выбросов.

Как уже было сказано выше, некоторые функции потерь чувствительнее относятся в выбросам, поэтому в таких ситуациях имеет смысл использовать более устойчивые функции потерь для обучения моделей.

4 Работа со специфичными фукнциями потерь

В некоторых задачах приходится сталкиваться с нестандартными функциями потерь (по сравнению с квадратичной). Например, средняя относительная ошибка (MAPE) или средняя абсолютная ошибка, в которой ошибки менее некоторой величины не учитываются. Иногда такую функцию потерь всё-таки можно использовать при обучении модели (можно взять градиент MAPE), однако обучение такой модели требует модификаций в используемые алгоритмы и может потребовать значительно больше времени для настройки.

Простым способом является использование стандартных фукнций потерь для непосредственного обучения модели (которые уже встроены в используемый алгоритм) и использование нужной фукнции потерь для подбора гиперпараметров, отбора признаков и настройки просто постобработки предсказаний.

В некоторых случаях, можно аппроксимировать функцию потерь через некоторую другую. Например, ещё в одном конкурсе в качестве метрики была дана следующая функция:

$$RMSPE = \sqrt{\frac{1}{l}\sum_{i=1}^{l}\left(\frac{\hat{y}_i - y_i}{y_i}\right)^2}$$

Привычные алгоритмы машинного обучения не умеют оптимизировать такую функцию потерь, но почти всегда умеют оптимизировать квадратичную функцию потерь. Пусть существует такая функция $f(x)$, что $RMSPE(y, \hat{y}) = RMSE(f(y), f(\hat{y}))$. То есть:

$$\frac{\hat{y} - y}{y} = f(\hat{y}) - f(y) \approx f(y) + f^{'}(y)(\hat{y} - y) - f(y) = f^{'}(y)(\hat{y} - y)$$

Тогда:

$$f^{'}(y) = \frac{1}{y}$$

Получаем, что $f(y) = \log y + C$. Константа C под квадратом взаимно уничтожится с другой. Таким образом, можно приближённо оптимизировать RMSPE, сделав логарифмическое преобразование целевой переменной.