EleWise проекты

Лекция 1

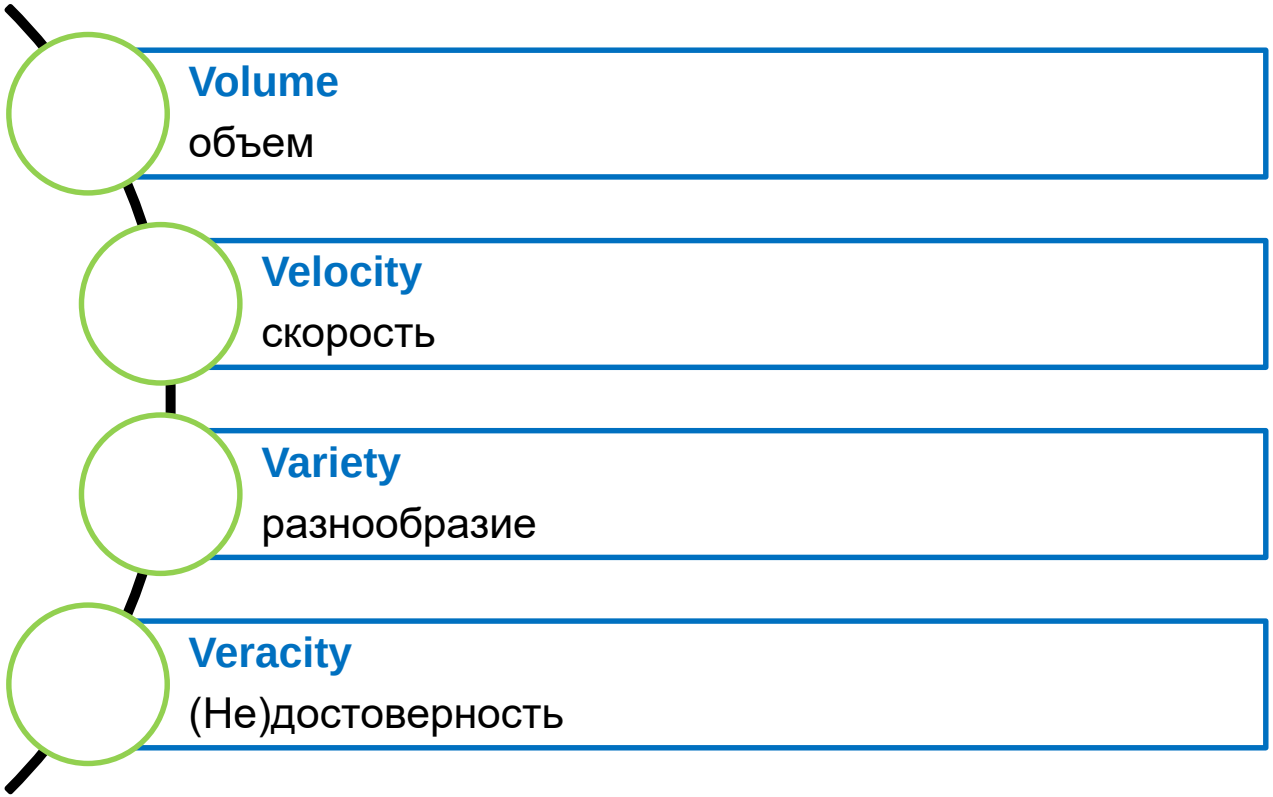
Знакомство с анализом данных и машинным обучением

Мир меняется


2005 – Папа Римский Бенедикт


2013 – Папа Римский Франциск

Big Data: 4V


Volume
объем
Velocity
скорость
Variety
разнообразие
Veracity
(Не)достоверность

Машинное обучение (machine learning = ML)

систематическое обучение алгоритмов и систем, в результате которого их знания или качество работы возрастают по мере накопления опыта

AI & Data Science

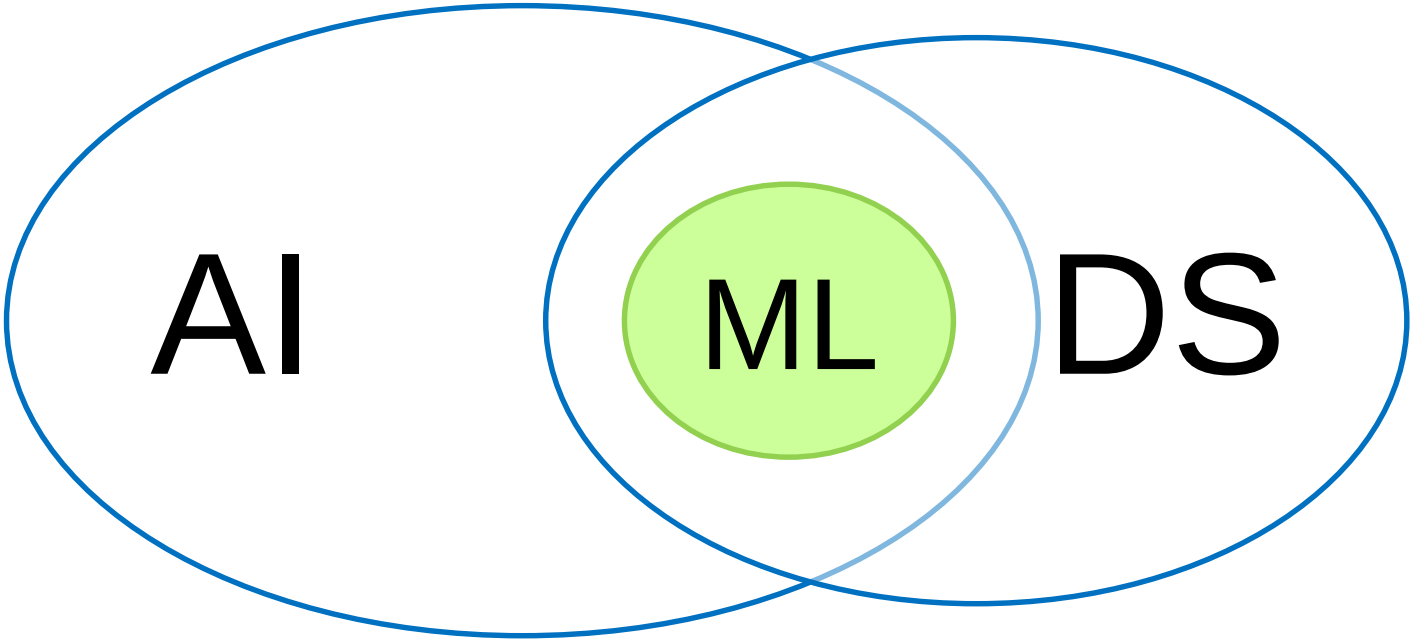
Искусственный интеллект

наука и технология создания интеллектуальных машин

Наука о данных

раздел информатики, изучающий проблемы анализа, обработки и представления данных в цифровой форме

AI & Data Science



1.1. Формальная постановка задачи машинного обучения

1.1. Формальная постановка задачи машинного обучения

X – множество объектов

Y – множество ответов

$y : X \to Y$ – неизвестная зависимость

Дано:

$\{x_1, \ldots, x_\ell\} \subset X$ – обучающая выборка

$y_i = y(x_i)$, $i = 1, \ldots, \ell$ - известные ответы

Найти:

$a : X \to Y$ – решающую функцию (алгоритм), приближающую y

1.1. Формальная постановка задачи машинного обучения

X – множество объектов Y – множество ответов Y' – множество ответов

$$X = \begin{pmatrix} x_1^{(1)} & x_2^{(1)} & \dots & x_k^{(1)} \\ x_1^{(2)} & \ddots & & \vdots \\ \vdots & & & \\ x_1^{(n)} & \dots & & x_k^{(n)} \end{pmatrix} \quad Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \Rightarrow \quad Y' = \begin{pmatrix} y_1' \\ y_2' \\ \vdots \\ y_n' \end{pmatrix},$$

Найти:

$a : X \to Y$ – решающую функцию (алгоритм), приближающую y

1.1. Формальная постановка задачи машинного обучения

$f_j : X \to D_j \;\; j = 1,\ldots, n$ – признаки объектов

Типы признаков:

- $D_j = \{0, 1\}$ – бинарный признак f_j;
- $|D_j| < \infty$ - номинальный признак f_j;
- $|D_j| < \infty$, Dj упорядочено – порядковый признак f_j;
- $D_j = R$ – количественный признак f_j.

Вектор $(f_1(x),\ldots, f_n(x))$ – признаковое описание объекта X

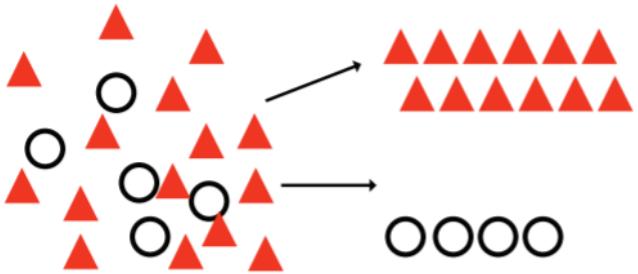
Матрица «объекты-признаки»

$$F = \|f_j(x_i)\|_{\ell \times n} = \begin{pmatrix} f_1(x_1) & \ldots & f_n(x_1) \\ \ldots & \ldots & \ldots \\ f_1(x_l) & \ldots & f_n(x_l) \end{pmatrix}$$

1.1. Формальная постановка задачи машинного обучения

Задачи классификации

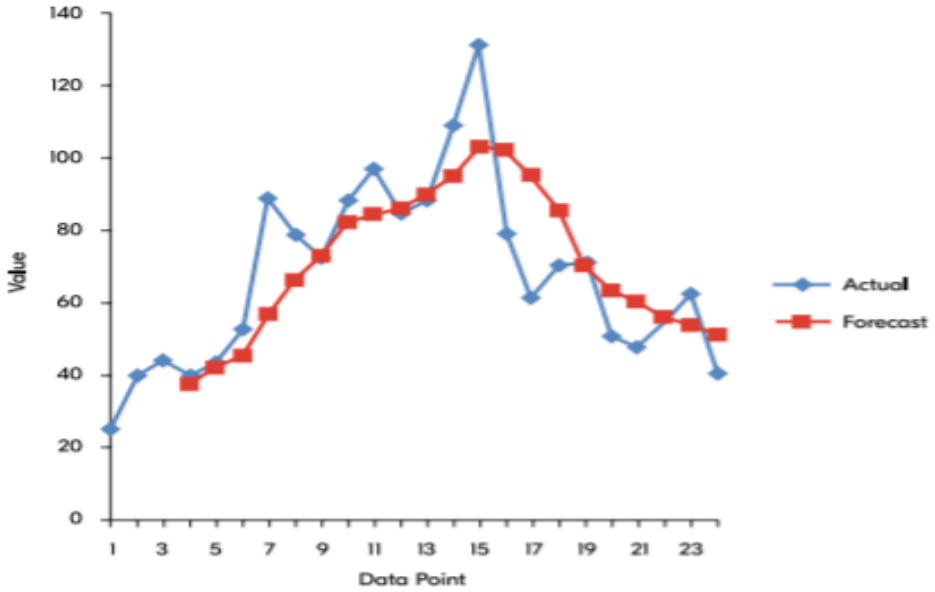
- $Y = \{-1, +1\}$ - классификация на два класса
- $Y = \{1, \ldots, M\}$ - классификация на M непересекающихся классов
- $Y = \{0, 1\}^M$ - классификация на M классов, которые могут пересекаться



1.1. Формальная постановка задачи машинного обучения

Задачи восстановления регрессии

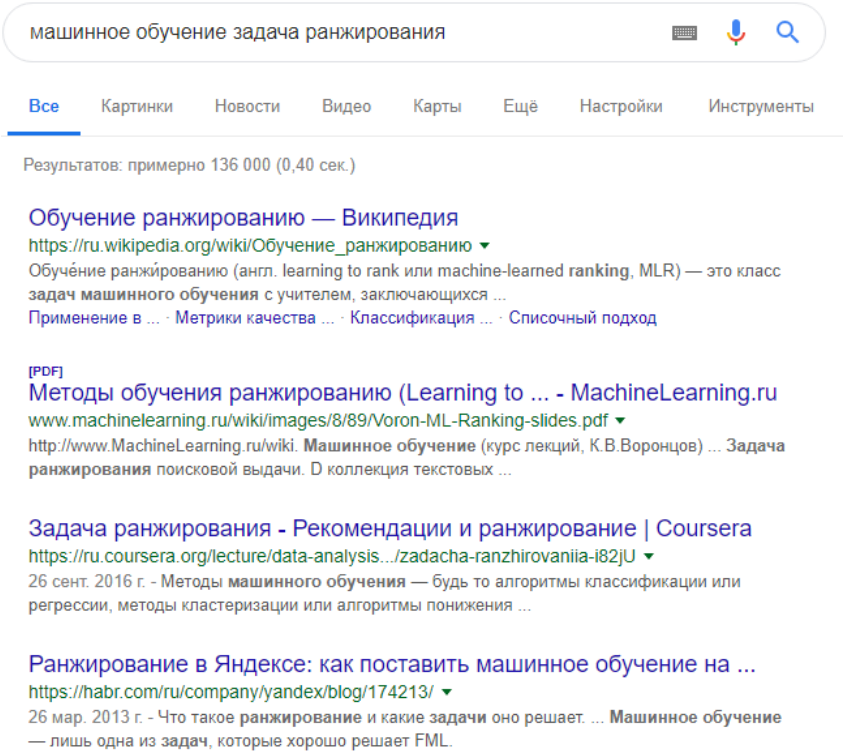
- $Y = R$ или $Y = R^m$



1.1. Формальная постановка задачи машинного обучения

Задачи ранжирования

Y - конечное упорядоченное множество



1.1. Формальная постановка задачи машинного обучения

Модель - параметрическое семейство функций

$$A = \{a(x) = g(x, \theta)\ )\ )\ \theta \in \Theta\},$$

где $g: X \times \Theta \to Y$ - фиксированная функция

Θ - множество допустимых параметров θ

Пример: линейная модель с вектором параметров $\theta = (\theta_1, \ldots, \theta_n)$ $\Theta = R^n$

$g(x, \theta) = \sum^n{}_{j=1}\ \theta_j f_j(x)$ - регрессия и ранжирование $Y = R$

$g(x, \theta) = sign \sum^n{}_{j=1}\ \theta_j f_j(x)$ - классификация $Y = \{-1, +1\}$

1.1. Формальная постановка задачи машинного обучения

$\mathcal{L}(a, x)$ - функция потерь, величина ошибки алгоритма $a \in A$ на объекте $x \in X$

Ф-ия потерь для задачи классификации

- $\mathcal{L}(a, x) = [a(x) \neq y(x)]$ - индикатор ошибки

Ф-ия потерь для задачи регрессии

- $\mathcal{L}(a, x) = |a(x) - y(x)|$ - абсолютное значение ошибки
- $\mathcal{L}(a, x) = (a(x) - y(x))^2$ - квадратичная ошибка

1.1. Формальная постановка задачи машинного обучения

Эмпирический риск - функционал качества алгоритма на x^ℓ

$$Q(a,X^\ell) = 1/\ell \sum_{i=1}^{\ell} \mathcal{L}(a, x_i)$$

Минимизация эмпирического риска

$$\mu(X^\ell) = \arg\min_{a \in A} Q(a,X^\ell)$$

Пример: метод наименьших квадратов ($Y = R$, $\mathcal{L}$ - квадратична)

$$\mu(X^\ell) = \arg\min_{\theta} \sum^{\ell}_{i=1} (g(x_i, \theta) - y_i)^2$$

1.2. Примеры применения

1.2. Примеры применения

- Задачи медицинской диагностики
- Задачи кредитного скоринга
- Задачи предсказания оттока клиентов
- Зачади рубрикации текстовых документов
- Задачи прогнозирования отказа оборудования

1.2. Примеры применения

EleWise проекты

1.3. Машинное обучение в прикладных задачах

1.3. Машинное обучение в прикладных задачах
1.3.1. Этапы анализа данных

CRISP-DM

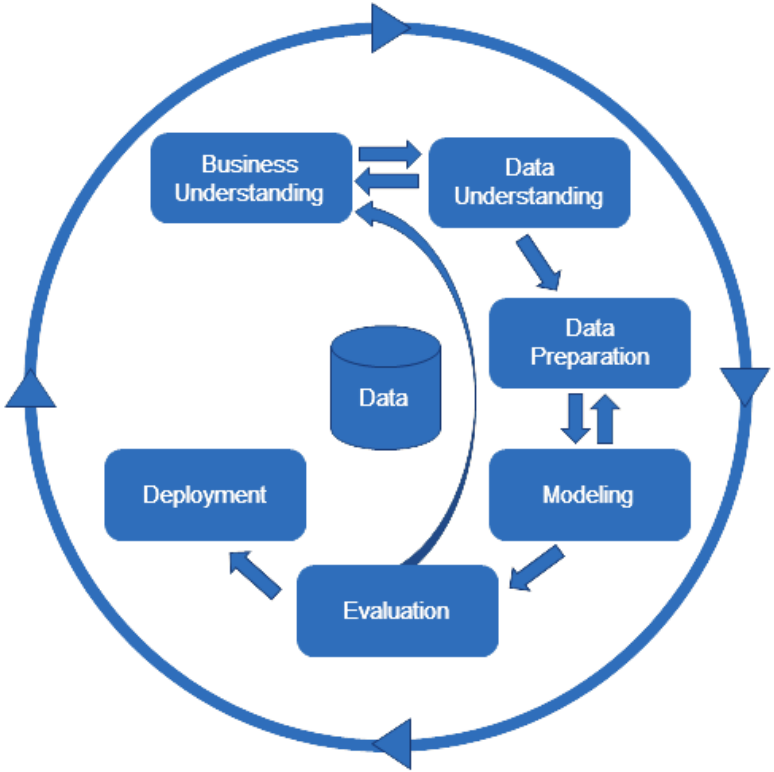
Cross-Industry Standard Process for Data Mining
межотраслевой стандартный процесс для исследования данных

CRISP-DM

- Понимание бизнес-целей (*Business Understanding*)
- Начальное изучение данных (*Data Understanding*)
- Подготовка данных (*Data Preparation*)
- Моделирование (*Modeling*)
- Оценка (*Evaluation*)
- Внедрение (*Deployment*)

1.3. Машинное обучение в прикладных задачах

1.3.1. Этапы анализа данных



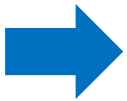
1.3. Машинное обучение в прикладных задачах

1.3.2. Препроцессинг

1. Создание векторного пространства признаков, где будут жить примеры обучающей выборки.
2. Нормализация данных. Вот самый классический пример нормализации данных: $X = (X — \mu)/\sigma$
3. Изменение размерности векторного пространства.

Пример:

color
серый
синий
зеленый



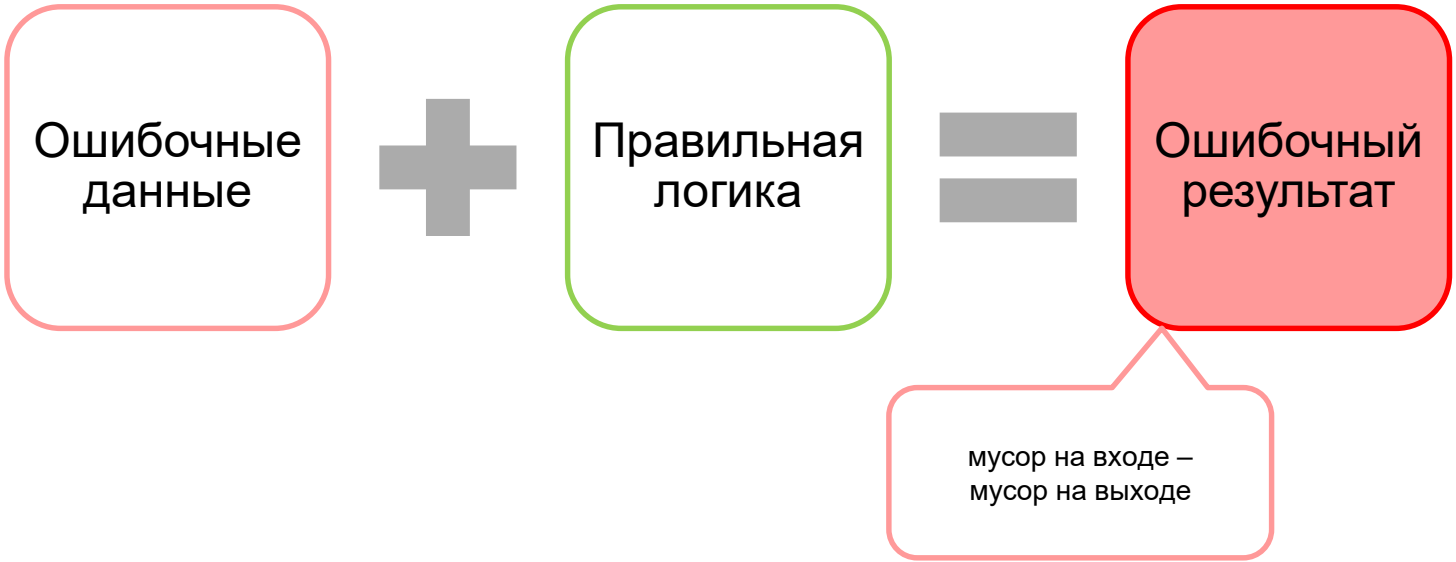
	color_ серый	color_ синий	color_ зеленый
серый	1	0	0
синий	0	1	0
зеленый	0	0	1

1.3. Машинное обучение в прикладных задачах
1.3.2. Препроцессинг

- количество
    - объем выборки
    - внутренние + внешние
- качество
    - фальсификация данных
    - пропущенные значения

1.3. Машинное обучение в прикладных задачах
1.3.2. Препроцессинг

Качество данных: проблема



1.3. Машинное обучение в прикладных задачах
1.3.2. Препроцессинг

Качество данных: проблема

Поле	Значение	Ошибка
Имя	Сергей	Первая буква - латинская
Фамилия	Петрович	Значение из другого поля
Город	Мсква	Опечатка
Доходы	100 руб.	Подозрительная сумма
Телефон	000-00-01	Несуществующий номер

1.3. Машинное обучение в прикладных задачах
1.3.2. Препроцессинг

Качество данных: проблема

1.3. Машинное обучение в прикладных задачах

1.3.3. Оценка качества (регрессия)

$$MAPE = \frac{1}{N}\sum_{t=1}^{N}\frac{|Z(t)-\hat{Z}(t)|}{Z(t)}\cdot 100\,\%$$ – средняя абсолютная ошибка в процентах

$$MAE = \frac{1}{N}\sum_{t=1}^{N}|Z(t)-\hat{Z}(t)|$$ – средняя абсолютная ошибка

$$MSE = \frac{1}{N}\sum_{t=1}^{N}\left(Z(t)-\hat{Z}(t)\right)^2$$ – среднеквадратичная ошибка

$$RMSE = \sqrt{MSE}$$ – кв. корень из среднеквадр. ошибки

$$ME = \frac{1}{N}\sum_{t=1}^{N}(Z(t)-\hat{Z}(t))$$ – средняя ошибка

$$SD = \sqrt{\frac{1}{N}\cdot\sum_{t=1}^{N}(\hat{Z}(t)-ME)^2}$$ – стандартное отклонение

1.3. Машинное обучение в прикладных задачах
1.3.3. Оценка качества (классификация)

Матрица сопряженности (Confusion matrix)

	Фактически	
Модель	+	-
+	TP	FP
-	FN	TN

1.3. Машинное обучение в прикладных задачах
1.3.3. Оценка качества (классификация)

Матрица сопряженности (Confusion matrix)

True Positives — верно классифицированные положительные объекты (истинно положительные случаи);

True Negatives — верно классифицированные отрицательные объекты (истинно отрицательные случаи);

False Negatives — положительные объекты, классифицированные как отрицательные (ошибка I рода, ложно отрицательные случаи);

False Positives — отрицательные объекты, классифицированные как положительные (ошибка II рода, ложно положительные случаи).

1.3. Машинное обучение в прикладных задачах
1.3.3. Оценка качества (классификация)

Accuracy

- **Точность (Accuarcy)** — число верно классифицированных объектов по модели:

$$Ac = \frac{TP + TN}{n}$$

Sensitivity & Specificity

- **Чувствительность (Sensitivity)** — доля истинно положительных объектов:

$$Se = TPR = \frac{TP}{TP + FN} \cdot 100\%$$

- **Специфичность (Specificity)** — доля истинно отрицательных объектов, которые были правильно идентифицированы моделью:

$$Sp = \frac{TN}{TN + FP} \cdot 100\%$$

Precision & Recall

- **Точность (Precision)** — доля положительных объектов среди тех, кого назвали положительными:

$$\text{Precision} = \frac{TP}{TP + FP} \cdot 100\%$$

- **Полнота (Recall)** — доля истинно положительных объектов:

$$\text{Recall} = \frac{TP}{TP + FN} \cdot 100\%$$

1.3. Машинное обучение в прикладных задачах
1.3.3. Оценка качества (классификация)

Внимательно:

- Accuarcy – Точность
- Precision – Точность

F-score

- **F-мера (F-score) –** среднее гармоническое precision и recall

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

$$F1 = (1 + \beta^2) \cdot \frac{Precision \cdot Recall}{\beta^2 \cdot Precision + Recall}$$

Уровень отсечения (cut-off)

Y	P	Y`
1	0.2	
1	0.4	
0	0.1	
0	0.3	
1	0.5	

EleWise проекты

Уровень отсечения 0.1

Y	P	Y`
1	0.2	1
1	0.4	1
0	0.1	0
0	0.3	1
1	0.5	1

EleWise проекты

Уровень отсечения 0.2

Y	P	Y`
1	0.2	0
1	0.4	1
0	0.1	0
0	0.3	1
1	0.5	1

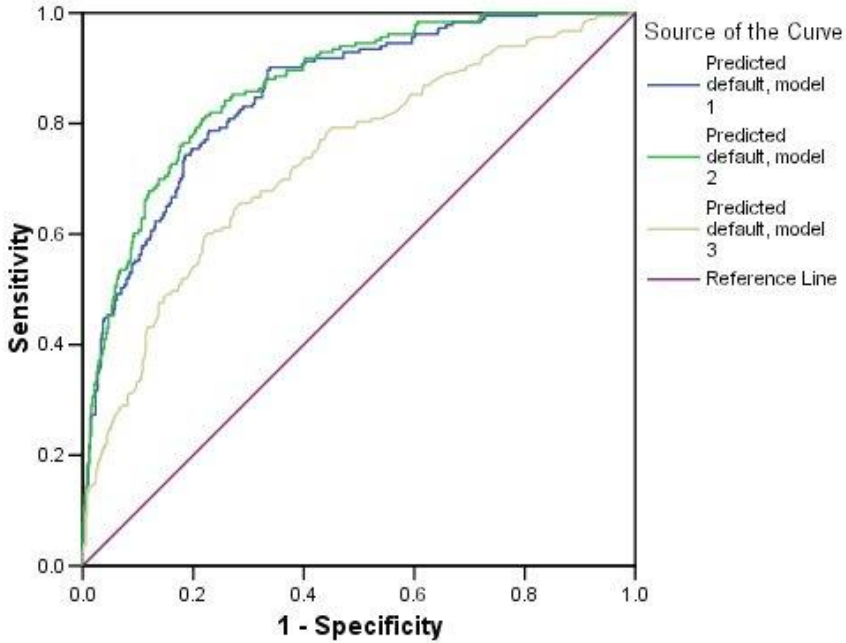
1.3. Машинное обучение в прикладных задачах
1.3.3. Оценка качества (классификация)

Уровень отсечения 0.9

Y	P	Y`
1	0.2	0
1	0.4	0
0	0.1	0
0	0.3	0
1	0.5	0

1.3. Машинное обучение в прикладных задачах
1.3.3. Оценка качества (классификация)

ROC-кривая



- ROC = receiver operating characteristic, иногда говорят «кривая ошибок»
- качество = площадь под этой кривой – **AUC** (AUC = area under the curve)

1.3. Машинное обучение в прикладных задачах

1.3.3. Оценка качества (классификация)

ROC-кривая

id	оценка	класс
1	0.5	0
2	0.1	0
3	0.2	0
4	0.6	1
5	0.2	1
6	0.3	1
7	0.0	0

Табл. 1

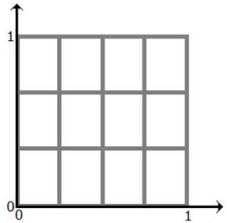
id	оценка	класс
4	0.6	1
1	0.5	0
6	0.3	1
3	0.2	0
5	0.2	1
2	0.1	0
7	0.0	0

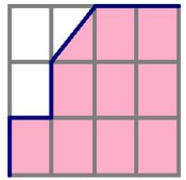
Табл. 2

id	> 0.25	класс
4	1	1
1	1	0
6	1	1
3	0	0
5	0	1
2	0	0
7	0	0

Табл. 3

1.3. Машинное обучение в прикладных задачах
1.3.3. Оценка качества (классификация)

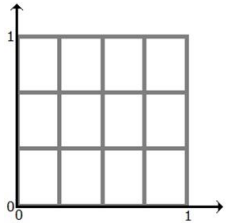


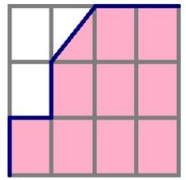


ROC-кривая

- Взять единичный квадрат на координатной плоскости, разбить его на m равных частей горизонтальными линиями и на ***n*** – вертикальными, где ***m*** – число 1 среди правильных меток теста (***m***=3), ***n*** – число нулей (***n***=4). В результате квадрат разбивается сеткой на ***m*×*n*** блоков.
- Просматривать строки табл. 2 сверху вниз и прорисовывать на сетке линии, переходя их одного узла в другой.
- Старт из точки (0, 0). Если значение метки класса в просматриваемой строке 1, то делаем шаг вверх; если 0, то делаем шаг вправо.
- В итоге мы попадём в точку (1, 1), т.к. сделаем в сумме ***m*** шагов вверх и ***n*** шагов вправо.

1.3. Машинное обучение в прикладных задачах
1.3.3. Оценка качества (классификация)

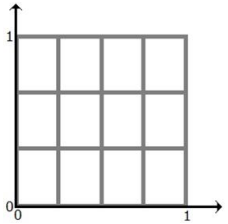


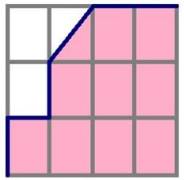


ROC-кривая

- Если у нескольких объектов значения оценок равны, то делаем шаг в точку, которая на ***a*** блоков выше и ***b*** блоков правее, где ***a*** – число единиц в группе объектов с одним значением метки, ***b***– число нулей в ней. В частности, если все объекты имеют одинаковую метку, то мы сразу шагаем из точки (0, 0) в точку (1, 1).

1.3. Машинное обучение в прикладных задачах
1.3.3. Оценка качества (классификация)





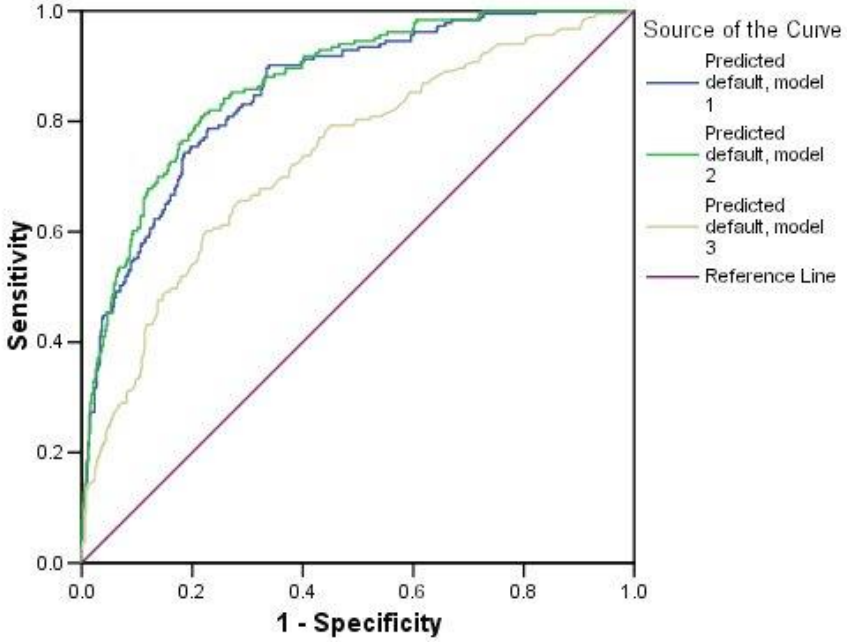
ROC-кривая

- **AUC ROC** – площадь под ROC-кривой – используют для оценивания качества упорядочивания алгоритмом объектов двух классов. Значение лежит на отрезке [0, 1].

- **AUC ROC равен доле пар объектов вида (объект класса 1, объект класса 0), которые алгоритм верно упорядочил**, т.е. первый объект идёт в упорядоченном списке раньше.

1.3. Машинное обучение в прикладных задачах
1.3.3. Оценка качества (классификация)

ROC-кривая



1.3. Машинное обучение в прикладных задачах

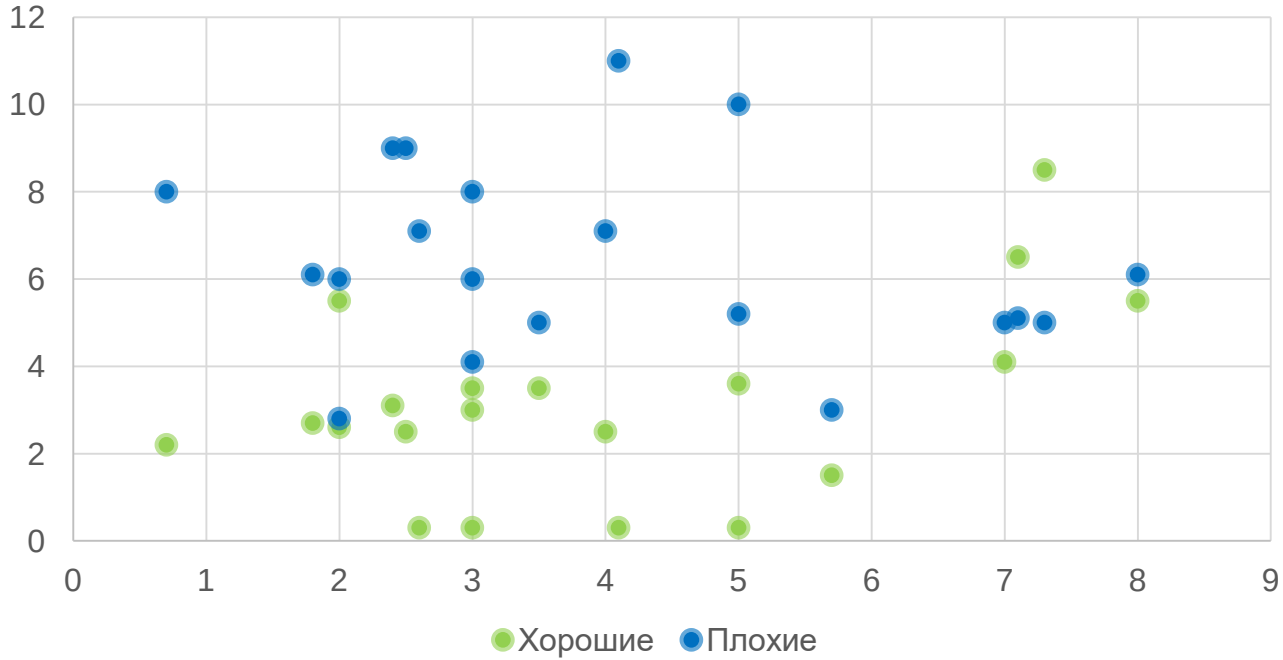
1.3.3. Оценка качества (классификация)

Интервал AUC	Качество модели
0.9-1.0	Отличное
0.8-0.9	Очень хорошее
0.7-0.8	Хорошее
0.6-0.7	Среднее
0.5-0.6	Неудовлетворительное

$$AUC = \int f(x)dx = \sum_i [\frac{X_{i+1} + X_i}{2}] \cdot (Y_{i+1} - Y_i)$$

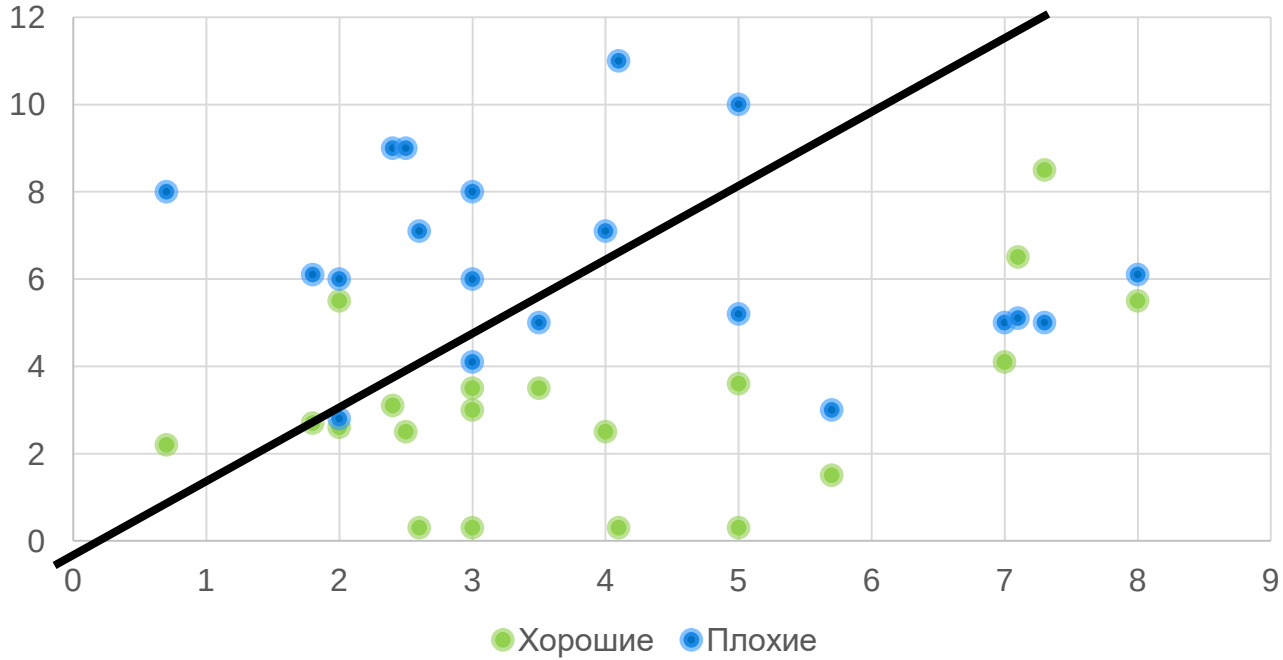
$$Gini = (AUC - 0.5) \cdot 2$$

1.3. Машинное обучение в прикладных задачах
1.3.4. Обзор алгоритмов

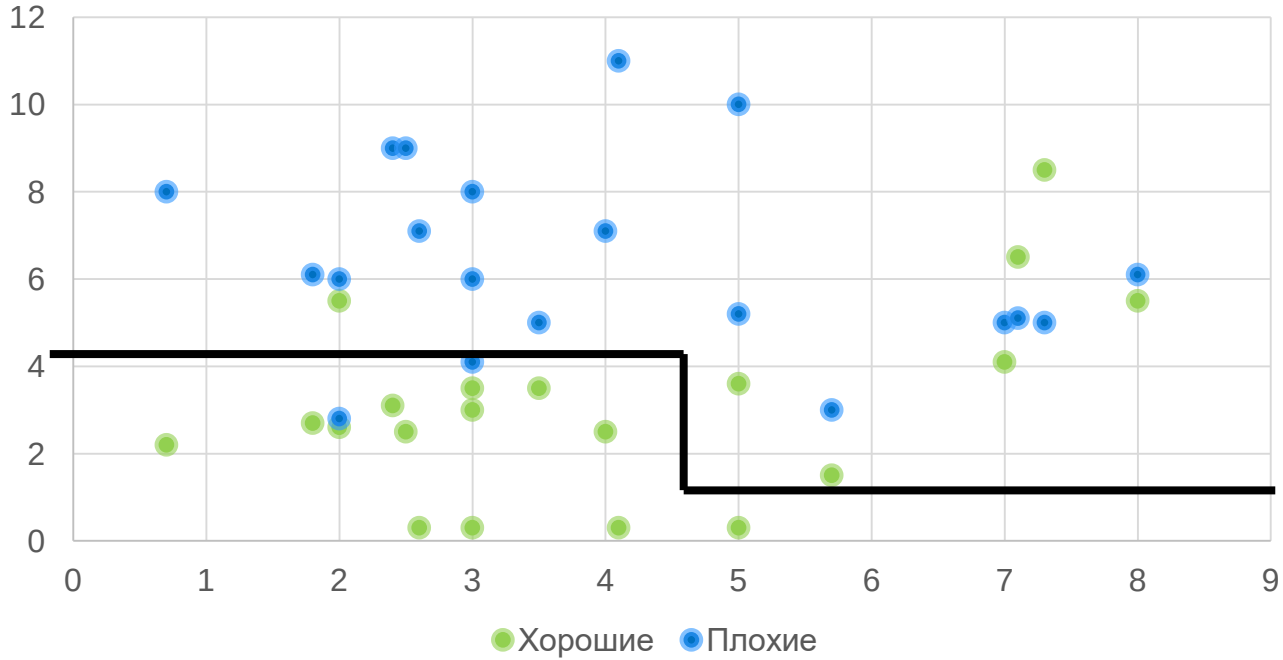


1.3. Машинное обучение в прикладных задачах

1.3.4. Обзор алгоритмов



1.3. Машинное обучение в прикладных задачах
1.3.4. Обзор алгоритмов



1.3. Машинное обучение в прикладных задачах

EleWise проекты

Китайская «система социального рейтинга»

К **2020** будет определять ценность людей:

- **SCS**: будет определен рейтинг жителей
- система будет определять позицию гражданина, отслеживая его социальное поведение: как он тратит деньги, регулярно ли оплачивает счета, даже то, как он взаимодействует с другими людьми
- от рейтинга гражданина будет зависеть, сможет ли он получить работу или ипотеку, а также в какой школе смогут учиться его дети


BLACK MIRROR