Machine Learning

myTRACKER

@ mail.ru group

Деревья решений

(регрессия)

Количественные признаки

X	y
1	0
2	1
3	1
4	0
5	1
6	1
7	0
8	1
9	1

Количественные признаки

X	y
1	0
2	1
3	1
4	0
5	1
6	1
7	0
8	1
9	1

Количественные признаки

X	y
1	0
2	1
3	1
4	0
5	1
6	1
7	0
8	1
9	1

X	y
10	0
21	1
37	1
42	0
54	1
69	1
71	0
83	1
95	1

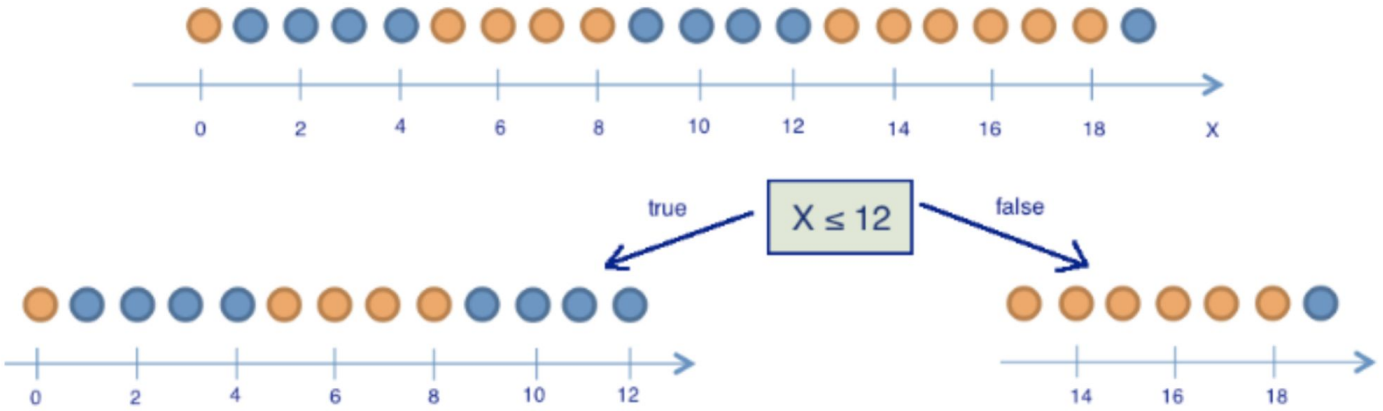
Количественные признаки

X	y
1	0
2	1
3	1
4	0
5	1
6	1
7	0
8	1
9	1

X	y	
10	0	
21	1	15.5
37	1	
42	0	39.5
54	1	48
69	1	
71	0	70
83	1	
95	1	

Деревья решений

Оранжевые: $p_1 = \frac{9}{20}$ Синие: $p_2 = \frac{11}{20}$ $S_0 = -\frac{9}{20} log_2\left(\frac{9}{20}\right) - \frac{11}{20} log_2\left(\frac{11}{20}\right) = 0.993$



$S_1 = -\frac{5}{13} log_2\left(\frac{5}{13}\right) - \frac{8}{13} log_2\left(\frac{8}{13}\right) = 0.96$ $S_2 = -\frac{1}{7} log_2\left(\frac{1}{7}\right) - \frac{6}{7} log_2\left(\frac{6}{7}\right) = 0.59$

$$\text{IG}(\text{“}X \leq 12\text{”}) = S_0 - \frac{13}{20} S_1 - \frac{7}{20} S_2 = 0.163$$

Алгоритм

Обходим все варианты и находим разбиение с наибольшим Information Gain (IG). После того повторяем операцию для каждого из разбиений, пока все объекты из разбиения не будут одного класса.

Деревья решений

ID	feature	target	
1	1	3	
2	2	6	
3	3	6	
4	5	6	
5	5	4	
6	7	3	
7	8	3	
8	9	3	
9	12	6	

Деревья решений

ID	feature	target	
1	1	3	
2	2	6	
3	3	6	
4	5	6	
5	5	4	
6	7	3	
7	8	3	
8	9	3	
9	12	6	

Деревья решений

ID	feature	target	
1	1	3	2.09
2	2	6	2.42
3	3	6	2.42
4	5	6	2.42
5	5	4	0.2
6	7	3	2.09
7	8	3	2.09
8	9	3	2.09
9	12	6	2.42

Выборочная дисперсия

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)^2$$

Деревья решений

ID	feature	target	
1	1	3	2.09
2	2	6	2.42
3	3	6	2.42
4	5	6	2.42
5	5	4	0.2
6	7	3	2.09
7	8	3	2.09
8	9	3	2.09
9	12	6	2.42

Выборочная дисперсия

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n}X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)^2$$

$$S^2_{top} = \frac{2.09+2.42+2.42+2.42+0.2}{5} = 1.91$$

$$S^2_{bottom} = \frac{2.09+2.09+2.09+2.42}{4} = 2.17$$

$$S^2_{all} = 2.03$$

$$IG_5 = S^2_{all} - \frac{5}{9}S^2_{top} - \frac{4}{9}S^2_{bottom} = 0.0044$$

Алгоритм

Обходим все варианты и находим разбиение с наибольшим Information Gain (IG). После того повторяем операцию для каждого из разбиений, пока все объекты из разбиения не будут одного класса.

Деревья решений

ID	feature	target	
1	1	3	2.09
2	2	6	2.42
3	3	6	2.42
4	5	6	2.42
5	5	4	0.2
6	7	3	2.09
7	8	3	2.09
8	9	3	2.09
9	12	6	2.42

Выборочная дисперсия

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n}X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)^2$$

$$IG_1 = 0.0022$$
$$IG_3 = 0.0$$
$$IG_5 = 0.0044$$
$$IG_8 = 0.0011$$

Алгоритм

Обходим все варианты и находим разбиение с наибольшим Information Gain (IG). После того повторяем операцию для каждого из разбиений, пока все объекты из разбиения не будут одного класса.

Метрики классификации

- Матрица ошибок (Confusion matrix)

	$y_{true} = 1$	$y_{true} = 0$
$y_{pred} = 1$	True Positive (TP)	False Positive (FP)
$y_{pred} = 0$	False Negative (FN)	True Negative (TN)

Метрики классификации

- Матрица ошибок (Confusion matrix)

	$y_{true} = 1$	$y_{true} = 0$
$y_{pred} = 1$	True Positive (TP)	Ошибка 2 рода (FP)
$y_{pred} = 0$	Ошибка 1 рода (FN)	True Negative (TN)

Метрики классификации

- Матрица ошибок (Confusion matrix)

	$y_{true} = 1$	$y_{true} = 0$
$y_{pred} = 1$	True Positive (TP)	False Positive (FP)
$y_{pred} = 0$	False Negative (FN)	True Negative (TN)

- Accuracy

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

Метрики классификации

- Матрица ошибок (Confusion matrix)

  Пусть 10 писем спам и 100 писем не спам

	$y_{true} = 1$	$y_{true} = 0$
$y_{pred} = 1$	5 (TP)	10 (FP)
$y_{pred} = 0$	5 (FN)	90 (TN)

- Accuracy

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} = \frac{5+90}{5+90+10+5} = 86.4$$

Метрики классификации

- Матрица ошибок (Confusion matrix)

  Пусть 10 писем спам и 100 писем не спам

	$y_{true} = 1$	$y_{true} = 0$
$y_{pred} = 1$	0 (TP)	0 (FP)
$y_{pred} = 0$	10 (FN)	100 (TN)

- Accuracy

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} = \frac{5+90}{5+90+10+5} = 86.4$$

Если предсказывать всегда не спам:

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} = \frac{0+100}{0+100+0+10} = 90.9$$

Метрики классификации

- Матрица ошибок (Confusion matrix)

	$y_{true} = 1$	$y_{true} = 0$
$y_{pred} = 1$	True Positive (TP)	False Positive (FP)
$y_{pred} = 0$	False Negative (FN)	True Negative (TN)

- Precision (точность)

$$precision = \frac{TP}{TP+FP}$$

- Recall (полнота)

$$recall = \frac{TP}{TP+FN}$$

- F-мера

$$f\_score = 2\frac{precision*recall}{precision+recall} = (\beta^2+1)\frac{precision*recall}{\beta^2 precision+recall}$$

Метрики классификации

- Матрица ошибок (Confusion matrix)

Пусть 10 писем спам и 100 писем не спам

	$y_{true} = 1$	$y_{true} = 0$
$y_{pred} = 1$	5 (TP)	10 (FP)
$y_{pred} = 0$	5 (FN)	90 (TN)

$$precision = \frac{TP}{TP+FP} = \frac{5}{5+10} = 0.33$$

$$recall = \frac{TP}{TP+FN} = \frac{5}{5+5} = 0.5$$

$$f\_score = 2\frac{precision*recall}{precision+recall} = 2\frac{0.33*0.5}{0.33+0.5} = 0.39$$

Метрики классификации

- Матрица ошибок (Confusion matrix)

  Пусть 10 писем спам и 100 писем не спам

	$y_{true} = 1$	$y_{true} = 0$
$y_{pred} = 1$	5 (TP)	10 (FP)
$y_{pred} = 0$	5 (FN)	90 (TN)

- True Positive Rate

$$TPR = \frac{TP}{TP+FN} = \frac{5}{5+5} = 0.5$$

- False Positive Rate

$$FPR = \frac{FP}{FP+TN} = \frac{10}{10+90} = 0.1$$

Метрики классификации

ID	Вероятность
1	0.99
2	0.9
3	0.51
4	0.49
5	0.44

- True Positive Rate

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

- False Positive Rate

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

Метрики классификации

ID	Вероятность
1	0.99
2	0.9
3	0.51
4	0.49
5	0.44

ID	Вероятность
1	0.99
2	0.9
3	0.51
4	0.49
5	0.44

ID	Вероятность
1	0.99
2	0.9
3	0.51
4	0.49
5	0.44

- True Positive Rate

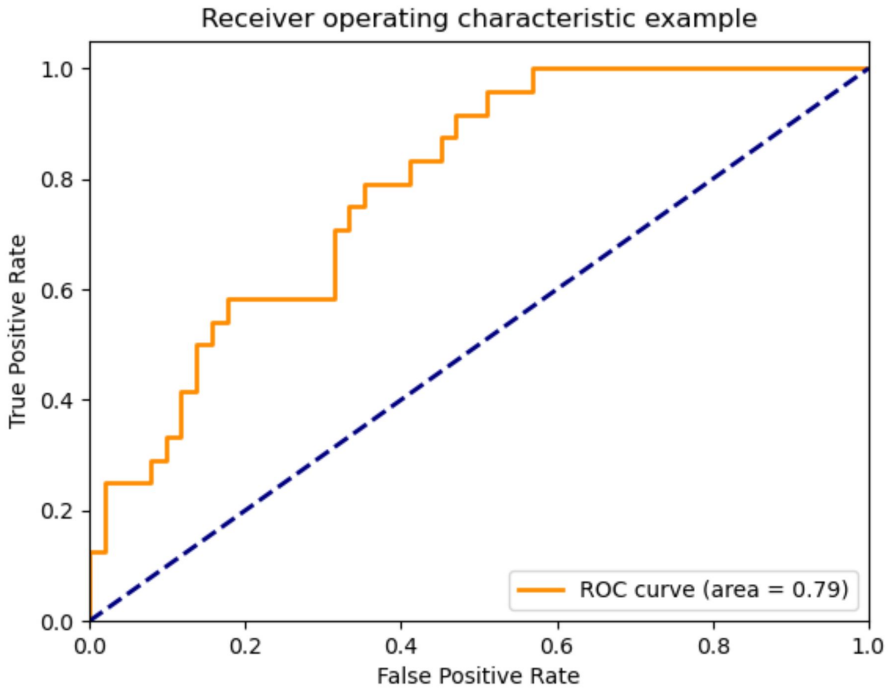
$$TPR = \frac{TP}{TP+FN}$$

- False Positive Rate

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

Метрики классификации

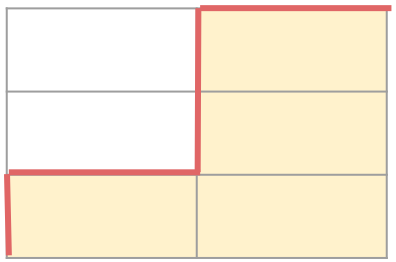
- Построение ROC AUC
  a. Для каждого порога посчитать TPR и FPR
  b. Построить график в осях TPR/FPR
- Кривая на графике: ROC curve
- Площадь под кривой: ROC AUC



Метрики классификации

ID	Вероятность	Ответ
1	0.99	1
2	0.9	0
3	0.51	1
4	0.49	1
5	0.44	0

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



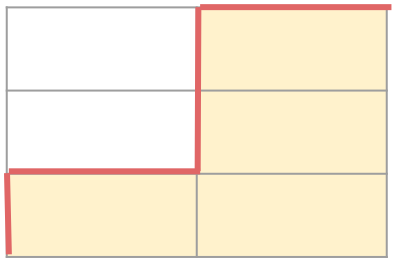
$$FPR = \frac{FP}{FP+TN}$$

$$\frac{4}{6}$$

Метрики классификации

ID	Вероятность	Ответ
1	0.99	1
2	0.9	0
3	0.51	1
4	0.49	1
5	0.44	0

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

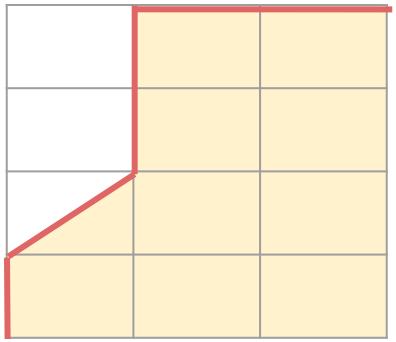


$$FPR = \frac{FP}{FP+TN}$$

$$\frac{4}{6}$$

ID	Вероятность	Ответ
1	0.99	1
2	0.9	0
3	0.9	1
4	0.51	1
5	0.49	1
6	0.44	0
7	0.3	0

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



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

$$\frac{9.5}{12}$$

Метрики классификации

Что делать если классов несколько?

- micro-метрики
  - Считаем общую матрицу ошибок
  - По ним вычисляем метрики
- macro-метрики
  - Вычисляем метрики на каждом классе “1 против всех”
  - Берем среднее
- weighted-метрики
  - Вычисляем метрики на каждом классе “1 против всех”
  - Берем средневзвешенное по классам
  - **Проблема:** F-мера может не лежать между Precision и Recall

Метрики классификации

Пусть у нас есть 4 класса:

- **Класс 1**. $TP = 1, FN = 1$
- **Класс 2**. $TP = 1, FN = 2$
- **Класс 3**. $TP = 5, FN = 1$
- **Класс 4**. $TP = 5, FN = 5$

$$recall = \frac{TP}{TP+FN}$$

$$recall_{micro} = \frac{1+1+5+5}{2+3+6+10} = 0.57$$

$$recall_{macro} = \frac{0.5+0.33+0.83+0.5}{4} = 0.54$$

$$recall_{weighted} = 0.095 * 0.5 + 0.143 * 0.33$$
$$+ 0.286 * 0.83 + 0.476 * 0.5 = 0.57$$

LogLoss

Определение. Вероятностью отнесения объекта к положительному классу будем рассчитывать по формуле:

$$p_+(x_i) = P(y_i = 1|x_i) = p_i$$

LogLoss

Определение. Вероятностью отнесения объекта к положительному классу будем рассчитывать по формуле:

$$p_+(x_i) = P(y_i = 1|x_i) = p_i$$

Определение. Вероятностью отнесения объекта к отрицательному классу будем рассчитывать по формуле:

$$p_-(x_i) = P(y_i = 0|x_i) = 1 - p_+(x_i) = 1 - p_i$$

Определение. Вероятностью отнесения объекта к своему классу будем рассчитывать по формуле:

$$p(x_i) = p_i^{y_i}(1 - p_i)^{1-y_i}$$

LogLoss

Определение. Вероятностью отнесения объекта к положительному классу будем рассчитывать по формуле:

$$p_+(x_i) = P(y_i = 1|x_i) = p_i$$

Определение. Вероятностью отнесения объекта к отрицательному классу будем рассчитывать по формуле:

$$p_-(x_i) = P(y_i = 0|x_i) = 1 - p_+(x_i) = 1 - p_i$$

Определение. Вероятностью отнесения объекта к своему классу будем рассчитывать по формуле:

$$p(x_i) = p_i^{y_i}(1 - p_i)^{1-y_i}$$

$$LogLoss(x_i) = -log(\prod_i p(x_i))$$

Метрики регрессии

- $MAE = \frac{\sum_i |y_i - \bar{y}_i|}{n}$
- $MSE = \frac{\sum_i (y_i - \bar{y}_i)^2}{n}$
- $RMSE = \sqrt{\frac{\sum_i (y_i - \bar{y}_i)^2}{n}} = \sqrt{MSE}$

Метрики регрессии

- $MAE = \frac{\sum_i |y_i - \bar{y}_i|}{n}$
- $MSE = \frac{\sum_i (y_i - \bar{y}_i)^2}{n}$
- $RMSE = \sqrt{\frac{\sum_i (y_i - \bar{y}_i)^2}{n}}$

$$MAE \leq RMSE \leq MAE\sqrt{n}$$

Метрики регрессии

- $MAE = \frac{\sum_i |y_i - \bar{y}_i|}{n}$
- $MSE = \frac{\sum_i (y_i - \bar{y}_i)^2}{n}$ $\quad MAE \leq RMSE \leq MAE\sqrt{n}$
- $RMSE = \sqrt{\frac{\sum_i (y_i - \bar{y}_i)^2}{n}}$
- $MAPE = \sum_i \left| \frac{y_i - \bar{y}_i}{y_i} \right|$
- $SMAPE = \frac{2}{n} \sum_i \frac{|y_i - \bar{y}_i|}{(y_i + \bar{y}_i)}$

Метрики регрессии

- $MAE = \frac{\sum_i |y_i - \bar{y}_i|}{n}$
- $MSE = \frac{\sum_i (y_i - \bar{y}_i)^2}{n}$
- $RMSE = \sqrt{\frac{\sum_i (y_i - \bar{y}_i)^2}{n}}$
- $MAPE = \sum_i \left| \frac{y_i - \bar{y}_i}{y_i} \right|$
- $SMAPE = \frac{2}{n} \sum_i \frac{|y_i - \bar{y}_i|}{(y_i + \bar{y}_i)}$

$$MAE \leq RMSE \leq MAE\sqrt{n}$$

A = 100 and F = 120. SMAPE = 18.2%
A = 100 and F = 80. SMAPE = 22.2%