Линейная регрессия

Машинное обучение

Александр Безносиков

ИСП РАН

20 февраля 2025

Линейная регрессия

- Вспомним прошлую лекцию: в машинном обучении мы ищем такое отображение $g\colon \mathbb{R}^d \to \mathbb{R}$, чтобы оно наилучшим образом приближало связь пространства объектов $\mathcal{X} \to \mathcal{Y}$.

Линейная регрессия

- Вспомним прошлую лекцию: в машинном обучении мы ищем такое отображение $g\colon \mathbb{R}^d \to \mathbb{R}$, чтобы оно наилучшим образом приближало связь пространства объектов $\mathcal{X} \to \mathcal{Y}$.
- В данной лекции мы работаем в предположении, что целевая переменная y_i **линейно** зависит от объектов x^i.

Линейная регрессия

- Вспомним прошлую лекцию: в машинном обучении мы ищем такое отображение $g\colon \mathbb{R}^d \to \mathbb{R}$, чтобы оно наилучшим образом приближало связь пространства объектов $\mathcal{X} \to \mathcal{Y}$.
- В данной лекции мы работаем в предположении, что целевая переменная y_i **линейно** зависит от объектов x^i. Более формально:

Постановка задачи линейной регрессии

Мы ищем такую функцию

$$g(x^i, w) = w_0 + \sum_{j=1}^{d} x_j^i w_j,$$

чтобы она максимально точно приближало значение целевой метки y_i.

Веса

- В дальнейшем мы всегда настраиваемые параметры w любой необязательно линейной модели g будем называть весами.
- В случае линейной модели название «веса» передает и четкий физический смысл.

Веса: важность предобработки

- Рассмотрим пример:

x_1	x_2	y
1	2000	2,2
2	3000	3,3
4	4000	4,5

Вопрос: с какими весами надо взять линейную модель, чтобы повторить такую зависимость?

Веса: важность предобработки

- Рассмотрим пример:

x_1	x_2	y
1	2000	2,2
2	3000	3,3
4	4000	4,5

  Вопрос: с какими весами надо взять линейную модель, чтобы повторить такую зависимость?
- $w_0 = 0,1,\ w_1 = 0,1,\ w_2 = 0,001.$
- **Вопрос:** исходя из размеров весов, можем ли мы что-то сказать о важности каждого из признаков?

Веса: важность предобработки

- Рассмотрим пример:

x_1	x_2	y
1	2000	2,2
2	3000	3,3
4	4000	4,5

  Вопрос: с какими весами надо взять линейную модель, чтобы повторить такую зависимость?
- $w_0 = 0,1$, $w_1 = 0,1$, $w_2 = 0,001$.
- **Вопрос:** исходя из размеров весов, можем ли мы что-то сказать о важности каждого из признаков? Хочется сказать, что первый признак более важный, так как имеет больший вес, но это ошибочное суждение, так как изменение второго признака на 50% привело к изменению итоговой метки почти на эти же 50%.

Веса: важность предобработки

- **Вопрос:** как сделать так, чтобы веса w несли информацию о важности признака?

Веса: важность предобработки

- **Вопрос:** как сделать так, чтобы веса w несли информацию о важности признака?
- Попробуем предобработать данные следующим образом, в пределах каждого из признаков отшкалируем так, чтобы все значения лежали в отрезке $[0; 1]$.
- Это просто сделать, например,

$$\tilde{x}_i^j = \frac{x_i^j}{\max\limits_{k \in [n]} \left| x_i^k \right|}$$

или

$$\tilde{x}_i^j = \frac{x_i^j - \min\limits_{k \in [n]} x_i^k}{\max\limits_{k \in [n]} \left| x_i^k \right| - \min\limits_{k \in [n]} x_i^k}$$

Веса = значимость

- Воспользуемся первым правилом и преобразуем таблицу из примера:

$\tilde{x}_1$	$\tilde{x}_2$	y
0,25	0,5	2,2
0,5	0,75	3,3
1	1	4,5

Веса = значимость

- Воспользуемся первым правилом и преобразуем таблицу из примера:

$\tilde{x}_1$	$\tilde{x}_2$	y
0,25	0,5	2,2
0,5	0,75	3,3
1	1	4,5

- Новые веса: $\tilde{w}_0 = 0,1$, $\tilde{w}_1 = 0,4$, $\tilde{w}_2 = 4$.
- Вот теперь веса $\tilde{w}$ лучше отражают значимость признаков. Видно, что $\tilde{w}_2$ значительно больше $\tilde{w}_1$, что всецело коррелирует с его влиянием на итоговую метку y.

Веса = значимость

- Воспользуемся первым правилом и преобразуем таблицу из примера:

$\tilde{x}_1$	$\tilde{x}_2$	y
0,25	0,5	2,2
0,5	0,75	3,3
1	1	4,5

- Новые веса: $\tilde{w}_0 = 0,1$, $\tilde{w}_1 = 0,4$, $\tilde{w}_2 = 4$.
- Вот теперь веса $\tilde{w}$ лучше отражают значимость признаков. Видно, что $\tilde{w}_2$ значительно больше $\tilde{w}_1$, что всецело коррелирует с его влиянием на итоговую метку y.
- В машинном обучении часто y так же является признаком и его можно преобразовывать аналогичным образом.
- Кроме приведенного примера существует масса других классических подходов.

Веса: больше предобработок

- Например, давайте потребуем:

$$\frac{1}{n}\sum_{i=1}^{n}\tilde{x}^i = 0, \quad \frac{1}{n}\sum_{i=1}^{n}\tilde{y}^i = 0.$$

- **Вопрос:** как такое осуществить?

Веса: больше предобработок

- Например, давайте потребуем:

$$\frac{1}{n}\sum_{i=1}^{n}\tilde{x}^i = 0, \quad \frac{1}{n}\sum_{i=1}^{n}\tilde{y}^i = 0.$$

- **Вопрос:** как такое осуществить? Посчитаем $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x^i$ и положим $\tilde{x}^i = x^i - \bar{x}$, аналогично для y.

Веса: больше предобработок

- Например, давайте потребуем:

$$\frac{1}{n}\sum_{i=1}^{n}\tilde{x}^i = 0, \quad \frac{1}{n}\sum_{i=1}^{n}\tilde{y}^i = 0.$$

- **Вопрос:** как такое осуществить? Посчитаем $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x^i$ и положим $\tilde{x}^i = x^i - \bar{x}$, аналогично для y.
- Утверждается, что $\tilde{w}_0 = 0$ для таких $\tilde{x}^i$ и $\tilde{y}^i$.

Веса: больше предобработок

- Например, давайте потребуем:

$$\frac{1}{n}\sum_{i=1}^{n}\tilde{x}^i = 0, \quad \frac{1}{n}\sum_{i=1}^{n}\tilde{y}^i = 0.$$

- **Вопрос:** как такое осуществить? Посчитаем $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x^i$ и положим $\tilde{x}^i = x^i - \bar{x}$, аналогично для y.
- Утверждается, что $\tilde{w}_0 = 0$ для таких $\tilde{x}^i$ и $\tilde{y}^i$. Докажем этот факт:

$$\frac{1}{n}\sum_{i=1}^{n}(w_0 + \langle x^i, w\rangle)^2 = \frac{1}{n}\sum_{i=1}^{n}\left[w_0^2 + \sum_{j=1}^{d}(x_j^i \cdot w_j)^2\right]$$

$$+ \frac{2}{n}\sum_{i=1}^{n}\left[w_0 \cdot \sum_{j=1}^{d}(x_j^i \cdot w_j)\right]$$

$$+ \text{плюс другие удвоенные без } w_0$$

Веса: больше предобработок

- Рассмотрим:

$$
\begin{aligned}
\frac{2}{n}\sum_{i=1}^{n}\left[w_0\cdot\sum_{j=1}^{d}(x_j^i\cdot w_j)\right]^2 &= \frac{2}{n}\cdot w_0\cdot\left[\sum_{i=1}^{n}\sum_{j=1}^{d}(x_j^i\cdot w_j)\right] \\
&= 2w_0\cdot\left[\sum_{j=1}^{d}\frac{1}{n}\sum_{i=1}^{n}(x_j^i\cdot w_j)\right] \\
&= 2w_0\cdot\left[\sum_{j=1}^{d}\left(\frac{1}{n}\sum_{i=1}^{n}x_j^i\right)\cdot w_j\right] \\
&= 0
\end{aligned}
$$

Веса: больше предобработок

- Поэтому исходная задача
$$\frac{1}{n}\sum_{i=1}^{n}(w_0 + \langle x^i, w\rangle)^2$$
с точки зрения оптимизации по w_0 есть просто $\min_{w_0 \in \mathbb{R}} w_0^2$.
- Получается, что при такой предобработке, не нужно искать w_0. Но такая выкладка справедлива только для квадратичной функции потерь.

Веса: больше предобработок

- Поэтому исходная задача
$$\frac{1}{n}\sum_{i=1}^{n}(w_0 + \langle x^i, w\rangle)^2$$
с точки зрения оптимизации по w_0 есть просто $\min_{w_0\in\mathbb{R}} w_0^2$.
- Получается, что при такой предобработке, не нужно искать w_0. Но такая выкладка справедлива только для квадратичной функции потерь.
- Существуют и другие классические техники. Например, вместо того, чтобы загонять каждый признак в отрезок длины 1 можно сделать похожий трюк, называемый нормализацией.
- Потребуем, чтобы
$$\frac{1}{n}\sum_{i=1}^{n}(x_j^i)^2 = 1.$$
Так будет, если умножим x_j^i на $s(j) = \sqrt{n/\sum_{i=1}^{n}(x_j^i)^2}$.

Максимум правдоподобия

- Но кажется, мы слишком усложняем. Предположим, что некоторая переменная y зависит от переменных $x_1, x_2, x_3, \ldots, x_d$ линейным образом:

$$y(x_1, \ldots, x_d) = w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_d x_d,$$

где коэффициенты $w_0, \ldots, w_d$ нам неизвестны. Предположим, что мы хотим найти эти коэффициенты, измеряя переменную y при различных значениях $x_1, \ldots, x_d$. Казалось бы, в этом нет ничего сложного, ведь для решения системы достаточно провести $d + 1$ измерений (как в примере из трех строк выше).
Вопрос: Какая проблема?

Максимум правдоподобия

В действительности может все быть значительно сложнее. Например,

1. в реальности зависимость далеко не линейная, но мы просто пытаемся приблизить ее линейной;
2. измерения производятся с некоторой погрешностью.

Максимум правдоподобия

- Рассмотрим вторую постановку. В частности, предположим, что для заданного набора $x_1^i, x_2^i, \ldots, x_d^i$ мы измеряем

$$y_i = w_0 + x_1^i w_1 + \ldots + x_d^i w_d + \xi_i,$$

где $\xi_i \sim \mathcal{N}(0, \sigma^2)$.

Максимум правдоподобия

- Рассмотрим вторую постановку. В частности, предположим, что для заданного набора $x_1^i, x_2^i, \ldots, x_d^i$ мы измеряем
$$y_i = w_0 + x_1^i w_1 + \ldots + x_d^i w_d + \xi_i,$$
где $\xi_i \sim \mathcal{N}(0, \sigma^2)$.
- Другими словами, мы предполагаем, что
$$y_i \sim \mathcal{N}(w_0 + x_1^i w_1 + \ldots + x_d^i w_d, \sigma^2),$$
где параметры $w = (w_0, \ldots, w_d)^\top$ должны быть найдены по выборке $\{y_i\}_{i=1}^n$ (мы будем считать, что $y_1, \ldots, y_n$ – независимые случайные величины).
Вопрос: Как лучше выбрать параметры $w_0, \ldots, w_d$?

Статистический подход: линейная регрессия

- Можно, например, рассмотреть оценку максимального правдоподобия:

$$\begin{aligned}\hat{w} &= \arg\max_{w \in \mathbb{R}^{d+1}} \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \langle x^i, w\rangle)^2\right) \\ &= \arg\max_{w \in \mathbb{R}^{d+1}} \left[\ln\left(\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \langle x^i, w\rangle)^2\right)\right)\right].\end{aligned}$$

Максимум правдоподобия

- Поскольку логарифм произведения равен сумме логарифмов, а аддитивные и мультипликативные константы не меняют точку оптимума, получаем:

$$\begin{aligned}\hat{w} &= \arg\max_{w \in \mathbb{R}^{d+1}} \left\{ \text{Const} + \sum_{i=1}^{n} -\frac{1}{2\sigma^2}(y_i - \langle x^i, w \rangle)^2 \right\} \\ &= \arg\min_{w \in \mathbb{R}^{d+1}} \sum_{i=1}^{n} (y_i - \langle x^i, w \rangle)^2 \\ &= \arg\min_{w \in \mathbb{R}^{d+1}} \frac{1}{n} \|Xw - y\|_2^2,\end{aligned}$$

  где X составлена из строк $(x^i)^\top$.
- Обнаружили связь минимизации эмпирического риска и статистического подхода.

Функции потерь: MSE

Полученная задача минимизации также называется задачей минимизации **квадратичных потерь** (Mean Squared Error, MSE).

MSE

Квадратичной функцией потерь (MSE) называется функция вида

$$\begin{aligned}\mathcal{L}_{\mathrm{MSE}} &= \frac{1}{2}\|Xw - y\|_2^2\\ &= \frac{1}{2n}\sum_{i=1}^{n}(y_i - \langle x^i, w\rangle)^2.\end{aligned}$$

Функции потерь: MSE

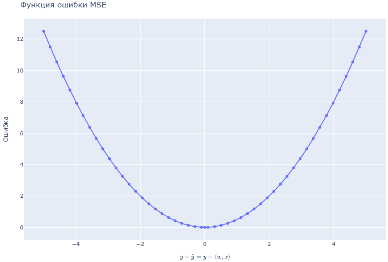
В случае переопределенной системы (когда $\min \mathcal{L}_{\mathrm{MSE}} = 0$) имеется явный вид решения $\hat{w}$. Это следует напрямую из условий оптимума:

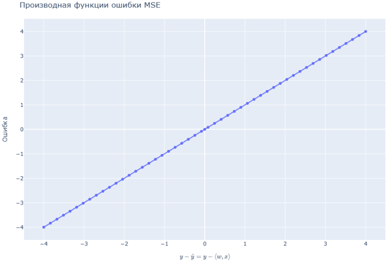
$$
\begin{aligned}
\nabla_w \mathcal{L}_{\mathrm{MSE}} \Big|_{\hat{w}} &= 0, \\
\nabla_w \left[\frac{1}{2} \|Xw - y\|_2^2 \right] \Bigg|_{\hat{w}} &= 0, \\
X^\top (X\hat{w} - y) &= 0, \\
\hat{w} &= (X^\top X)^{-1} X^\top y.
\end{aligned}
$$

Однако, несмотря на распространенность, MSE далеко не единственный способ измерения расстояния.

Функции потерь: MSE (пример)

Визуализация квадратичной функции потерь:





Функции потерь: MAE

Если MSE, по сути, является евклидовым расстоянием, то MAE (Mean Absolute Error) – это расстояние по ℓ_1-норме.

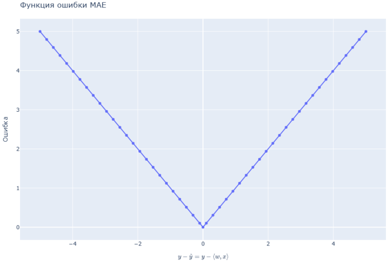
MAE

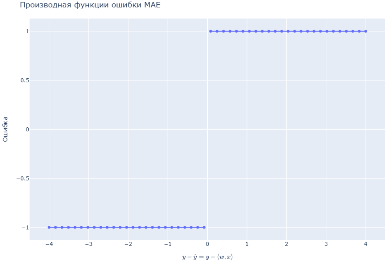
Абсолютной функцией потерь (MAE) называется функция вида

$$\begin{aligned} \mathcal{L}_{\mathrm{MAE}} &= \|Xw - y\|_1 \\ &= \frac{1}{n} \sum_{i=1}^{n} \left| y_i - \langle x^i, w \rangle \right|. \end{aligned}$$

Функции потерь: MAE (пример)

Визуализация абсолютной функции потерь:



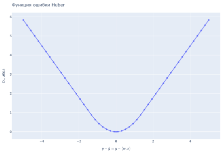


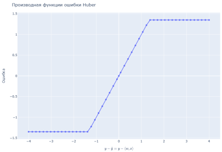
Функции потерь: иные

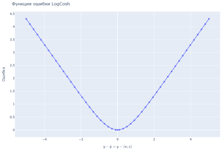
Есть еще парочка широко используемых функций потерь:

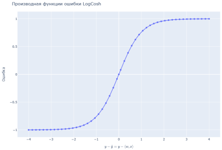
$$
\begin{aligned}
\mathcal{L}_{\text{Huber}} &= \begin{cases} \frac{1}{2n}\sum\limits_{i=1}^{n}(y_i - \langle x^i, w\rangle)^2, & \text{if } y_i - \langle x^i, w\rangle \le \delta, \\ \delta \cdot \left(\frac{1}{n}\sum\limits_{i=1}^{n}\left|y_i - \langle x^i, w\rangle\right| - \frac{1}{2}\delta\right), & \text{else.} \end{cases} \\
\mathcal{L}_{\text{LogCosh}} &= \frac{1}{n}\sum_{i=1}^{n}\log\left[\cosh(y_i - \langle x^i, w\rangle)\right] \\
\mathcal{L}_{\text{MAPE}} &= \frac{1}{n}\sum_{i=1}^{n}\frac{y_i - \langle x^i, w\rangle}{y_i}\cdot 100\%
\end{aligned}
$$

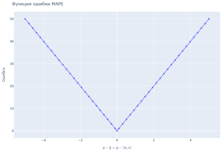
Визуализация функций потерь (слева функции, справа – их производные):

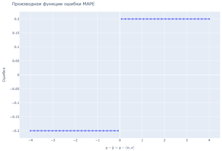




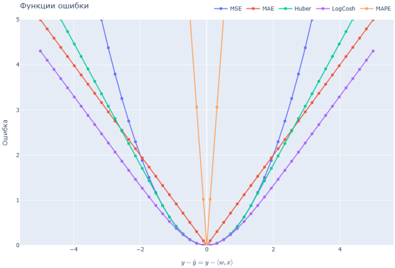


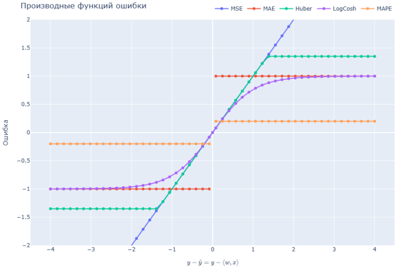






Функции потерь: совместные графики





Классификация

Перейдем к теперь постановки задачи бинарной классификации ($|\mathcal{Y}| = 2$) с метками классов $\{-1, +1\}$.

Вопрос. Как перейти от результатов регрессии с квадратичной функцией потерь к получению предсказания меток классов?

Классификация

Перейдем к теперь постановки задачи бинарной классификации ($|\mathcal{Y}| = 2$) с метками классов $\{-1, +1\}$.

Вопрос. Как перейти от результатов регрессии с квадратичной функцией потерь к получению предсказания меток классов?

Нетрудно догадаться, что используя функцию sign (взятие знака), мы можем преобразовать наше непрерывное предсказание в дискретное:

$$\mathrm{sign}\left(w_0 + \sum_{j=1}^{d} x_j w_j\right) \to \{-1, +1\}\,.$$

Классификация

Перейдем к теперь постановки задачи бинарной классификации ($|\mathcal{Y}| = 2$) с метками классов $\{-1, +1\}$.

Вопрос. Как перейти от результатов регрессии с квадратичной функцией потерь к получению предсказания меток классов?

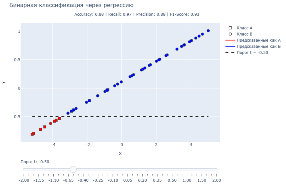
Нетрудно догадаться, что используя функцию sign (взятие знака), мы можем преобразовать наше непрерывное предсказание в дискретное:

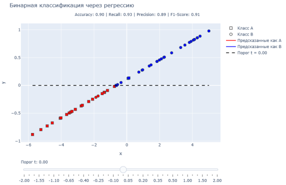
$$\mathrm{sign}\left(w_0 + \sum_{j=1}^{d} x_j w_j\right) \to \{-1, +1\}\,.$$

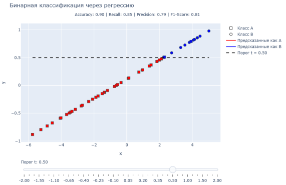
В качестве threshold-а (разделяющего параметра) здесь используется 0, однако, мы вольны выбирать его произвольно (например, положив равным 0.5).

Классификация

Рассмотрим различные значения threshold-а на синтетическом датасете для задачи бинарной классификации. Зависимости от выбранного значения сильно меняются предсказания меток классов. (интерактивный график доступен в приложенном ноутбуке)







Классификация

Вопрос: А почему нам тогда сразу не минимизировать функции вида

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\left[\text{sign}\left(w_0 + \langle x_i, w\rangle\right) \neq y_i\right]?$$

Классификация

Вопрос: А почему нам тогда сразу не минимизировать функции вида

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\left[\text{sign}\left(w_0+\langle x_i,w\rangle\right)\neq y_i\right]?$$

Такую задачу сложно решать численными методами (о них на следующей лекции): считать градиент, а значит в качестве функции потерь ее использовать проблематично. Однако, если мы немного изменим ее вид на

$$1-\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\left[\text{sign}\left(w_0+\langle x_i,w\rangle\right)\neq y_i\right]\to\max_{w},$$

то получим крайне интуитивно понятную структуру. Мы пытаемся максимизировать нашу точность предсказаний, уменьшая количество неверно предсказанных меток. Данная функция является *метрикой* качества нашего предсказания и называется **accuracy**.