Лекция 10

Логистическая регрессия. Метрики в задаче классификации

Курс: Введение в DS на УБ и МиРА (весна, 2022)
Преподаватель: Владимир Омелюсик

6 июня 2022 г.

В предыдущих сериях

- Линейная регрессия с точки зрения машинного обучения.
- Метод k ближайших соседей.

Логистическая регрессия: вводные данные

- Дано: X и y.
- Будем решать задачу бинарной классификации: $y_i \in \{-1, 1\}$.

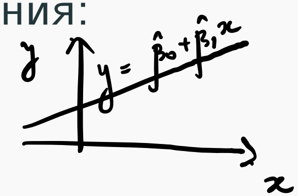
Линейная классификация

- Попробуем как-то использовать линейную модель. Для этого перепишем её в виде скалярного произведения:

$\hat{\beta} = [\hat{\beta}_0, \hat{\beta}_1 \dots \hat{\beta}_k]$

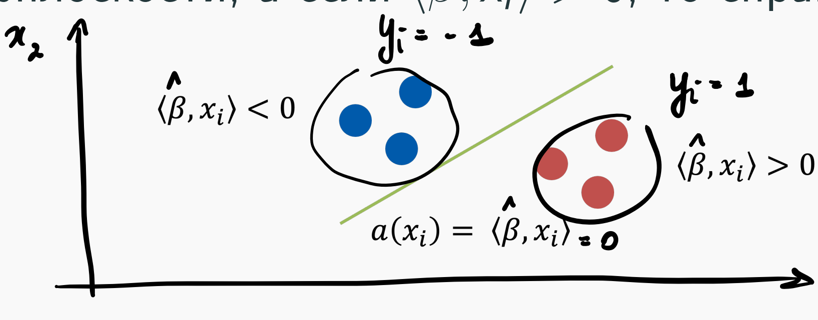
$x_i = [1 \ x_i^1 \ x_i^2 \dots x_i^k]$

$$\boxed{a(x_i) = \langle \hat{\beta}, x_i \rangle = \hat{\beta}_0 + \hat{\beta}_1 x_i^1 + \dots}$$



- Уравнение $\langle \hat{\beta}, x_i \rangle$ задаёт гиперплоскость в пространстве признаков (аналог прямой в двумерном пространстве).
- Если $\langle \hat{\beta}, x_i \rangle < 0$, то объект находится слева от гиперплоскости, а если $\langle \hat{\beta}, x_i \rangle > 0$, то справа.

$$\text{sign}(x) = \begin{cases} 1, & x > 0 \\ 0, & x = 0 \\ -1, & x < 0 \end{cases}$$



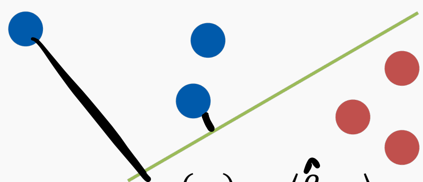
- Предсказания будем делать как $\hat{y}_i = \text{sign}\, a(x_i) = \text{sign}\langle \hat{\beta}, x_i \rangle$

Линейная классификация

- Расстояние от точки до гиперплоскости $\langle\hat{\beta}, x_i\rangle = 0$:

$$\frac{|\langle\hat{\beta}, x_i\rangle|}{\cancel{\|\hat{\beta}\|_2}}$$

- Чем больше $\langle\hat{\beta}, x_i\rangle$, тем дальше объект от гиперплоскости и тем увереннее ответ классификатора.



-1 / 1

\+ \+

- Эту идею можно переписать в виде **отступа** $M_i = y_i\langle\hat{\beta}, x_i\rangle$.
- $M_i > 0 \Rightarrow$ классификатор даёт верный ответ, $M_i < 0 \Rightarrow$ классификатор ошибается.

Обучение линейного классификатора

- Будем штрафовать за неправильную классификацию.

[верно] = 1
[неверно] = 0

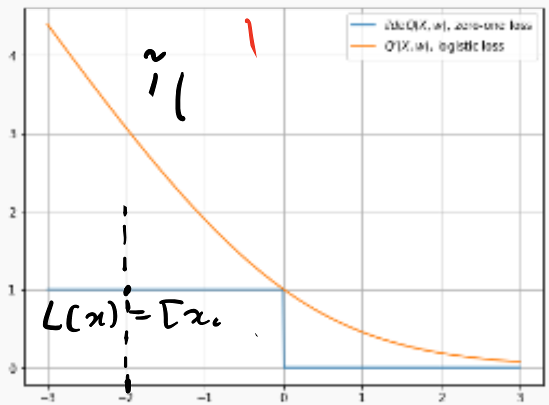
$$L(y_i, \hat{y}_i) = \frac{1}{N}\sum_{i=1}^{N}[\operatorname{sign}(\langle\hat{\beta}, x_i\rangle) \neq y_i]$$

- Функцию потерь также можно записать в терминах отступа:

$$L(y_i, \hat{y}_i) = \frac{1}{N}\sum_{i=1}^{N}[y_i\langle\hat{\beta}, x_i\rangle < 0]$$

Верхняя оценка на функцию потерь

- Проблема: функцию $L(x) = [x < 0]$ нельзя продифференцировать.
- Решение: оценим сверху дифференцируемой функцией потерь:
$$L(X) = [x < 0] \leqslant \tilde{L}(x)$$
- Будем минимизировать $\tilde{L}(x)$ и надеяться, что также проминимизируем $L(x)$.



Логистическая функция потерь

$\log \equiv \ln$

- Используем логистическую функцию потерь.

$$\tilde{L}(x) = \log(1 + e^{-x})$$

- Тогда задачу можно переписать как

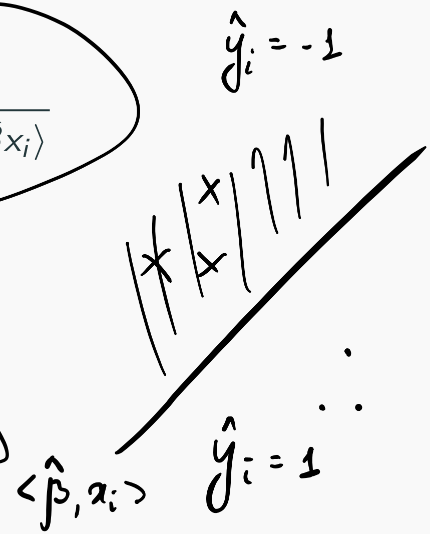
$$\tilde{L}(y_i, \hat{y}_i) = \frac{1}{N}\sum_{i=1}^{N} \log(1 + e^{-\underbrace{y_i\langle\hat{\beta}, x_i\rangle}}) \to \min_{\hat{\beta}}$$

- Такую задачу мы уже имеем решать!

Оптимизация

- Можно посчитать градиент:

$$\nabla_{\hat\beta}\tilde L(y_i,\hat y_i) = -\frac{1}{N}\sum_{i=1}^{N}\frac{y_i x_i}{1+\exp^{y_i\langle\hat\beta x_i\rangle}}$$



- И запустить градиентный спуск:

$$\underbrace{\hat\beta_{t+1} = \hat\beta_t - \alpha\nabla_{\hat\beta}\tilde L(y_i,\hat y_i)|_{\hat\beta_t}}$$

- Можно добавить регуляризацию:

$$\frac{1}{N}\sum_{i=1}^{N}\log(1+e^{-y_i\langle\hat\beta,x_i\rangle}) + \lambda\|\hat\beta\|_2^2 \to \min_{\hat\beta}$$

$\hat y_i = \mathrm{sign}(\langle\hat\beta, x_i\rangle)$

Мягкая и жёсткая классификация

- **Жёсткая классификация** – предсказываем метку класса.
- **Мягкая классификация** – предсказываем вероятность принадлежности к классу.
- От мягкой классификации можно перейти к жёсткой, сравнив вероятность с некоторым порогом t:

$$y_i = [\hat{p}_i > t]$$

$$\hat{p}(x_i \in \text{кл. } 1) = 0.7 \Rightarrow \hat{y}_i = 1$$

$t = 0.5$

Предсказания вероятностей

- Предсказываем метки как

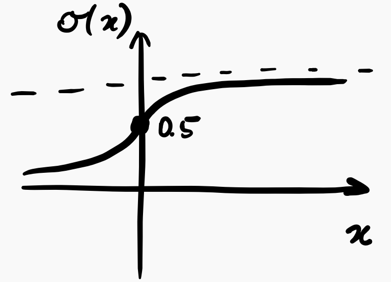
$$\hat{y}_i = \operatorname{sign} a(x_i) = \operatorname{sign}\langle\hat{\beta}x_i\rangle$$

$\hat{p}_i = \hat{P}(x_i \in \text{кл. } 1) =$

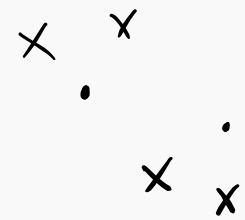
$= \langle \hat{\beta} x_i \rangle$

- Можно ли использовать $a(x_i) = \langle\hat{\beta}x_i\rangle$ для предсказания вероятностей? Нет!
- Для предсказания вероятностей обернём выход модели в некоторую функцию, которая выдаёт числа от 0 до 1. Например, сигмоиду:

$$\hat{p}_i = \sigma(\langle\hat{\beta}x_i\rangle) = \frac{1}{1 + e^{-\langle\hat{\beta}x_i\rangle}}$$



Предсказания вероятностей



- Можно показать, что

$$\tilde{L}(y_i, \hat{y}_i) = \frac{1}{N}\sum_{i=1}^{N} \log(1 + e^{-y_i \langle \hat{\beta}, x_i \rangle}) =$$

$$= -\frac{1}{N}\sum_{i=1}^{N} [y_i = 1] \log \sigma(\langle \hat{\beta} x_i \rangle) + [y_i = -1] \log(1 - \sigma(\langle \hat{\beta} x_i \rangle))$$

- То есть обучая модель по схеме выше, мы обучаем её и правильно предсказывать вероятности.

Accuracy (доля правильных ответов)

≠ точность

* Бинар. классиф.

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}[y_i = \hat{y}_i]$$

$= 1,\ y_i = \hat{y}_i$

$0,\ y_i \neq \hat{y}_i$

- Проста и хорошо интерпретируема.
- Неустойчива к дисбалансу классов.

Модель: $\hat{y}_i = 1$

acc $= 0.99$

100 набл.

99 набл.
$y_i = 1$

1 набл.
$y_i = -1$

Матрица ошибок

истин.

		$y_i = 1$	$y_i = -1$
предск.	$a(x_i) = 1$	True Positive	False Positive
	$a(x_i) = -1$	False Negative	True Negative

В терминах матрицы ошибок:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

Precision (точность)

	$y_i = 1$	$y_i = -1$
$a(x_i) = 1$	True Positive	False Positive
$a(x_i) = -1$	False Negative	True Negative

$$\text{Precision} = \frac{TP}{TP + FP}$$

- Показывает, сколько среди предсказанных положительных объектов действительно положительных.
- Насколько можно доверять классификатору, когда он предсказывает положительный класс?

Recall (полнота)

	$y_i = 1$	$y_i = -1$
$a(x_i) = 1$	True Positive	False Positive
$a(x_i) = -1$	False Negative	True Negative

$$\underset{\text{Recall}}{\cancel{\text{Precision}}} = \frac{TP}{TP + FN}$$

- Показывает, сколько объектов положительного класса было найдено классификатором.

Выбор между Precision и Recall

$y_i = 1 \Rightarrow$ полож.

- Между Precision и Recall существует выбор.

$y_i = -1 \Rightarrow$ отриц.

- Высокая Precision, низкая Recall:
  - Редко ошибаемся при предсказании положительного класса.
  - Находим мало объектов положительного класса.

- Precision и Recall устойчивы к дисбалансу классов.

	$y_i = 1$	$y_i = -1$
$a(x_i) = 1$	10	20
$a(x_i) = -1$	90	10 000

Accuracy = 0.99
Precision = 0.33
Recall = 0.1

100
$y_i = 1$

10 020
$y_i = -1$

F-мера

- Precision и Recall можно объединить в одну метрику:

$$\boxed{F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}}$$

F_β

Порог

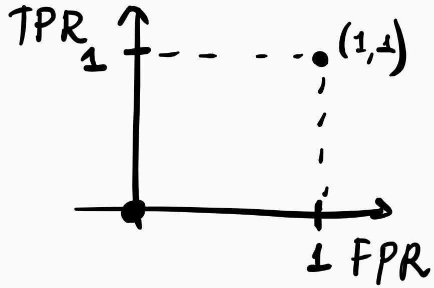
- От мягкой классификации можно перейти к жёсткой, сравнив вероятность с некоторым порогом:

$$y_i = [\sigma(\langle \hat{\beta}, x_i \rangle) > t]$$

- В зависимости от порога будут получаться разные значения TP, FP, FN, TN в матрице ошибок.
- Высокий порог: Precision выше, Recall ниже.

ROC-кривая

- Receiving Operating Characteristic.
- По оси X – False Positive Rate:



$$FPR = \frac{FP}{FP + TN}$$

доля неправ. предск. об. отриц. кл.

- По оси Y – True Positive Rate (Recall):

$$TPR = \text{Recall} = \frac{TP}{TP + FN}$$

доля прав. предск. объек. полож. кл.

	$y_i = 1$	$y_i = -1$
$a(x_i) = 1$	True Positive	False Positive
$a(x_i) = -1$	False Negative	True Negative

Пример

+1 | −1 −1 −1 −1 −1

y	-1	1	-1	1	1	1
$\hat{p}$	0.9	0.7	0.5	0.2	0.1	0.1

$t = 0.8 \Rightarrow$ $\hat{p} > 0.8 \Rightarrow +1$
$\hat{p} \leq 0.8 \Rightarrow -1$

$$\boldsymbol{t = 1}: TPR = \frac{0}{4} = 0, FPR = \frac{0}{2} = 0$$

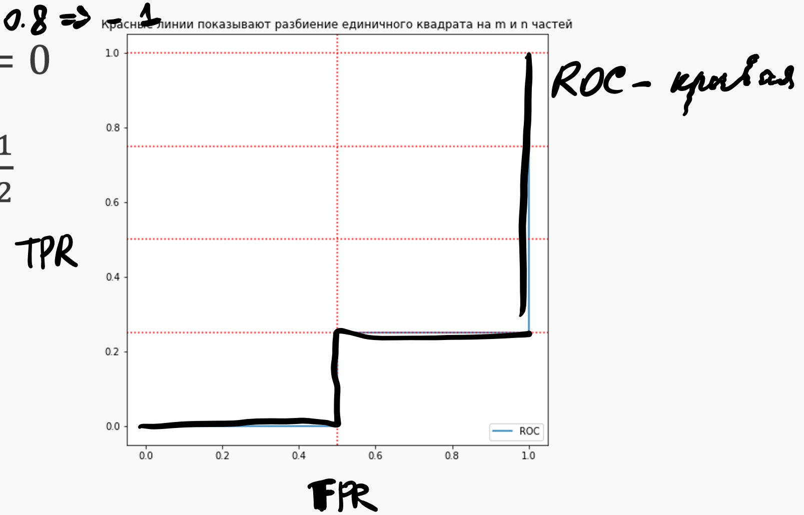
$$\boldsymbol{t = 0.8}: TPR = \frac{0}{4} = 0, FPR = \frac{1}{2}$$

$$\boldsymbol{t = 0.6}: TPR = \frac{1}{4}, FPR = \frac{1}{2}$$

$$\boldsymbol{t = 0.4}: TPR = \frac{1}{4}, FPR = \frac{2}{2}$$

$$\boldsymbol{t = 0.15}: TPR = \frac{2}{4}, FPR = \frac{2}{2}$$

…



Свойства

- Лежит в единичном квадрате.
- Для идеального классификатора проходит через $(0, 1)$.
- AUC ROC – площадь под ROC-кривой.
- Для идеального классификатора AUC ROC $= 1$.
- Для худшего классификатора AUC ROC ~ 0.5 (или 0).