В сторону от линейной модели

Сейчас мы посмотрим, как можно в нашем общем контексте немного отойти в сторону от классической линейной регрессии в поисках всё ещё интерпретируемой модели, хотя уже и не так сильно покрытой статистическими критериями.

9.1 Что предсказывают линейные модели?

До сих пор мы рассматривали классическую регрессионную модель, основанную на методе наименьших квадратов. Мы сумели построить явно оценку параметров и показать, как в таком сеттинге можно проверять гипотезы. Теперь же нас интересует, как можно, с одной стороны, улучшить в некоторых аспектах линейную модель, отказавшись от метода наименьших квадратов, а с другой – предложить что-то ещё для решения задач, возникающих в машинном обучении, основываясь на линейной идее и по-другому записав вероятностную модель.

Кокнретнее, в машинном обучении принято рассматривать две большие группы задач, где по набору признаков требуется что-то предсказать про объект.

1. Регрессия: предсказать числовое значение или несколько.
2. Классификация: предсказать класс(или тип, метку) объекта.

Мы начнём с регрессии, а потом предложим один интересный метод для классификации.

9.2 В чём проблема MSE?

Функция потерь, называемая MSE, задаётся как

$$MSE(\theta) = \frac{1}{n}\|X\theta - Y\|_2^2,$$

как мы видели раньше, она возникает естественно в модели линейной регрессии

$$Y = X\theta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

при оценке параметров θ. Оказалось, что

$$\hat{\theta} = \arg\min_{\theta} MSE(\theta)$$

дало нам оценку с большим количеством интересных статистических свойств. Какой прогноз предлагает нам модель с такой оценкой параметров?

При фиксированных X переменная Y имеет некоторое условное распределение $p(y|x)$ и когда мы оптимизируем MSE, мы на самом деле хотели бы минимизировать, подбирая функцию прогноза $\hat{Y}(x)$, матожидание квадрата ошибки

$$\mathbb{E}\left[(Y - \hat{Y}(x))^2 \mid X = x\right] = \int (y - \hat{Y}(x))^2 p(y|x) dy.$$

Если продифференцируем по $\hat{Y}$ и приравняем к нулю, то получим

$$\int y p(y|x) dy - \int \hat{Y} p(y|x) dy = 0$$

или, по-другому,

$$\hat{Y}(x) = \mathbb{E}\left[Y \mid X = x\right]$$

и есть прогноз, который получается из MSE. Поскольку это условное матожидание, то есть, среднее, то такой прогноз оказывается очень неустойчивым к выбросам в данных и от выбросов приходится избавляться отдельно.

9.3 L1-регрессия

Один из вариантов построения более устойчивой к выбросам линейной модели – поменять функцию потерь на

$$MAE(\theta) = \frac{1}{n}\|Y - X\theta\|_1,$$

которая называется *Mean Absolute Error(MAE)* или просто суммой модулей отклонений. Такая функция, как ни странно, тоже возникает из оценки метода максимального правдоподобия для регрессионной модели

$$Y = X\theta + \varepsilon, \quad \varepsilon_i \sim_{iid} Laplace(\alpha).$$

Распределение Лапласа – это экспоненциальное распределение, которое симметрично продолжено влево от нуля и правильно отнормировано:

$$p(x) = \frac{\alpha}{2} e^{-\alpha|x|}.$$

Как и в гауссовском случае, лог-правдоподобие будет иметь вид

$$l(\theta) = -\alpha \sum_{i=1}^{n} |Y_i - X_{i,\cdot}\theta|,$$

поэтому максимизация по θ лог-правдоподобия даст ту же оценку, что и

$$\hat{\theta}_{MAE} = \arg\min_{\theta} MAE(\theta).$$

Давайте посмотрим, какой проноз нам продвигает такая модель, конкретнее, мы заинтересованы в

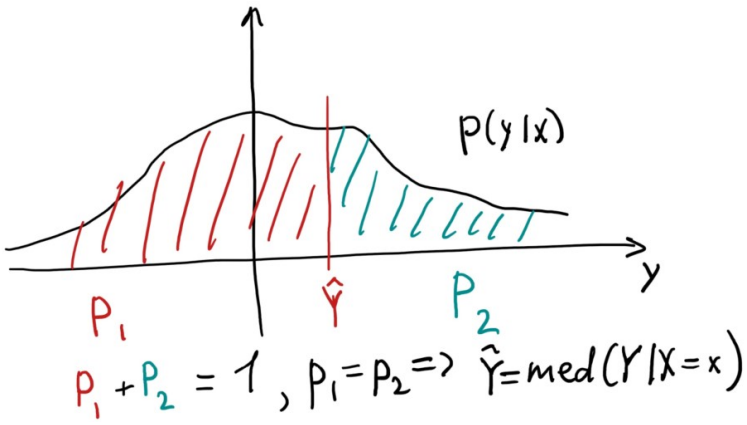
$$\min_{\hat{Y}} \mathbb{E}\left[|Y - \hat{Y}| \mid X = x\right].$$

Поскольку модуль недифференцируем в одной точке, то нам придётся поделить интеграл на две части:

$$\mathbb{E}\left[|Y - \hat{Y}| \mid X = x\right] = \int_{\hat{Y}<y} (y - \hat{Y})p(y|x)dy + \int_{\hat{Y}>y} (\hat{Y} - y)|p(y|x)dy.$$

Когда продифференцируем (не забываем, что пределы интегрирования зависят от $\hat{Y}$) и приравняем к нулю, получим, что $\hat{Y}$ такая, что

$$\int_{\hat{Y}<y} p(y|x)dy = \int_{\hat{Y}>y} p(y|x)dy,$$

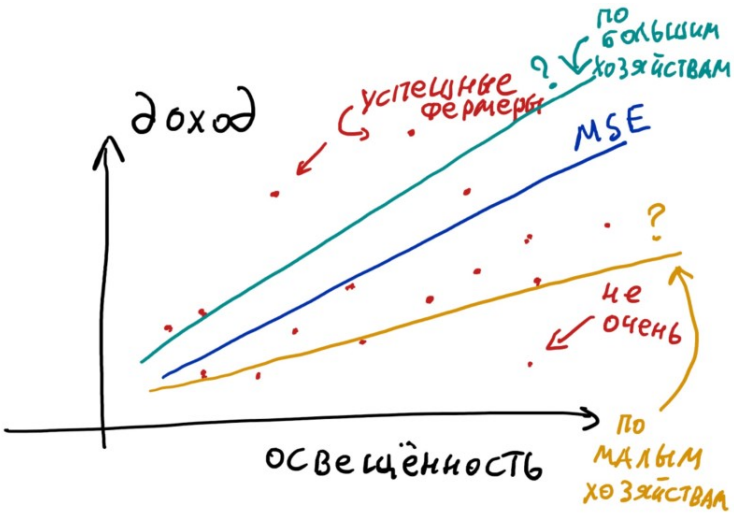


то есть, прогноз $\hat{Y}(x) = Med(Y|X = x)$. Медиана гораздо более устойчива к выбросам за счёт более тяжёлых хвостов распределения Лапласа. В силу того, что $\hat{\theta}_{MAE}$ – оценка метода максимального правдоподобия, многие хорошие свойства сохраняются, в том числе, проверка гипотез LR-, LM- и W-тестами для методов максимального правдоподобия.

9.4 Квантильная регрессия

Если мы только что выяснили, что мы можем научиться предсказывать условное матожидание, условную медиану, то интересный вопрос состоит в следующем: можем ли мы предсказывать квантиль? Медиана – это квантиль уровня 0.5, можно ли научиться выдавать квантиль уровня 0.2?

Пример 9.1. *Алексей панически боится, что его вложения в новую бразильскую кофейную плантацию, которую хотят открыть его знакомые, не принесут ему ожидаемой прибыли или даже принесут убытки. Он нашёл на просторах интернета датасет, в котором собраны кофейные плантации и приведены различные характеристики (геопозиция, тип почвы, площадь, среднегодовое количество осадков, количество солнечных дней, количество работников....), стартовый капитал и окупаемость после года работы. Его интересует предсказание худших доходов, например, квантили уровня* 0.2.



Эта идея лежит в основе меры риска (понятие в финансовой математике), которая называется Value-at-Risk

$$VaR_\gamma(Y) = q_\gamma(Y).$$

Если мы бы имели такой метод, мы умели бы ещё предсказывать доверительные интервалы и даже целые гистограммы! Для этого давайте вспомним, что такое квантиль случайной величины Y уровня γ, формально это

$$q_Y(\gamma) = \inf\{y \ : \ \mathbb{P}(Y \le y) \ge \gamma\}.$$

Другим образом его можно определить как

$$q_Y(\gamma) = \arg\min_q \left\{ \gamma \int_q^\infty (y-q) dF_Y(y) + (1-\gamma) \int_{-\infty}^q (q-y) dF_Y(y) \right\},$$

потому что, если продифференцировать по прогнозу q, то мы получим

$$\gamma \int_q^\infty dF_Y(y) = (1-\gamma) \int_{-\infty}^q dF_Y(y),$$

вероятность справа неизбежно равна γ, а слева $(1-\gamma)$. Если присмотреться, то при $\gamma = 0.5$ мы такой результат уже видели в MAE-регрессии; действительно, квантиль уровня 0.5 – это как раз медиана.

Для поиска медианного прогноза мы минимизировали $MAE(\theta)$, попробуем что-то похожее найти для квантили. Оказывается, что сумму выше можно представить как

$$\gamma \int_q^\infty (y-q) dF_Y(y) + (1-\gamma) \int_{-\infty}^q (q-y) dF_Y(y) = \mathbb{E}\left[\rho_\gamma(Y-q)\right],$$

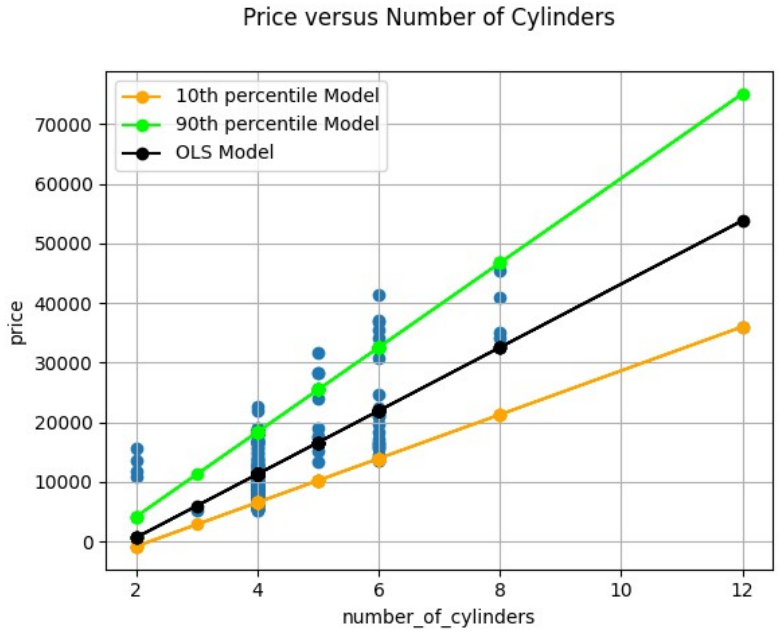
где

$$\rho_\gamma(t) = t(\gamma - \mathbb{1}(t<0)).$$

То есть, для поиска оценки можем решать эмпирический аналог

$$\hat{q}_\gamma = \arg\min_q \frac{1}{n} \sum_{i=1}^n \rho_\gamma(Y_i - q).$$

В результате получим, что мы можем, меняя γ, менять оцениваемую квантиль и даже оценивать несколько квантилей отдельно. Сравните результат с MSE(в `statsmodels` это Ordinary Least Squares, OLS) для датасета `automobiles_dataset_uciml`:



9.5 Логистическая регрессия

Теперь давайте рассмотрим ещё один интересный подход, который тоже почти линейный и является одной из возможных адаптаций линейной модели для задачи классификации.

Положим, мы хотим решать задачу бинарной классификации с классами $Y_i \in \{0, 1\}$.

Пример 9.2. *Алексей долго задумывается о том, чтобы купить машину, так как видит, что много коллег ездит на ней на работу, но долго не может решиться. Часто откладывает до своего повышения... Тут ему стало интересно: как влияет уровень дохода на наличие автомобиля в семье? Поискав по интернету, он нашёл датасет из Германии и решил поставить задачу классификации: предсказать наличие автомобиля Y по параметрам человека X. Его первой простой идеей было задействовать только доход и он предложил линейную модель*

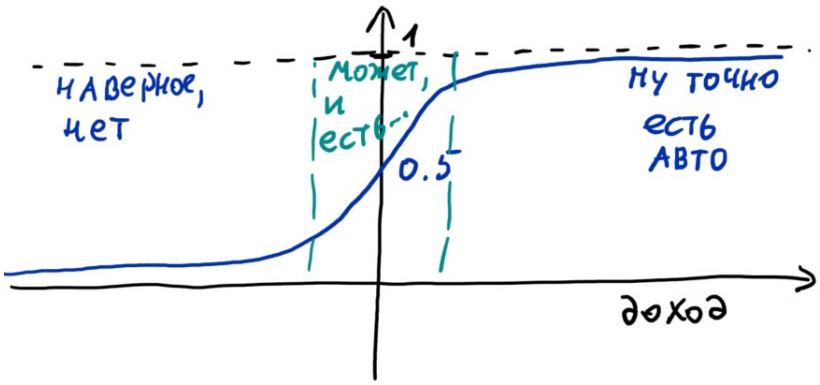
$$Y = \theta X + \varepsilon,$$

но быстро понял, что она ничего толком не позволяет ему сделать – целевая переменная категориальная. Тогда он подумал ещё немного, и ему пришла в голову мысль: а что если попробовать как-то моделировать именно вероятность?

В таком случае оказывается, что чисто линейная регрессия не отражает того, что нужно, но если мы преобразуем выход модели, то можем строить решения вида

$$\mathbb{P}\left(Y = 1 | X = x\right) = F(X\theta),$$

где F – некоторая функция, образ которой составляет отрезок $[0, 1]$ (вспомните, например, софтмакс). Один легендарный вариант – логистическая функция:



$$F(z) = \frac{1}{1 + e^{-z}}, \; -$$

приводящая к модели *логистической регрессии*. Если посмотреть на картинку, то у графика такой функции есть ярко выраженный левый убывающий и правый растущий хвосты, имеющие понятную интерпретацию. Слева денег так мало, что автомобиля точно не может быть, справа так много, что, вероятно, он есть, посередине зона неопределённости – все люди разные и не всем на этом уровне позарез нужен автомобиль.

Оценку параметров в такой модели мы, конечно, будем делать с помощью метода максимального правдоподобия, в котором после использования независимости наблюдений (X_i, Y_i) и явного вида модели мы можем немного хитро записать правдоподобие

$$L(\theta) = \prod_{i=1}^{n} F(X_{i,\cdot}\theta)^{Y_i}(1 - F(X_{i,\cdot}\theta))^{1-Y_i}$$

и дальше его прологарифмировать

$$l(\theta) = \sum_{i=1}^{n} \left(Y_i \ln F(X_{i,\cdot}\theta) + (1 - Y_i) \ln(1 - F(X_{i,\cdot}\theta))\right).$$

Иными словами,

$$l(\theta) = \sum_{i:Y_i=1} \ln F(X_{i,\cdot}\theta) + \sum_{i:Y_i=0} \ln(1 - F(X_{i,\cdot}\theta))$$

и эту функцию уже можно оптимизировать численными методами, получая достаточно популярный на практике вероятностный классификатор. Если мы зададим пороговую вероятность, то мы сможем разделять наблюдения на классы строго.