Логистическая регрессия

Название курса

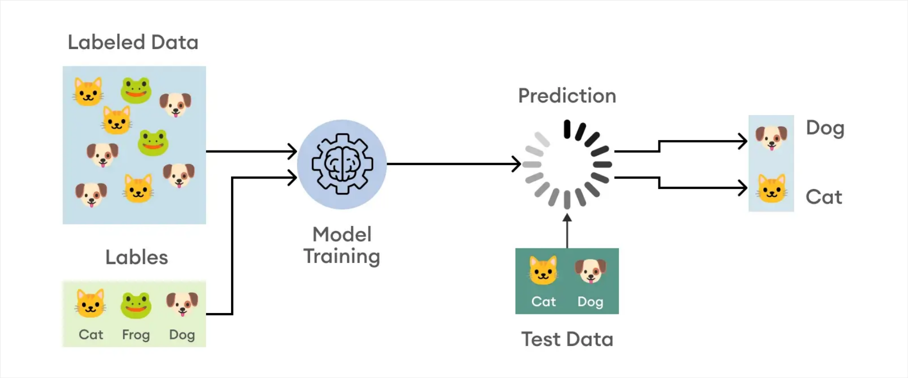
Абдурахмон Садиев

ИСП РАН

13 марта 2025

Примеры задач классификации

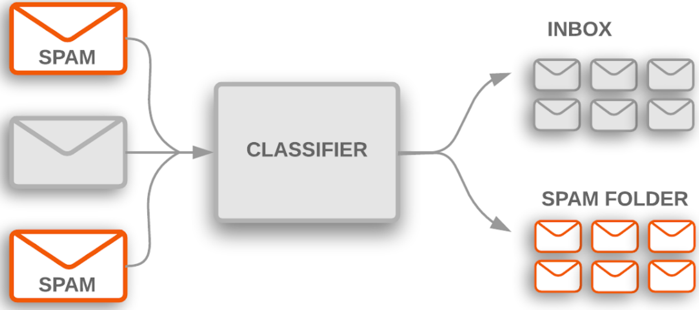
Классификация изображений



Источник

Примеры задач классификации

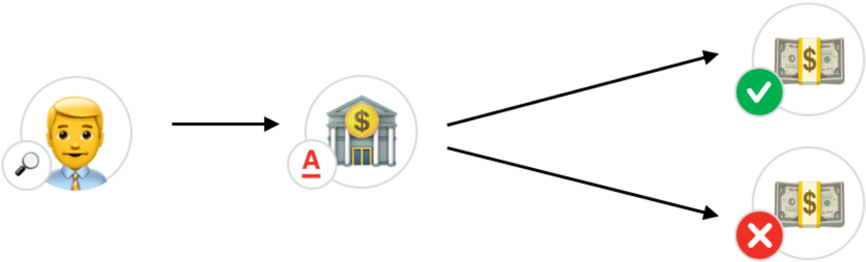
Классификация спама



Источник

Примеры задач классификации

Задача кредитного скоринга



Источник

Постановка задачи

Задача *Бинарная классификация*

Постановка задачи

Задача *Бинарная классификация*

- Пусть $\mathcal{X} = \mathbb{R}^d$ пространство объектов;

Постановка задачи

Задача *Бинарная классификация*

- Пусть $\mathcal{X} = \mathbb{R}^d$ пространство объектов;
- Пусть $\mathcal{Y} = \{-1, 1\}$ (либо $\{0, 1\}$) множество допустимых ответов;

Постановка задачи

Задача *Бинарная классификация*

- Пусть $\mathcal{X} = \mathbb{R}^d$ пространство объектов;
- Пусть $\mathcal{Y} = \{-1, 1\}$ (либо $\{0, 1\}$) множество допустимых ответов;
- $X = \{(x^{(i)}, y^{(i)})\}_{i=1}^{n}$ - обучающая выборка.

Постановка задачи

Задача *Бинарная классификация*

- Пусть $\mathcal{X} = \mathbb{R}^d$ пространство объектов;
- Пусть $\mathcal{Y} = \{-1, 1\}$ (либо $\{0, 1\}$) множество допустимых ответов;
- $X = \{(x^{(i)}, y^{(i)})\}_{i=1}^{n}$ - обучающая выборка.

Наблюдение

Данную задачу можно решать линейной регрессией, **НО**

Постановка задачи

Задача *Бинарная классификация*

- Пусть $\mathcal{X} = \mathbb{R}^d$ пространство объектов;
- Пусть $\mathcal{Y} = \{-1, 1\}$ (либо $\{0, 1\}$) множество допустимых ответов;
- $X = \{(x^{(i)}, y^{(i)})\}_{i=1}^n$ - обучающая выборка.

Наблюдение

Данную задачу можно решать линейной регрессией, **НО**

- Можно построить пример, где данный метод работает плохо;

Постановка задачи

Задача *Бинарная классификация*

- Пусть $\mathcal{X} = \mathbb{R}^d$ пространство объектов;
- Пусть $\mathcal{Y} = \{-1, 1\}$ (либо $\{0, 1\}$) множество допустимых ответов;
- $X = \{(x^{(i)}, y^{(i)})\}_{i=1}^{n}$ - обучающая выборка.

Наблюдение

Данную задачу можно решать линейной регрессией, **НО**

- Можно построить пример, где данный метод работает плохо;
- Интуитивно это не имеет смысла.

Линейная модель классификации

Определение

Линейная модель классификации определяется следующим образом:

$$\operatorname{sign}\left(\langle w, x\rangle + w_0\right) = \operatorname{sign}\left(\sum_{j=1}^{d} w_j x_j + w_0\right), \tag{1}$$

где $w \in \mathbb{R}^d$ - вектор весов, $w_0 \in \mathbb{R}$ - сдвиг.

Линейная модель классификации

Определение

Линейная модель классификации определяется следующим образом:

$$\operatorname{sign}\left(\langle w, x\rangle + w_0\right) = \operatorname{sign}\left(\sum_{j=1}^{d} w_j x_j + w_0\right), \qquad (1)$$

где $w \in \mathbb{R}^d$ - вектор весов, $w_0 \in \mathbb{R}$ - сдвиг. Если предположить, что в данных есть $x_0 = 1$, то нет необходимости вводить сдвиг w_0, т.е.

$$\operatorname{sign}\left(\langle w, x\rangle\right).$$

Функция потерь

Вопрос: *Как обучать?*

Функция потерь

Вопрос: *Как обучать?*

Ответ: *Максимизировать долю правильных ответов:*

Функция потерь

Вопрос: *Как обучать?*

Ответ: *Максимизировать долю правильных ответов:*

Доля правильных ответов (accuracy)

$$\max_w \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}[\mathrm{sign}(\langle w, x^{(i)} \rangle) = y^{(i)}]. \tag{2}$$

Функция потерь

Вопрос: *Как обучать?*

Ответ: *Максимизировать долю правильных ответов:*

Доля правильных ответов (accuracy)

$$\max_{w} \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}[\operatorname{sign}(\langle w, x^{(i)} \rangle) = y^{(i)}]. \tag{2}$$

Или эквивалентно минимизировать долю неверных ответов

Функция потерь

Вопрос: *Как обучать?*

Ответ: *Максимизировать долю правильных ответов:*

Доля правильных ответов (accuracy)

$$\max_w \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[\operatorname{sign}(\langle w, x^{(i)}\rangle) = y^{(i)}]. \tag{2}$$

Или эквивалентно минимизировать долю неверных ответов

$$\max_w \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[\operatorname{sign}(\langle w, x^{(i)}\rangle) = y^{(i)}] = 1 - \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[\operatorname{sign}(\langle w, x^{(i)}\rangle) \neq y^{(i)}]$$

Функция потерь

Вопрос: *Как обучать?*

Ответ: *Максимизировать долю правильных ответов:*

Доля правильных ответов (accuracy)

$$\max_w \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[\operatorname{sign}(\langle w, x^{(i)}\rangle) = y^{(i)}]. \tag{2}$$

Или эквивалентно минимизировать долю неверных ответов

$$\max_w \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[\operatorname{sign}(\langle w, x^{(i)}\rangle) = y^{(i)}] = 1 - \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[\operatorname{sign}(\langle w, x^{(i)}\rangle) \neq y^{(i)}]$$

$$\Rightarrow \quad \min_w \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[\operatorname{sign}(\langle w, x^{(i)}\rangle) \neq y^{(i)}].$$

Как обучать?

Задача оптимизации:

$$\min_{w}\left\{\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}[\operatorname{sign}\left(\langle w, x^{(i)}\rangle\right)\neq y^{(i)}]\right\}.$$

Как обучать?

Задача оптимизации:

$$\min_w \left\{ \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}[\operatorname{sign}\left(\langle w, x^{(i)} \rangle\right) \neq y^{(i)}] \right\}.$$

Проблемы:

- Целевая функция дискретна относительно весов.
- Возможно наличие множества глобальных минимумов

Как обучать?

Задача оптимизации:

$$\min_w \left\{ \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}[\operatorname{sign}\left(\langle w, x^{(i)} \rangle\right) \neq y^{(i)}] \right\}.$$

Проблемы:

- Целевая функция дискретна относительно весов.
- Возможно наличие множества глобальных минимумов

Решение: Свести задачу к минимизации гладкого функционала.

Отступы

Задача оптимизации:

$$\min_w \left\{ \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}[\text{sign}\left(\langle w, x^{(i)} \rangle\right) \neq y^{(i)}] \right\}.$$

Отступы

Задача оптимизации:

$$\min_w \left\{ \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}[\text{sign}\left(\langle w, x^{(i)} \rangle\right) \neq y^{(i)}] \right\}.$$

Наблюдение: Заметим, что

Отступы

Задача оптимизации:

$$\min_{w} \left\{ \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}[\operatorname{sign}\left(\langle w, x^{(i)} \rangle\right) \neq y^{(i)}] \right\}.$$

Наблюдение: Заметим, что

$$y^{(i)} \cdot \langle w, x^{(i)} \rangle \quad > \quad 0, \text{ если } y^{(i)} = \operatorname{sign}\left(\langle w, x^{(i)} \rangle\right);$$

Отступы

Задача оптимизации:

$$\min_{w}\left\{\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}[\operatorname{sign}\left(\langle w, x^{(i)}\rangle\right) \neq y^{(i)}]\right\}.$$

Наблюдение: Заметим, что

$$\begin{aligned} y^{(i)}\cdot\langle w, x^{(i)}\rangle &> 0, \text{ если } y^{(i)} = \operatorname{sign}\left(\langle w, x^{(i)}\rangle\right); \\ y^{(i)}\cdot\langle w, x^{(i)}\rangle &< 0, \text{ иначе.} \end{aligned}$$

Отступы

Задача оптимизации:

$$\min_{w}\left\{\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}[\operatorname{sign}\left(\langle w, x^{(i)}\rangle\right) \neq y^{(i)}]\right\}.$$

Наблюдение: Заметим, что

$$\begin{aligned} y^{(i)}\cdot\langle w, x^{(i)}\rangle &> 0, \text{ если } y^{(i)} = \operatorname{sign}\left(\langle w, x^{(i)}\rangle\right); \\ y^{(i)}\cdot\langle w, x^{(i)}\rangle &< 0, \text{ иначе.} \end{aligned}$$

Величина $M_i = y^{(i)}\cdot\langle w, x^{(i)}\rangle$ называется *отступом*.

Верхние оценки

Задача оптимизации

$$\min_{w} \frac{1}{n} \sum_{i=1}^{n} h\left(M_i\right),$$

где $M_i = y^{(i)} \cdot \langle w, x^{(i)} \rangle$ и $h(M) = [M < 0]$.

Верхние оценки

Задача оптимизации

$$\min_w \frac{1}{n} \sum_{i=1}^{n} h\left(M_i\right),$$

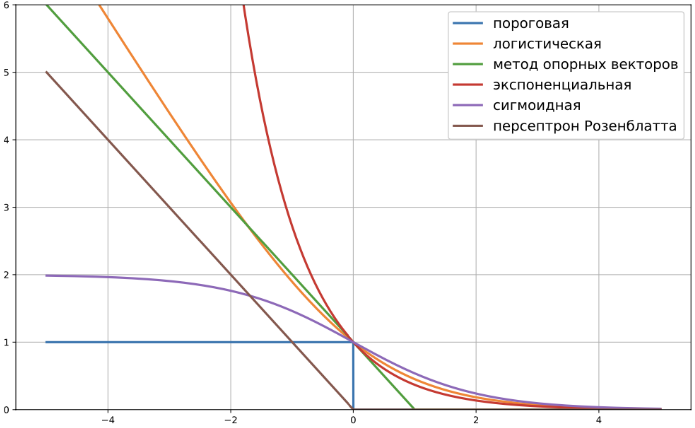
где $M_i = y^{(i)} \cdot \langle w, x^{(i)} \rangle$ и $h(M) = [M < 0]$.

Верхние оценки: Оценим $h(M)$ сверху гладкой функций $\widetilde{h}(M)$, т.е.

$$h(M) \ \leq \ \widetilde{h}(M).$$

Верхние оценки: Оценим $h(M)$ сверху гладкой функций $\widetilde{h}(M)$, т.е.

$$h(M) \leq \widetilde{h}(M).$$



Верхние оценки

Примеры верхних оценок

- $\widetilde{h}(M) = \log\left(1 + e^{-M}\right)$ - логистическая функция потерь;
- $\widetilde{h}(M) = (1 - M)_+ = \max\{0, 1 - M\}$ - кусочно-линейная функция потерь (используется в методе опорных векторов);
- $\widetilde{h}(M) = (-M)_+ = \max\{0, -M\}$ - кусочно-линейная функция потерь (соответствует персептрону Розенблатта);
- $\widetilde{h}(M) = e^{-M}$ - экспоненциальная функция потерь;
- $\widetilde{h}(M) = \frac{2}{1+e^M}$ - сигмоидная функция потерь.

Логистическая регрессия

Задача оптимизации:

$$\min_{w} \left\{ \frac{1}{n} \sum_{i=1}^{n} \widetilde{h} \left(y^{(i)} \langle w, x^{(i)} \rangle \right) \right\}, \tag{3}$$

где для логистической регрессии $\widetilde{h}(M) = \log \left(1 + e^{-M}\right)$.

Логистическая регрессия

Задача оптимизации:

$$\min_w \left\{ \frac{1}{n} \sum_{i=1}^{n} \widetilde{h}\left(y^{(i)} \langle w, x^{(i)} \rangle\right) \right\}, \tag{3}$$

где для логистической регрессии $\widetilde{h}(M) = \log\left(1 + e^{-M}\right)$.

Логистическая регрессия

$$\min_w \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + \exp\left(-y^{(i)} \langle w, x^{(i)} . \rangle\right)\right)$$

Логистическая регрессия

Задача оптимизации:

$$\min_{w}\left\{\mathcal{L}(w)=\frac{1}{n}\sum_{i=1}^{n}\underbrace{\log\left(1+exp\left(-y^{(i)}\langle w, x^{(i)}\rangle\right)\right)}_{=\ell_i(w)}\right\}.$$

Некоторые свойства:

- Каждая функция ℓ_i является выпуклой и $\frac{\|x^{(i)}\|^2}{4}$-гладкой;
- Функция $\mathcal{L}$ является выпуклой и $\frac{1}{4}\lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^{n}x^{(i)}(x^{(i)})^\top\right)$-гладкой.

Оценивание вероятностей

Свойство

Основное свойство логистической регрессии: она корректно оценивает вероятность принадлежности объекта к каждому из классов.

Оценивание вероятностей

Свойство

Основное свойство логистической регрессии: она корректно оценивает вероятность принадлежности объекта к каждому из классов.

- Зафиксируем $x \in \mathcal{X}$;

Оценивание вероятностей

Свойство

Основное свойство логистической регрессии: она корректно оценивает вероятность принадлежности объекта к каждому из классов.

- Зафиксируем $x \in \mathcal{X}$;
- $p(y = 1|x)$ - вероятность того, что объект x будет принадлежать классу 1;

Оценивание вероятностей

Свойство

Основное свойство логистической регрессии: она корректно оценивает вероятность принадлежности объекта к каждому из классов.

- Зафиксируем $x \in \mathcal{X}$;
- $p(y = 1|x)$ - вероятность того, что объект x будет принадлежать классу 1;
- Алгоритм $b(x)$ возвращает числа из отрезка $[0, 1]$.

Цель: выбрать для него такую процедуру обучения, что в точке x ему будет оптимально выдавать число $p(y = 1|x)$.

Оценивание вероятностей

Если в выборке объект x встречается m раз с ответом $\{y_1, \ldots, y_m\}$, то имеем следующее требование

Оценивание вероятностей

Если в выборке объект x встречается m раз с ответом $\{y_1, \dots, y_m\}$, то имеем следующее требование

$$\underset{b \in \mathbb{R}}{\operatorname{argmin}} \frac{1}{m} \sum_{i=1}^{m} L(y, b) \approx p(y = 1|x). \tag{4}$$

Оценивание вероятностей

Если в выборке объект x встречается m раз с ответом $\{y_1, \dots, y_m\}$, то имеем следующее требование

$$\underset{b \in \mathbb{R}}{\operatorname{argmin}} \frac{1}{m} \sum_{i=1}^{m} L(y, b) \approx p(y = 1|x). \qquad (4)$$

При стремлении m к бесконечности получим, что функционал стремится к матожиданию ошибки:

Оценивание вероятностей

Если в выборке объект x встречается m раз с ответом $\{y_1, \dots, y_m\}$, то имеем следующее требование

$$\underset{b \in \mathbb{R}}{\operatorname{argmin}} \frac{1}{m} \sum_{i=1}^{m} L(y, b) \approx p(y = 1|x). \tag{4}$$

При стремлении m к бесконечности получим, что функционал стремится к матожиданию ошибки:

$$\underset{b \in \mathbb{R}}{\operatorname{argmin}} \mathbb{E}\left[L(y_i, b)|x\right] = p(y = 1|x). \tag{5}$$

Оценивание вероятностей

Пример 1

Покажите, что квадратичная функция потерь $L(y, b) = (\mathbb{I}[y = +1] - b)^2$ позволяет предсказывать корректные вероятности.

Оценивание вероятностей

Пример 1

Покажите, что квадратичная функция потерь $L(y, b) = (\mathbb{I}[y = +1] - b)^2$ позволяет предсказывать корректные вероятности.

Запишем матожидание функции потерь в точке x:

$$\begin{aligned} \mathbb{E}\left[L(y, b)|x\right] &= \mathbb{E}\left[\mathbb{I}[y = +1](1 - b)^2 + \mathbb{I}[y = -1]b^2|x\right] \\ &= p(y = +1|x)(1 - b)^2 + (1 - p(y = +1|x))\, b^2. \end{aligned}$$

Оценивание вероятностей

Пример 1

Покажите, что квадратичная функция потерь
$L(y, b) = (\mathbb{I}[y = +1] - b)^2$ позволяет предсказывать корректные вероятности.

Оценивание вероятностей

Пример 1

Покажите, что квадратичная функция потерь $L(y, b) = (\mathbb{I}[y = +1] - b)^2$ позволяет предсказывать корректные вероятности.

Матожидание функции потерь в точке x:

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

Оценивание вероятностей

Пример 1

Покажите, что квадратичная функция потерь $L(y, b) = (\mathbb{I}[y = +1] - b)^2$ позволяет предсказывать корректные вероятности.

Матожидание функции потерь в точке x:

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

Продифференцируем по b:

$$\frac{\partial}{\partial b}\mathbb{E}\left[L(y, b)|x\right]$$

Оценивание вероятностей

Пример 1

Покажите, что квадратичная функция потерь $L(y, b) = (\mathbb{I}[y = +1] - b)^2$ позволяет предсказывать корректные вероятности.

Матожидание функции потерь в точке x:

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

Продифференцируем по b:

$$\frac{\partial}{\partial b}\mathbb{E}\left[L(y, b)|x\right] \quad = \quad 2p(y = +1|x)(b - 1) + 2\left(1 - p(y = +1|x)\right) b$$

Оценивание вероятностей

Пример 1

Покажите, что квадратичная функция потерь $L(y, b) = (\mathbb{I}[y = +1] - b)^2$ позволяет предсказывать корректные вероятности.

Матожидание функции потерь в точке x:

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

Продифференцируем по b:

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

Оценивание вероятностей

Пример 1

Покажите, что квадратичная функция потерь $L(y, b) = (\mathbb{I}[y = +1] - b)^2$ позволяет предсказывать корректные вероятности.

Матожидание функции потерь в точке x:

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

Продифференцируем по b:

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

Легко видеть, что оптимальный ответ алгоритма действительно равен вероятности: $b = p(y = +1|x)$

Оценивание вероятностей

Пример 2

Покажите, что абсолютная функция потерь $L(y, b) = |\mathbb{I}[y = +1]b|, \ b \in [0; 1]$, не позволяет предсказывать корректные вероятности.

Оценивание вероятностей

Пример 2

Покажите, что абсолютная функция потерь $L(y, b) = |\mathbb{I}[y = +1]b|, \;\; b \in [0; 1]$, не позволяет предсказывать корректные вероятности.

Запишем матожидание функции потерь в точке x:

$$\mathbb{E}\left[L(y, b)|x\right]$$

Оценивание вероятностей

Пример 2

Покажите, что абсолютная функция потерь $L(y, b) = |\mathbb{I}[y = +1]b|, \; b \in [0; 1]$, не позволяет предсказывать корректные вероятности.

Запишем матожидание функции потерь в точке x:

$$\mathbb{E}\left[L(y, b)|x\right] \quad = \quad \mathbb{E}\left[\mathbb{I}[y = +1]|1 - b| + \mathbb{I}[y = -1]|b||x\right]$$

Оценивание вероятностей

Пример 2

Покажите, что абсолютная функция потерь $L(y, b) = |\mathbb{I}[y = +1]b|, \ b \in [0; 1]$, не позволяет предсказывать корректные вероятности.

Запишем матожидание функции потерь в точке x:

$$\begin{aligned}
\mathbb{E}\left[L(y, b)|x\right] &= \mathbb{E}\left[\mathbb{I}[y = +1]|1 - b| + \mathbb{I}[y = -1]|b||x\right] \\
&= p(y = +1|x)(1 - b) + (1 - p(y = +1|x))\, b.
\end{aligned}$$

Оценивание вероятностей

Пример 1

Покажите, что квадратичная функция потерь $L(y, b) = (\mathbb{I}[y = +1] - b)^2$ позволяет предсказывать корректные вероятности.

Матожидание функции потерь в точке x:

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

Оценивание вероятностей

Пример 1

Покажите, что квадратичная функция потерь $L(y, b) = (\mathbb{I}[y = +1] - b)^2$ позволяет предсказывать корректные вероятности.

Матожидание функции потерь в точке x:

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

Продифференцируем по b:

Оценивание вероятностей

Пример 1

Покажите, что квадратичная функция потерь $L(y, b) = (\mathbb{I}[y = +1] - b)^2$ позволяет предсказывать корректные вероятности.

Матожидание функции потерь в точке x:

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

Продифференцируем по b:

$$\frac{\partial}{\partial b} \mathbb{E}\left[L(y, b) | x\right]$$

Оценивание вероятностей

Пример 1

Покажите, что квадратичная функция потерь $L(y, b) = (\mathbb{I}[y = +1] - b)^2$ позволяет предсказывать корректные вероятности.

Матожидание функции потерь в точке x:

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

Продифференцируем по b:

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

Оценивание вероятностей

Пример 1

Покажите, что квадратичная функция потерь $L(y, b) = (\mathbb{I}[y = +1] - b)^2$ позволяет предсказывать корректные вероятности.

Матожидание функции потерь в точке x:

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

Продифференцируем по b:

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

Оценивание вероятностей

Пример 1

Покажите, что квадратичная функция потерь $L(y, b) = (\mathbb{I}[y = +1] - b)^2$ позволяет предсказывать корректные вероятности.

Оценивание вероятностей

Пример 1

Покажите, что квадратичная функция потерь $L(y, b) = (\mathbb{I}[y = +1] - b)^2$ позволяет предсказывать корректные вероятности.

Рассмотрим 2 случая:

Оценивание вероятностей

Пример 1

Покажите, что квадратичная функция потерь $L(y, b) = (\mathbb{I}[y = +1] - b)^2$ позволяет предсказывать корректные вероятности.

Рассмотрим 2 случая:

- $p(y = +1|x) = \frac{1}{2} \Rightarrow$ классификатор не позволяет предсказывать корректную вероятность в точке x (Почему?);

Оценивание вероятностей

Пример 1

Покажите, что квадратичная функция потерь $L(y, b) = (\mathbb{I}[y = +1] - b)^2$ позволяет предсказывать корректные вероятности.

Рассмотрим 2 случая:

- $p(y = +1|x) = \frac{1}{2} \Rightarrow$ классификатор не позволяет предсказывать корректную вероятность в точке x (Почему?);
- $p(y = +1|x) \neq \frac{1}{2} \Rightarrow$ классификатор также не позволяет предсказывать корректную вероятность в точке. (Почему?)

Правдоподобие

Если алгоритм $b(x) \in [0, 1]$ выдает вероятности, то они должны согласовываться с выборкой.

Правдоподобие

Если алгоритм $b(x) \in [0, 1]$ 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

$$b(x^{(i)})^{\mathbb{I}[y_i=+1]}(1 - b(x^{(i)}))^{[y_i=-1]}. \tag{6}$$

Правдоподобие

Если алгоритм $b(x) \in [0,1]$ 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

$$b(x^{(i)})^{\mathbb{I}[y_i=+1]}(1-b(x^{(i)}))^{[y_i=-1]}. \tag{6}$$

Тогда *правдоподобие* выборки:

$$\prod_{i=1}^{n} b(x^{(i)})^{\mathbb{I}[y_i=+1]}(1-b(x^{(i)}))^{\mathbb{I}[y_i=-1]}. \tag{7}$$

Правдоподобие

Минимизация минус логарифма правдоподобия:

$$\min_b \left\{ -\sum_{i=1}^{n} \left(\mathbb{I}[y^{(i)} = +1] \log b(x^{(i)}) + \mathbb{I}[y^{(i)} = -1] \log \left(1 - b(x^{(i)})\right) \right) \right\}.$$

Правдоподобие

Минимизация минус логарифма правдоподобия:

$$\min_b \left\{ -\sum_{i=1}^{n} \left(\mathbb{I}[y^{(i)} = +1] \log b(x^{(i)}) + \mathbb{I}[y^{(i)} = -1] \log \left(1 - b(x^{(i)}) \right) \right) \right\}.$$

Покажем, что она также позволяет корректно предсказывать вероятности:

Правдоподобие

Минимизация минус логарифма правдоподобия:

$$\min_b \left\{ -\sum_{i=1}^{n} \left(\mathbb{I}[y^{(i)} = +1] \log b(x^{(i)}) + \mathbb{I}[y^{(i)} = -1] \log \left(1 - b(x^{(i)})\right) \right) \right\}.$$

Покажем, что она также позволяет корректно предсказывать вероятности:

- Функция потерь:

$$L(y, b) = \mathbb{I}[y = +1] \log b - \mathbb{I}[y = -1] \log (1 - b);$$

Правдоподобие

Минимизация минус логарифма правдоподобия:

$$\min_b \left\{ -\sum_{i=1}^{n} \left(\mathbb{I}[y^{(i)} = +1] \log b(x^{(i)}) + \mathbb{I}[y^{(i)} = -1] \log \left(1 - b(x^{(i)}) \right) \right) \right\}.$$

Покажем, что она также позволяет корректно предсказывать вероятности:

- Функция потерь:

$$L(y, b) = \mathbb{I}[y = +1] \log b - \mathbb{I}[y = -1] \log (1 - b);$$

- Запишем матожидание функции потерь в точке x:

$$\begin{aligned} \mathbb{E}\left[L(y, b) | x\right] &= \mathbb{E}\left[-\mathbb{I}[y = +1] \log b - \mathbb{I}[y = -1] \log (1 - b) \, | x\right] \\ &= -p(y = +1|x) \log b - (1 - p(y = +1|x)) \log(1 - b). \end{aligned}$$

Правдоподобие

Функция потерь:

$$L(y, b) = \mathbb{I}[y = +1] \log b - \mathbb{I}[y = -1] \log (1 - b)\,;$$

Правдоподобие

Функция потерь:

$$L(y, b) = \mathbb{I}[y = +1] \log b - \mathbb{I}[y = -1] \log (1 - b)\,;$$

Матожидание функции потерь в точке x:

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

Правдоподобие

Функция потерь:

$$L(y, b) = \mathbb{I}[y = +1] \log b - \mathbb{I}[y = -1] \log (1 - b) ;$$

Матожидание функции потерь в точке x:

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

Продифференцируем по b:

$$\frac{\partial}{\partial b} \mathbb{E}\left[L(y, b) | x\right]$$

Правдоподобие

Функция потерь:

$$L(y, b) = \mathbb{I}[y = +1] \log b - \mathbb{I}[y = -1] \log (1 - b)\,;$$

Матожидание функции потерь в точке x:

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

Продифференцируем по b:

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

Правдоподобие

Функция потерь:

$$L(y, b) = \mathbb{I}[y = +1] \log b - \mathbb{I}[y = -1] \log (1 - b)\,;$$

Матожидание функции потерь в точке x:

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

Продифференцируем по b:

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

Легко видеть, что оптимальный ответ алгоритма равен вероятности положительного класса: $b = p(y = +1|x)$.

Логистическая регрессия

- Чтобы алгоритм $b(x)$ возвращал числа из отрезка $[0, 1]$, можно положить

$$b(x) = \sigma\left(\langle w, x\rangle\right),$$

где σ - любая монотонно неубывающая функция с областью значений $[0, 1]$.

Логистическая регрессия

- Чтобы алгоритм $b(x)$ возвращал числа из отрезка $[0, 1]$, можно положить

$$b(x) = \sigma\left(\langle w, x\rangle\right),$$

где σ - любая монотонно неубывающая функция с областью значений $[0, 1]$.
- Мы будем использовать **сигмоидную функцию**:

$$\sigma(z) = \frac{1}{1 + e^{-z}}. \tag{8}$$

Логистическая регрессия

Сигмоидная функция

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

Логистическая регрессия

Сигмоидная функция

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

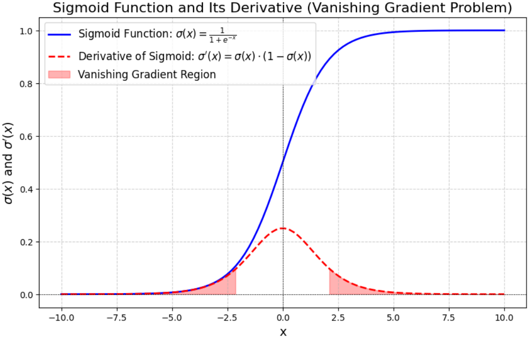
- Ее производная: $\sigma'(z) = \sigma(z)(1 - \sigma(z))$.

Логистическая регрессия

Сигмоидная функция

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

- Ее производная: $\sigma'(z) = \sigma(z)(1 - \sigma(z))$.



Логистическая регрессия

Тогда мы имеем:

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

Логистическая регрессия

Тогда мы имеем:

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

Подставим трансформированный ответ линейной модели в логарифмическую функцию потерь:

$$\mathcal{L}(w, X) \quad =$$

Логистическая регрессия

Тогда мы имеем:

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

Подставим трансформированный ответ линейной модели в логарифмическую функцию потерь:

$$\mathcal{L}(w, X) \quad = \quad -\sum_{i=1}^{n} \left([y^{(i)} = +1] \log \frac{1}{1 + e^{-\langle w,x^{(i)} \rangle}} + [y^{(i)} = -1] \log \frac{e^{-\langle w,x^{(i)} \rangle}}{1 + e^{-\langle w,x^{(i)} \rangle}} \right)$$

Логистическая регрессия

Тогда мы имеем:

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

Подставим трансформированный ответ линейной модели в логарифмическую функцию потерь:

$$\begin{aligned}
\mathcal{L}(w, X) &= -\sum_{i=1}^{n}\left([y^{(i)} = +1]\log\frac{1}{1 + e^{-\langle w,x^{(i)}\rangle}} + [y^{(i)} = -1]\log\frac{e^{-\langle w,x^{(i)}\rangle}}{1 + e^{-\langle w,x^{(i)}\rangle}}\right)\\
&= -\sum_{i=1}^{n}\left([y^{(i)} = +1]\log\frac{1}{1 + e^{-\langle w,x^{(i)}\rangle}} + [y^{(i)} = -1]\log\frac{1}{1 + e^{\langle w,x^{(i)}\rangle}}\right)
\end{aligned}$$

Логистическая регрессия

Тогда мы имеем:

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

Подставим трансформированный ответ линейной модели в логарифмическую функцию потерь:

$$\begin{aligned}
\mathcal{L}(w,X) &= -\sum_{i=1}^{n}\left([y^{(i)}=+1]\log\frac{1}{1+e^{-\langle w,x^{(i)}\rangle}}+[y^{(i)}=-1]\log\frac{e^{-\langle w,x^{(i)}\rangle}}{1+e^{-\langle w,x^{(i)}\rangle}}\right)\\
&= -\sum_{i=1}^{n}\left([y^{(i)}=+1]\log\frac{1}{1+e^{-\langle w,x^{(i)}\rangle}}+[y^{(i)}=-1]\log\frac{1}{1+e^{\langle w,x^{(i)}\rangle}}\right)\\
&= \sum_{i=1}^{n}\log\left(1+\exp\left(-y^{(i)}\langle w,x^{(i)}\rangle\right)\right)
\end{aligned}$$

Постановка задачи

Задача *Многоклассовая классификация*

Постановка задачи

Задача *Многоклассовая классификация*

- Пусть $\mathcal{X} = \mathbb{R}^d$ пространство объектов;

Постановка задачи

Задача *Многоклассовая классификация*

- Пусть $\mathcal{X} = \mathbb{R}^d$ пространство объектов;
- Пусть $\mathcal{Y} = \{1, \cdot, K\}$ множество допустимых ответов;

Постановка задачи

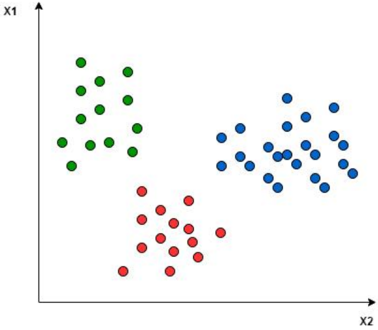
Задача *Многоклассовая классификация*

- Пусть $\mathcal{X} = \mathbb{R}^d$ пространство объектов;
- Пусть $\mathcal{Y} = \{1, \cdot, K\}$ множество допустимых ответов;
- $X = \{(x^{(i)}, y^{(i)})\}_{i=1}^{n}$ - обучающая выборка.

Постановка задачи

Задача *Многоклассовая классификация*

- Пусть $\mathcal{X} = \mathbb{R}^d$ пространство объектов;
- Пусть $\mathcal{Y} = \{1, \cdot, K\}$ множество допустимых ответов;
- $X = \{(x^{(i)}, y^{(i)})\}_{i=1}^n$ - обучающая выборка.



Один против всех (one-versus-all)

Один против всех (one-versus-all)

- Построим K линейных моделей: $b_k(x) = \langle w_k, x \rangle + w_{0,k}$;

Один против всех (one-versus-all)

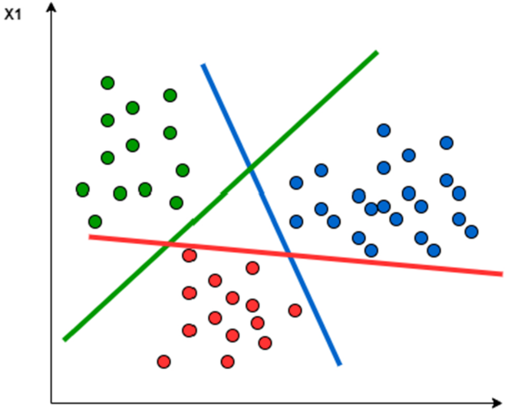
- Построим K линейных моделей: $b_k(x) = \langle w_k, x \rangle + w_{0,k}$;
- b_k будем обучать по выборке $\{(x_i, 2\mathbb{I}[y_i = k] - 1)\}_{i=1}^{n}$;

Один против всех (one-versus-all)

- Построим K линейных моделей: $b_k(x) = \langle w_k, x \rangle + w_{0,k}$;
- b_k будем обучать по выборке $\{(x_i, 2\mathbb{I}[y_i = k] - 1)\}_{i=1}^n$;
- Итоговый классификатор: $a(x) = \operatorname{argmax}_{k \in \{1,\dots,K\}} b_k(x)$.

Один против всех (one-versus-all)

- Построим K линейных моделей: $b_k(x) = \langle w_k, x \rangle + w_{0,k}$;
- b_k будем обучать по выборке $\{(x_i, 2\mathbb{I}[y_i = k] - 1)\}_{i=1}^{n}$;
- Итоговый классификатор: $a(x) = \operatorname{argmax}_{k \in \{1,\dots,K\}} b_k(x)$.



Все против всех (all-versus-all)

Все против всех (all-versus-all)

- Построим C_K^2 линейных моделей: $a_{i,j}(x) = \langle w_{i,j}, x \rangle + w_{0,i,j}$, где $\forall i, j \in \{1, \ldots, K\} : \ i \neq j$;

Все против всех (all-versus-all)

- Построим C_K^2 линейных моделей: $a_{i,j}(x) = \langle w_{i,j}, x\rangle + w_{0,i,j}$, где $\forall i, j \in \{1, \dots, K\} : \; i \neq j$;
- b_k будем обучать по подвыборке $X_{i,j} = \{(x_m, y_m) \in X \mid \mathbb{I}[y_m = i] \text{ или } \mathbb{I}[y_m = j]\}$;

Все против всех (all-versus-all)

- Построим C_K^2 линейных моделей: $a_{i,j}(x) = \langle w_{i,j}, x\rangle + w_{0,i,j}$, где $\forall i, j \in \{1, \ldots, K\} : \; i \neq j$;
- b_k будем обучать по подвыборке $X_{i,j} = \{(x_m, y_m) \in X \mid \mathbb{I}[y_m = i] \text{ или } \mathbb{I}[y_m = j]\}$;
- Итоговый классификатор: $a(x) = \operatorname{argmax}_{k\in\{1,\ldots,K\}} \sum\limits_{i,j:i\neq j}^{K} \mathbb{I}[a_{i,j} = k]$.

Многоклассовая логистическая регрессия

Бинарная логистическая регрессия:

- Построили линейную модель: $b(x) = \langle w, x \rangle + w_0$;
- Перевели прогноз в вероятность с помощью сигмоидной функции;

Многоклассовая логистическая регрессия

Бинарная логистическая регрессия:

- Построили линейную модель: $b(x) = \langle w, x \rangle + w_0$;
- Перевели прогноз в вероятность с помощью сигмоидной функции;

Многоклассовая логистическая регрессия:

- Построим K линейных моделей: $b_k(x) = \langle w_k, x \rangle + w_{0,k}$;
- Как преобразовывать вектор оценок в вектор вероятностей?

SoftMax

SoftMax

Definition

$$\mathrm{SoftMax}\left(z_1, \ldots, z_K\right) = \left(\frac{\exp(z_1)}{\sum_{k=1}^{K} \exp(z_k)}, \ldots, \frac{\exp(z_K)}{\sum_{k=1}^{K} \exp(z_k)}\right) \tag{9}$$

SoftMax

Definition

$$\mathrm{SoftMax}\left(z_1, \dots, z_K\right) = \left(\frac{\exp(z_1)}{\sum_{k=1}^{K}\exp(z_k)}, \dots, \frac{\exp(z_K)}{\sum_{k=1}^{K}\exp(z_k)}\right) \tag{9}$$

В этом случае вероятность k-го класса будет выражаться как

$$P\left(y = k|x, w\right) = \frac{\exp(\langle w_k, x\rangle + w_{0,k})}{\sum_{j=1}^{K}\exp(\langle w_j, x\rangle + w_{0,j})}.$$

SoftMax

Definition

$$\text{SoftMax}\left(z_1, \ldots, z_K\right) = \left(\frac{\exp(z_1)}{\sum_{k=1}^{K} \exp(z_k)}, \ldots, \frac{\exp(z_K)}{\sum_{k=1}^{K} \exp(z_k)}\right) \qquad (9)$$

В этом случае вероятность k-го класса будет выражаться как

$$P\left(y = k | x, w\right) = \frac{\exp(\langle w_k, x\rangle + w_{0,k})}{\sum_{j=1}^{K} \exp(\langle w_j, x\rangle + w_{0,j})}.$$

Обучать эти веса предлагается с помощью метода максимального правдоподобия:

$$\max_{w_1, \ldots, w_K} \sum_{i=1}^{n} P\left(y = y_i | x_i, w\right).$$