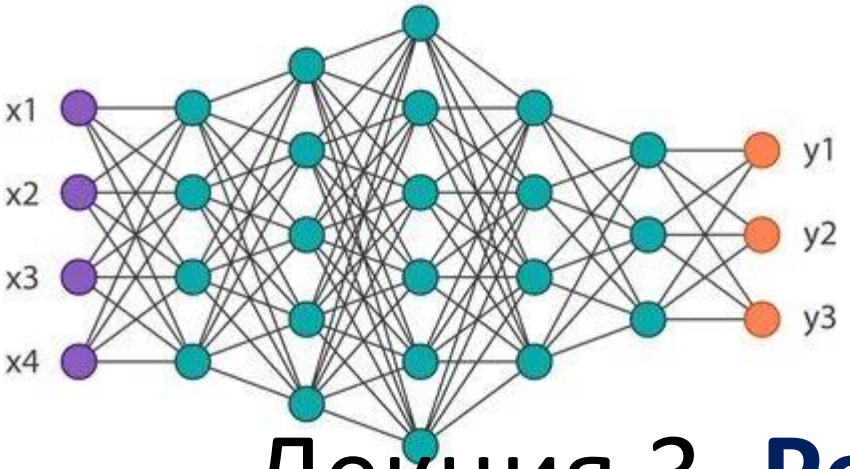
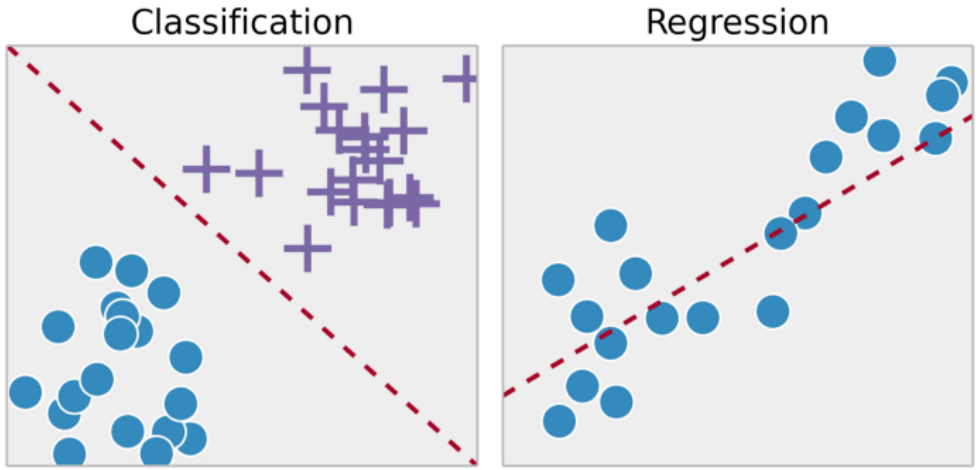


Лекция 3. **Решение задач регрессии и классификации с помощью искусственных нейронных сетей**

Дисциплина: **Интеллектуальный анализ данных, текстов и изображений**
Лектор: к.т.н. **Буров Сергей Александрович**
burov-sa@ranepa.ru

Введение


МАШИННОЕ ОБУЧЕНИЕ
Данные заранее категоризированы
или численные
Данные никак не размечены
С учителем
Без учителя
Предсказать категорию
Предсказать значение
Разделить по схожести
Найти зависимости
Выявить последовательности
Классификация
«Разложи носки по цвету»
Регрессия
«Разложи галстуки по длине»
Кластеризация
«Разложи похожие вещи по кучкам»
Ассоциация
«Найди какие вещи я часто ношу вместе»
Уменьшение Размерности (обобщение)
«Собери из вещей лучшие наряды»


Classification
Regression

Введение

В целом, классификация и регрессия достаточно схожие задачи (регрессию можно представить как классификацию, а классификацию, как регрессию).

В зависимости от области применения, обозначают и другие задачи, которые могут решаться с помощью классификации и регрессии.

Вопросы лекции:

1. Задача регрессии. Метрики качества для задачи регрессии
2. Задача классификации. Метрики качества для задачи классификации

Регрессионный анализ

Регрессионный анализ — метод моделирования измеряемых данных и исследования их свойств. Данные состоят из пар значений **зависимой переменной** (переменной отклика) и **независимой переменной** (объясняющей переменной).

Регрессионная модель — функция независимой переменной и параметров с добавленной случайной переменной. Параметры модели настраиваются таким образом, что модель наилучшим образом приближает данные.

Наборы данных для машинного обучения

Dataset (датасет, набор данных)– это обработанная и структурированная информация в табличном виде. Строки такой таблицы называются объектами, а столбцы – признакам

Предикторы

Целевые признаки

X1	X2	.....	Xn	Y1	Y2	...	Yn

Наборы данных для машинного обучения

Датасет для регрессионного анализа

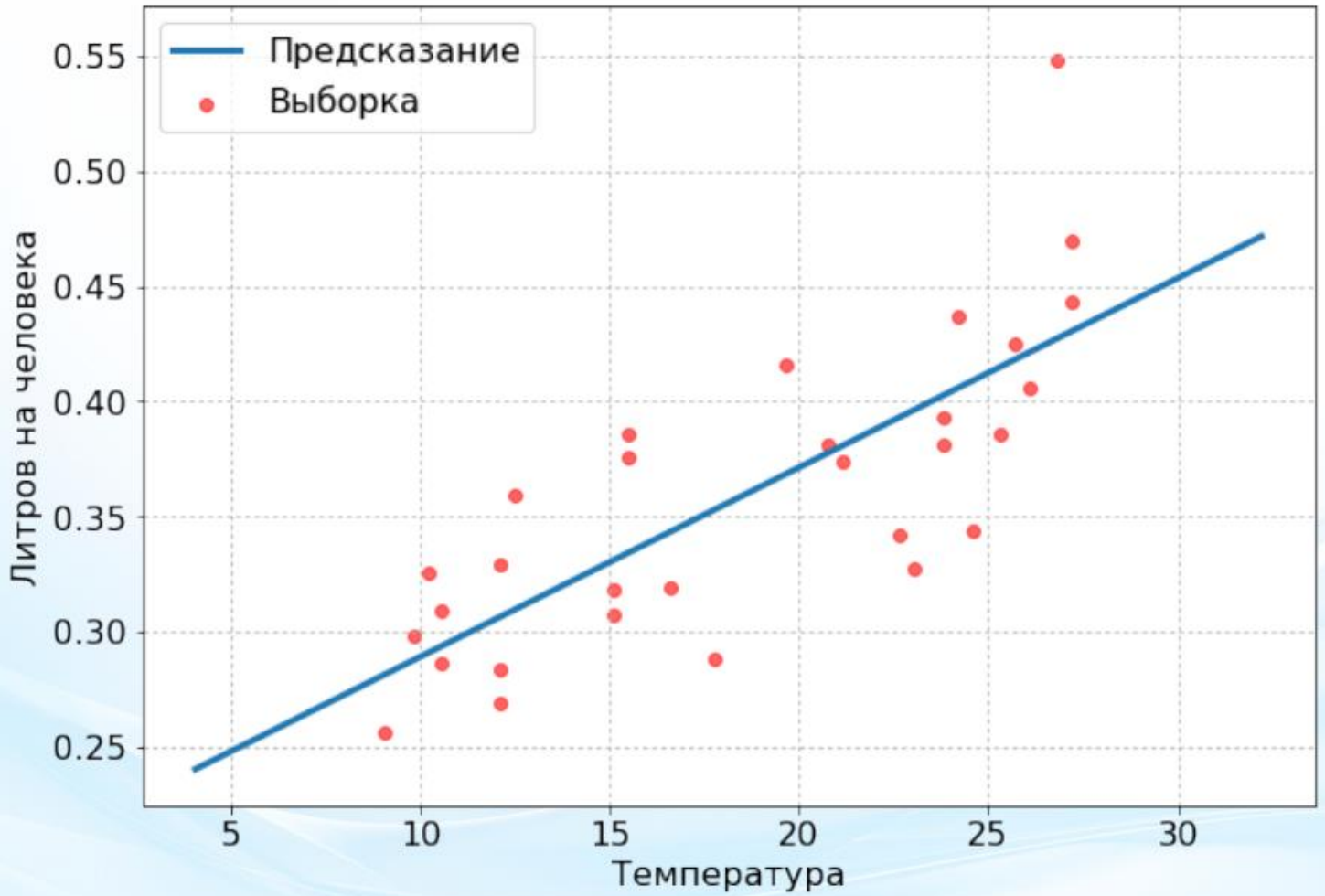
X1	X2	.....	Xn	Y1
12	Красный		1.7	1.6
14	Синий		6.4	0.8
2	Зелёный		12.3	12.3
14	Синий		12.4	1.54
12	Синий		1.2	1.45

Наборы данных для машинного обучения

Датасет для множественной регресси

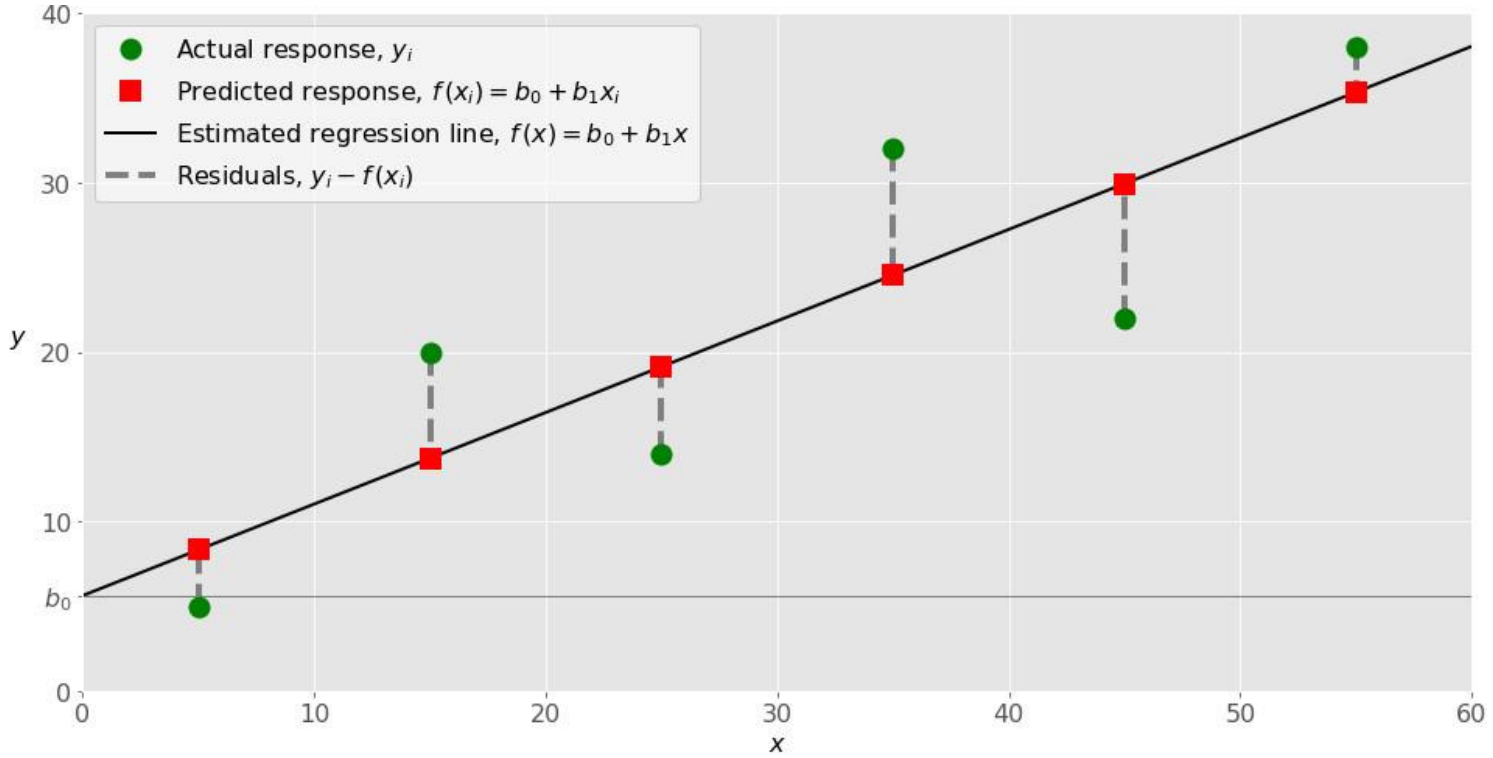
X1	X2	.....	Xn	Y1	Y2
12	Красный		1.7	1.6	2.4
14	Синий		6.4	0.8	1.2
2	Зелёный		12.3	12.3	1.5
14	Синий		12.4	1.54	1.4
12	Синий		1.2	1.45	1.23

Пример – потребление мороженного



Город	Температура	Литров на человека
Москва	8	0.26
Санкт-Петербург	10	0.30
Москва	11	0.35
Самара	12	0.31
Тамбов	11	0.37
……	……	……
Анапа	27	0.55

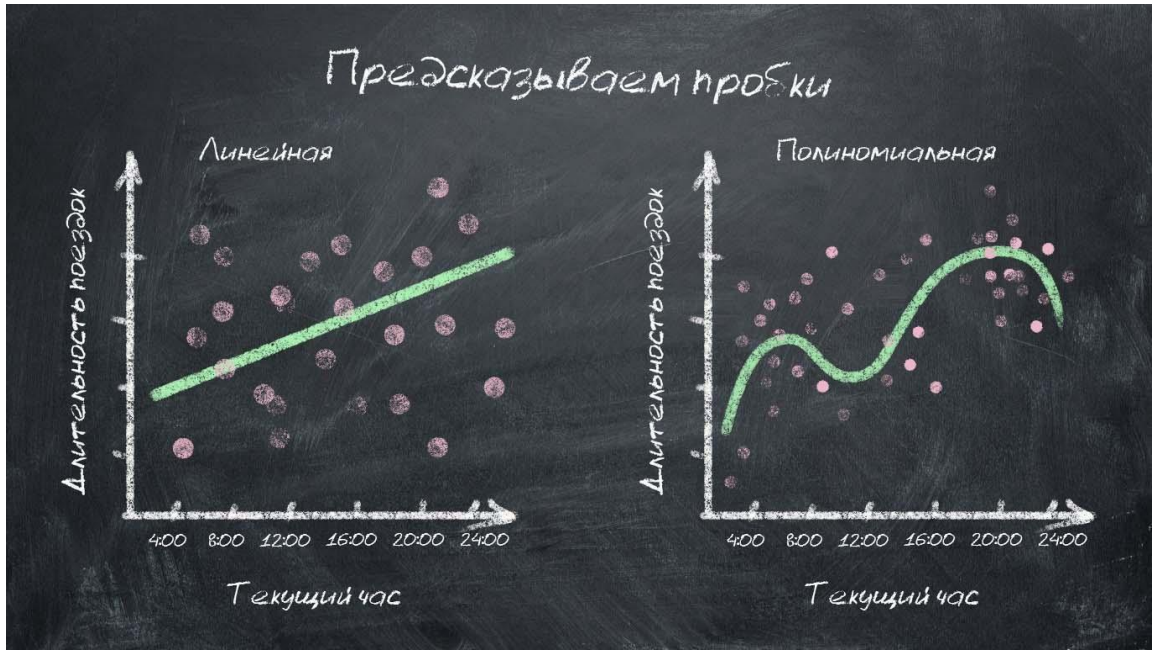
Линейная регрессия методом наименьших квадратов




```
m = sklearn.linear_model.LinearRegression(fit_intercept=True)
#Обучение модели:
m.fit(X, Y)
#Вектор коэффициентов:
m.coef_
#Свободный коэффициент:
m.intercept_
#Предсказания:
m.predict(X)
```

Функциональные зависимости при регрессионном анализе

Функции	Описание
Линейная	$y = a + bx$
Парабола второго порядка	$y = a + bx + cx^2$
Кубическая парабола	$y = a + bx + cx^2 + dx^3$
Показательная	$y = ab^x$
Экспоненциальная	$y = ae^{bx}$
Логарифмическая	$y = a + b \lg x$
Гиперболическая	$y = a + b\frac{1}{x}$
Кривая Гомперца	$y = ab^{c^x}$
Логистическая	$y = \frac{d}{1 + e^{a+bx}}$



Регрессия с помощью искусственных нейронных сетей – иной подход

Регрессия с помощью искусственных нейронных сетей – иной подход

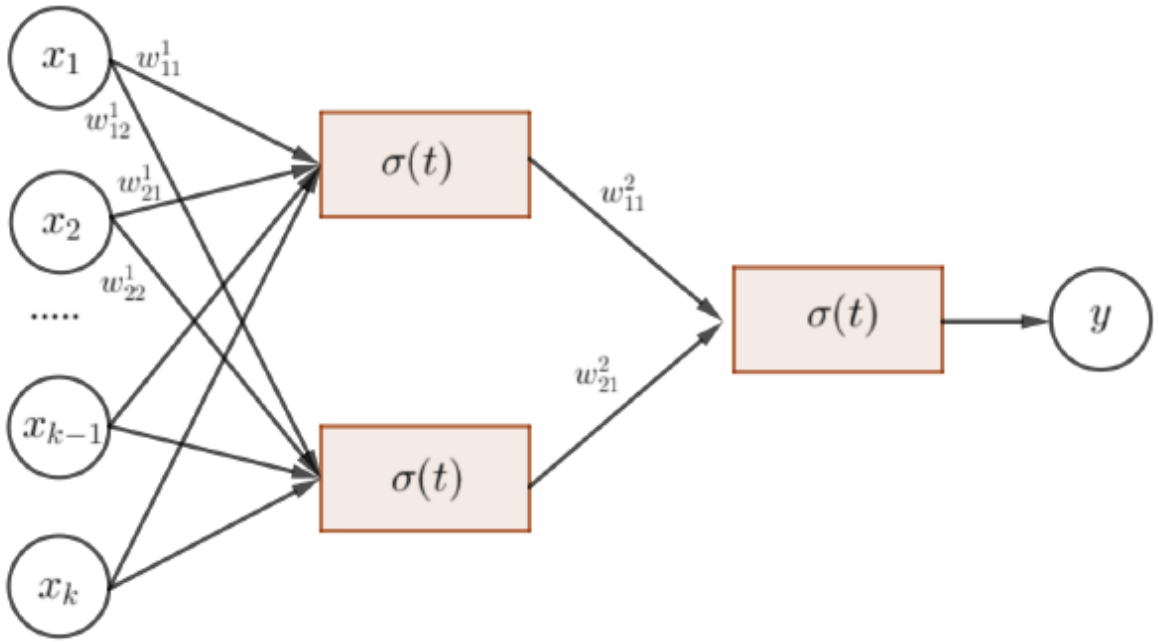
Что такое нейросеть?

Вспоминаем предыдущую лекцию

Что такое нейросеть?

Нейросеть это последовательная комбинация регрессий

Две регрессии скреплены третьей



$$h_j = \sigma(w_0 + w_{j1}^1 \cdot x_1 + ... + w_{jk}^1 \cdot x_k)$$
$$y = \sigma(w_{11}^2 \cdot h_1 + w_{21}^2 \cdot h_2)$$

Как оценить качество регрессии?

Используем метрики оценки качества регрессии

Метрики качества регрессии

Средняя квадратичная ошибка- MSE,Mean Squared Error

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(a(x_i) - y_i)^2$$

MSE применяется в ситуациях, когда нам надо подчеркнуть большие ошибки и выбрать модель, которая дает меньше больших ошибок прогноза

Метрики качества регрессии

Средняя абсолютная ошибка- MAE, Mean Absolute Error

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|a(x_i) - y_i|$$

MAE менее чувствителен к выбросам, чем **MSE.** Подходит для сравнения двух моделей, но позволяет сделать вывод на сколько хорошо решается задача.

Пример.
MSE=10 – плохо, если целевая переменная в диапазоне от 0 до 1.
MSE=10 – хорошо, если целевая переменная в диапазоне от 10000 до 10000

Метрики качества регрессии

Коэффициент детерминации – R^2

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (a(x_i) - y_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}$$

Коэффициент детерминации - нормированная среднеквадратичная ошибка.

Коэффициент детерминации измеряет долю дисперсии, объясненную моделью, в общей дисперсии целевой переменной.

Пример.
R^2≈0 – плохо
R^2≈1 – хорошо, модель объясняет данные

Метрики качества регрессии

Средняя абсолютная процентная ошибка – MAPE, Mean Absolute Percentage Error

$$MAPE = 100\% \times \frac{1}{n} \sum_{i=1}^{n} \frac{|y_i - a(x_i)|}{|y_i|}$$

Не имеет размерности, но очень просто интерпретируется.

Пример.

MAPE=11.4% - ошибка составила 11,4% от фактических значений.

Основная проблема данной ошибки – нестабильность

Симметричная MAPE – SMAPE, Symmetric MAPE

$$SMAPE = \frac{1}{n} \sum_{i=1}^{n} \frac{2 \times |y_i - a(x_i)|}{|y_i| + |a(x_i)|}$$

Метрики качества регрессии

Корень из средней квадратичной ошибки – RMSE, Root Mean Squared Error)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(a(x_i) - y_i)^2}$$

каждое отклонение возводится в квадрат, любое небольшое отклонение может значительно повлиять на показатель ошибки

Метрики качества регрессии

Средняя абсолютная масштабированная ошибка – MASE, Mean absolute scaled error)

$$MASE = \frac{\sum_{i=1}^{n} |Y_i - e_i|}{\frac{n}{n-1} \sum_{i=2}^{n} |Y_i - Y_{i-1}|}$$

Хороший вариант для расчета точности, так как сама ошибка не зависит от масштабов данных и является симметричной: то есть положительные и отрицательные отклонения от факта рассматриваются в равной степени.

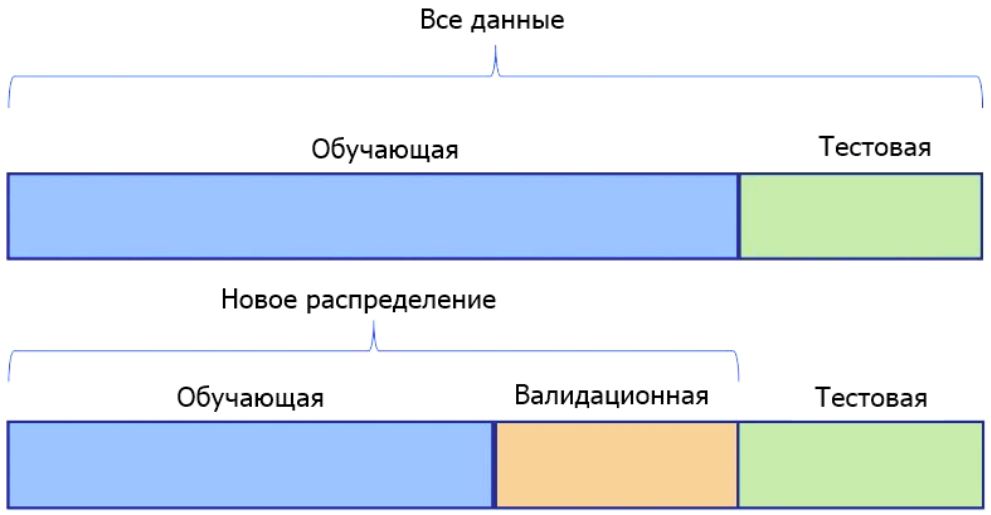
Недостаток – сложность в интерпретации

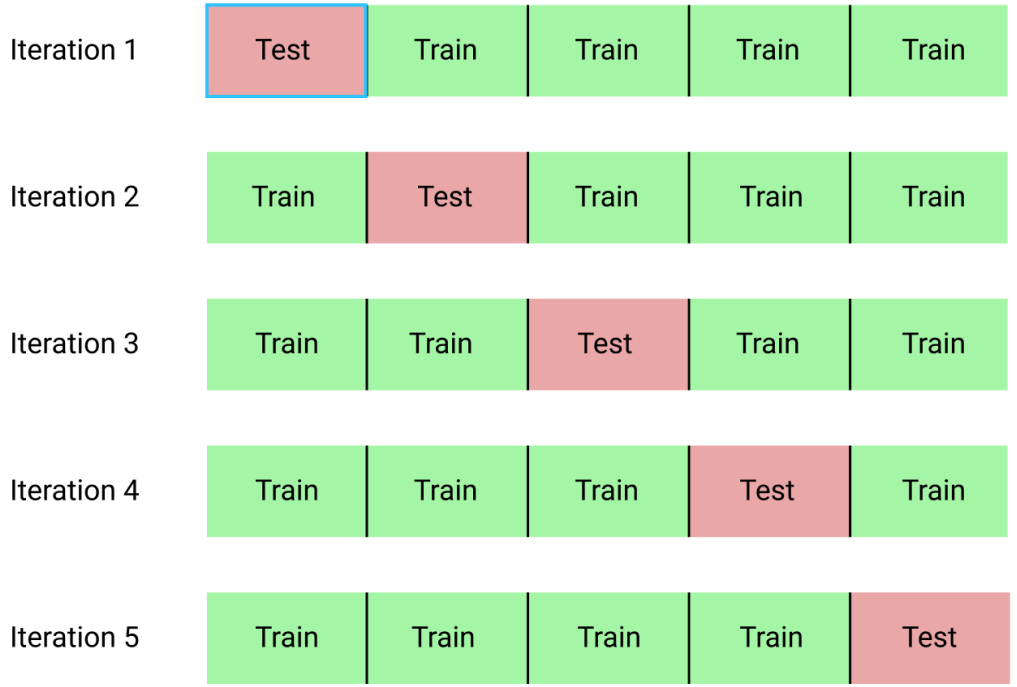
Метрики качества регрессии

Кросс-валидация

Перекрёстная проверка (кросс-проверка, скользящий контроль, **cross-validation**) — метод оценки аналитической модели и её поведения на независимых данных. При оценке модели имеющиеся в наличии данные разбиваются на k частей. Затем на k−1 частях данных производится обучение модели, а оставшаяся часть данных используется для тестирования. Процедура повторяется k раз; в итоге каждая из k частей данных используется для тестирования. В результате получается оценка эффективности выбранной модели с наиболее равномерным использованием имеющихся данных.

Тестовая и обучающая выборки

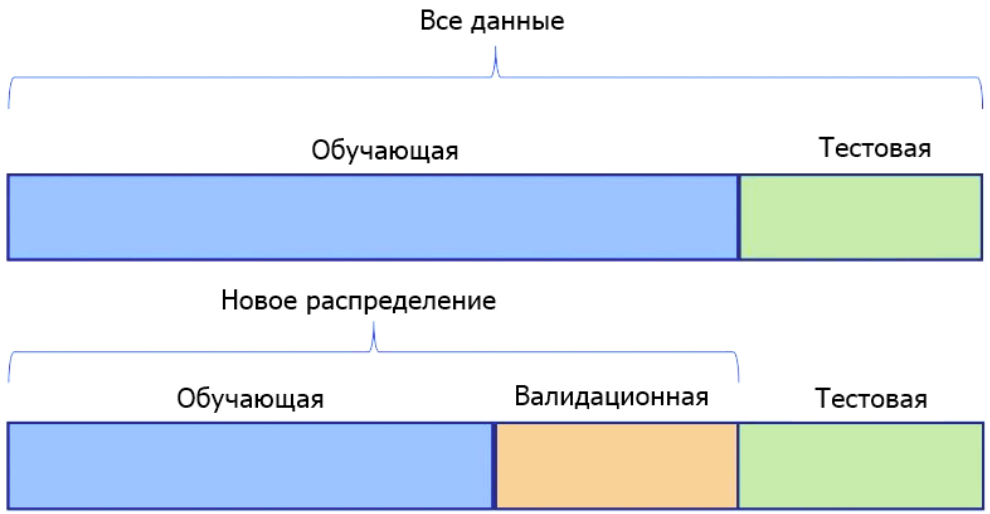




Метрики качества регрессии

Для разбиения выборки на обучающую и тестовую, есть встроенные функции

Тестовая и обучающая выборки



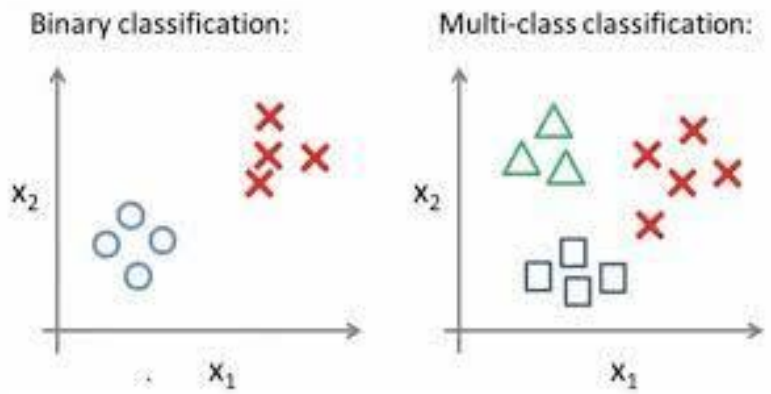

```
КОПИРОВАТЬ
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y,
                                                    train_size=0.67,
                                                    random_state=42)


print(f"Классы в y_train:\n{y_train}")
print(f"Классы в y_test:\n{y_test}")
```

Задача классификации

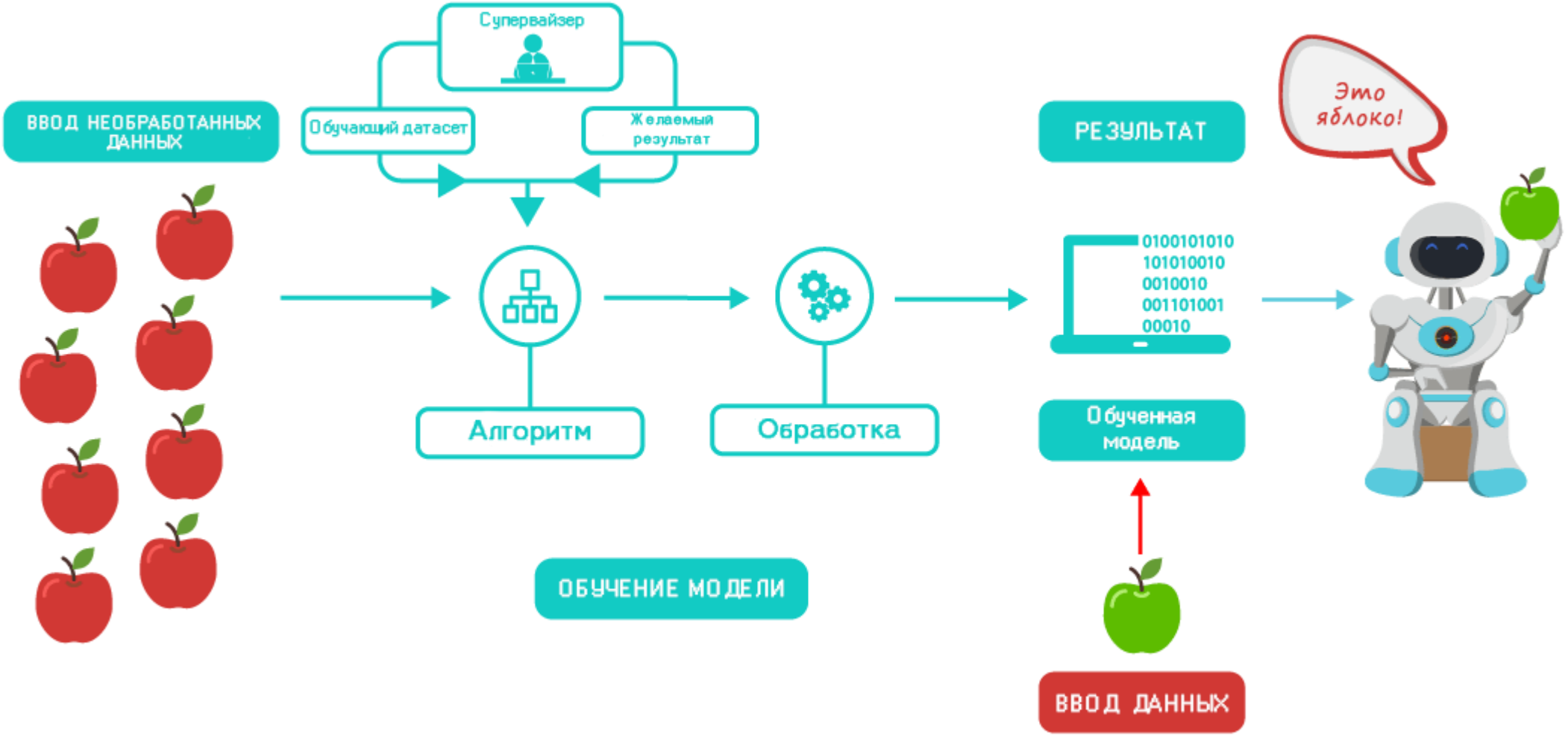
Классификация:
Имеется множество *объектов* (ситуаций), разделённых некоторым образом на *классы*. Задано конечное множество объектов, для которых известно, к каким классам они относятся (**обучающая выборка**). Классовая принадлежность остальных объектов не известна. Требуется построить **алгоритм**, способный классифицировать произвольный объект из исходного множества.



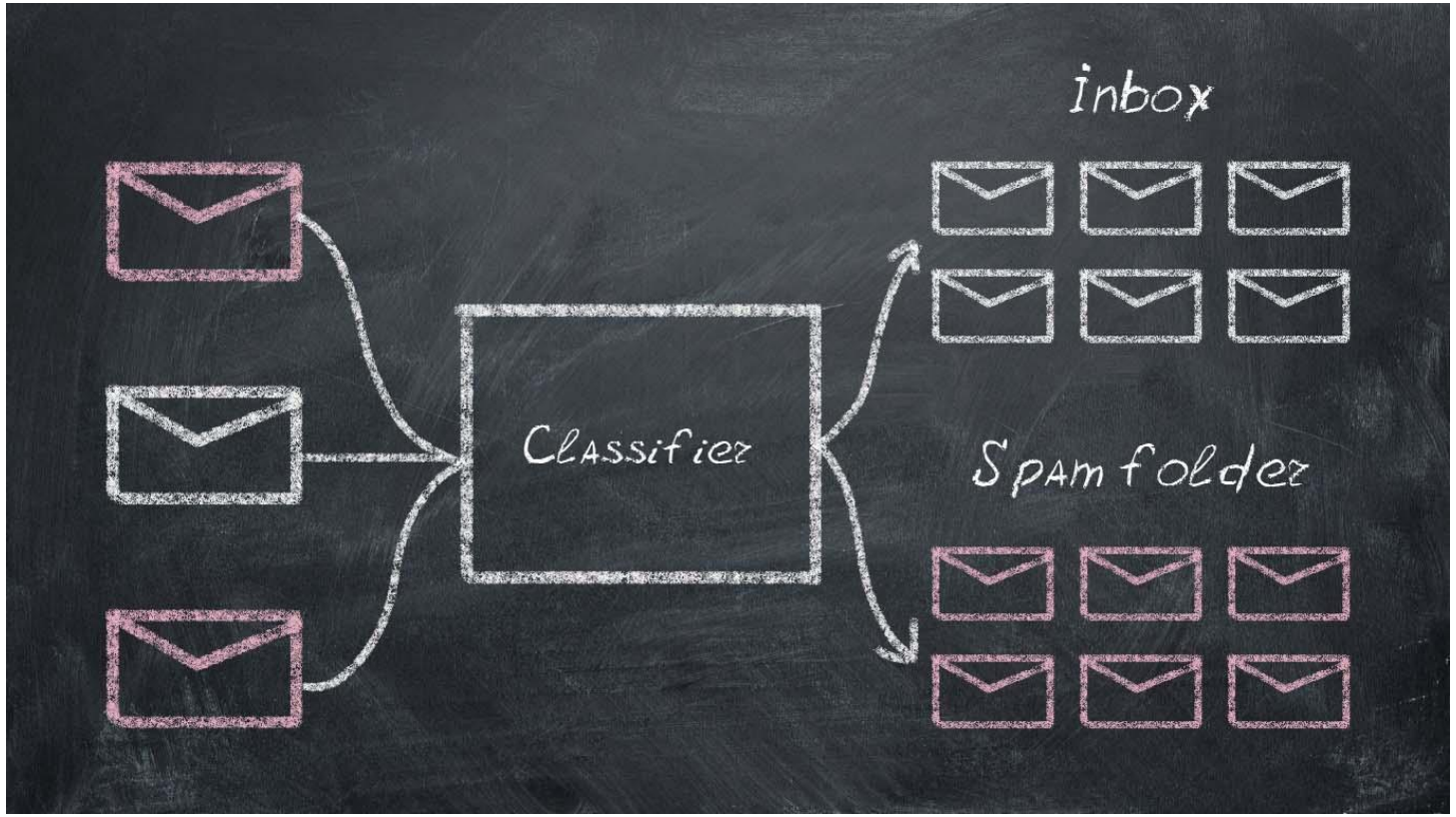
Классифицировать объект — значит, указать номер (или наименование класса), к которому относится данный объект.

Классификация объекта — номер или наименование класса, выдаваемый алгоритмом классификации в результате его применения к данному конкретному объекту.

Задача классификации



Задача классификации



Типы входных данных

Признаковое описание — наиболее распространённый случай. Каждый объект описывается набором своих характеристик, называемых признаками. Признаки могут быть числовыми или нечисловыми.

Матрица расстояний между объектами. Каждый объект описывается расстояниями до всех остальных объектов обучающей выборки. С этим типом входных данных работают немногие методы, в частности, метод ближайших соседей, метод парзеновского окна, метод потенциальных функций.

Временной ряд или **сигнал** представляет собой последовательность измерений во времени. Каждое измерение может представляться числом, вектором, а в общем случае — признаковым описанием исследуемого объекта в данный момент времени.

Изображение или **видеоряд**.

Встречаются и более сложные случаи, когда входные данные представляются в виде **графов**, **текстов**, **результатов запросов к базе данных**, и т. д. Как правило, они приводятся к первому или второму случаю путём предварительной обработки данных и извлечения признаков.

Классификацию сигналов и изображений называют также **распознаванием образов**.

Типы классов

Двухклассовая классификация. Наиболее простой в техническом отношении случай, который служит основой для решения более сложных задач.

Многоклассовая классификация. Когда число классов достигает многих тысяч (например, при распознавании иероглифов или слитной речи), задача классификации становится существенно более трудной.

Непересекающиеся классы.

Пересекающиеся классы. Объект может относиться одновременно к нескольким классам.

Нечёткие классы. Требуется определять степень принадлежности объекта каждому из классов, обычно это действительное число от 0 до 1.

Наборы данных для машинного обучения

Датасет для бинарной (двухклассовой) классификации binary classification

X1	X2	.....	Xn	Y1
12	Красный		1.7	1
14	Синий		6.4	0
2	Зелёный		12.3	0
14	Синий		12.4	1
12	Синий		1.2	0

X1	X2	.....	Xn	Y1
12	Красный		1.7	TRUE
14	Синий		6.4	FALSE
2	Зелёный		12.3	TRUE
14	Синий		12.4	TRUE
12	Синий		1.2	FALSE

X1	.....	Xn	Y1
12	Красный	0	Умный
14	Синий	1	Умный
2	Зелёный	1	Глупый
14	Синий	0	...
12	Синий	0	Умный

Наборы данных для машинного обучения

Датасет для мультиклассовой (двухклассовой) классификации multi-label classification

X1	X2	.....	Xn	Y1	Y2
12	Красный		1.7	1	0
14	Синий		6.4	0	1
2	Зелёный		12.3	0	1
14	Синий		12.4	1	0
12	Синий		1.2	0	1

X1	X2	.....	Xn	Y1	Y2
12	Красный		1.7	1	1
14	Синий		6.4	1	1
2	Зелёный		12.3	0	1
14	Синий		12.4	1	0
12	Синий		1.2	0	1

X1	.....	Xn	Y1
12	Красный	0	1
14	Синий	1	2
2	Зелёный	1	3
14	Синий	0	...
12	Синий	0	2

Метрики качества классификации

Матрица ошибок – Confusion matrix

(матрица «путаницы»)

Истинные значения

Предсказанные значения

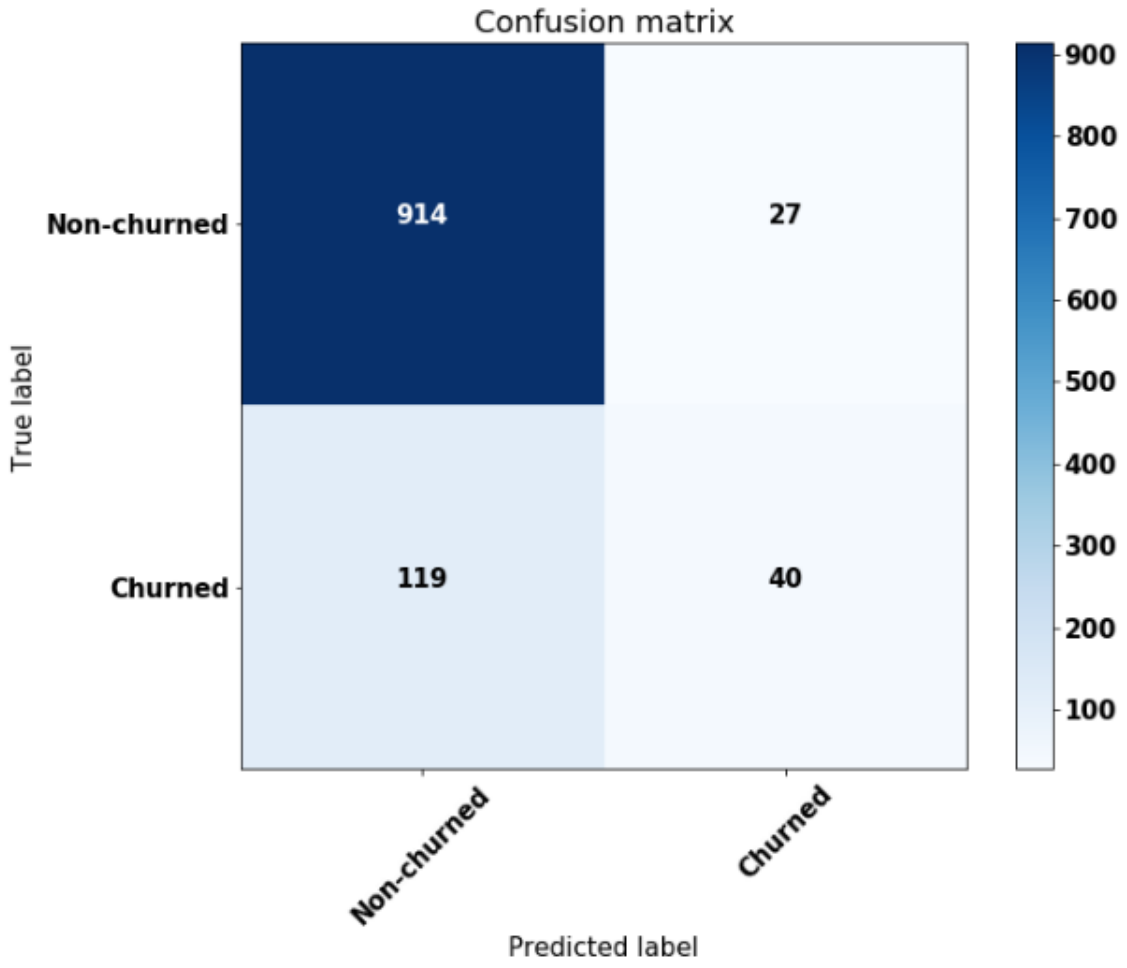
	$y = 1$	$y = 0$
$\hat{y} = 1$	True Positive (TP)	False Positive (FP)
$\hat{y} = 0$	False Negative (FN)	True Negative (TN)

Ошибка II рода

Ошибка I рода

Метрики качества классификации

Матрица ошибок – Confusion matrix (матрица «путаницы»)



Класс 1 – кредитоспособный заёмщик
Класс 0 – некредитоспособный заёмщик

1) Некредитоспособный заёмщик классифицирован как некредитоспособный (**True Positive** — **TP**).
2) Кредитоспособный заёмщик классифицирован как кредитоспособный, (**True Negative** — **TN**).
3) Кредитоспособный заёмщик классифицирован как некредитоспособный (**False Positive** — **FP**), **ошибка I рода**.
4) Некредитоспособный заёмщик распознан как кредитоспособный, (**False Negative** — **FN**), **ошибка II рода**.

Метрики качества классификации

Аккуратность (Accuracy) – доля правильных ответов

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

Пример: Имеется алгоритм классификации писем на 2 класса: спам, не-спам. Результаты классификации 100 не-спам писем, 90 из которых наш классификатор определил верно (True Negative = 90, False Positive = 10), и 10 спам-писем, 5 из которых классификатор также определил верно (True Positive = 5, False Negative = 5).

Тогда accuracy:

$$accuracy = \frac{5 + 90}{5 + 90 + 10 + 5} = 86,4$$

Метрики качества классификации

Точность (Precision) – доля истинно положительных классификаций(доля объектов, классифицированных как положительные и действительно являющиеся таковыми)

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

Особенности оценки качества классификации

1. Метрика `accuracy` зависят от соотношения классов.
2. Метрики `precision` и `recall` не зависят от соотношения классов и потому применимы в условиях несбалансированных выборок.
3. Высокие `precision` и `recall` `это здорово, но одновременно увеличить обе метрики невозможно – ищем баланс.`

Вывод: Нужна метрика, которая объединяет `precision` и `recall`

Метрики качества классификации

F-мера (F-score) – гармоническое среднее между точностью и полнотой

$$F_\beta = (1 + \beta^2) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall}$$

β – вес точности в метрике (как правило берём β = 1)

F-мера достигает максимума при полноте и точности, равными единице, и близка к нулю, если один из аргументов близок к нулю

Метрики качества классификации

F-мера (F-score)

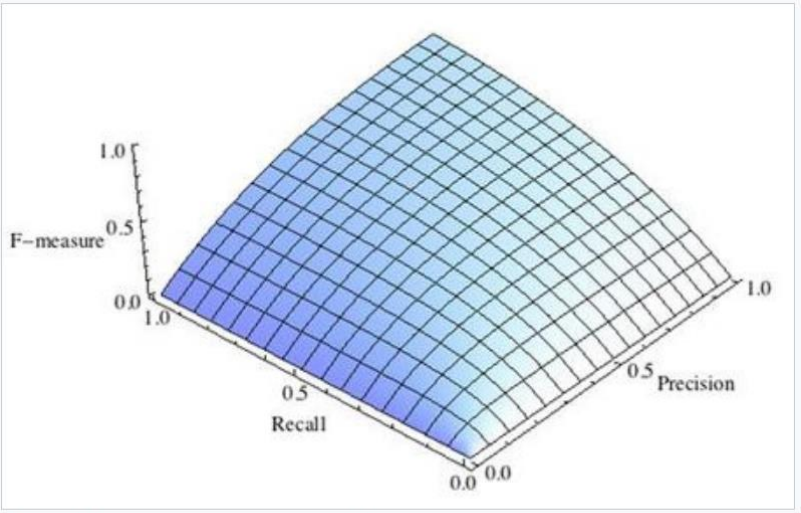


Рис.1 Сбалансированная F-мера, $\beta = 1$

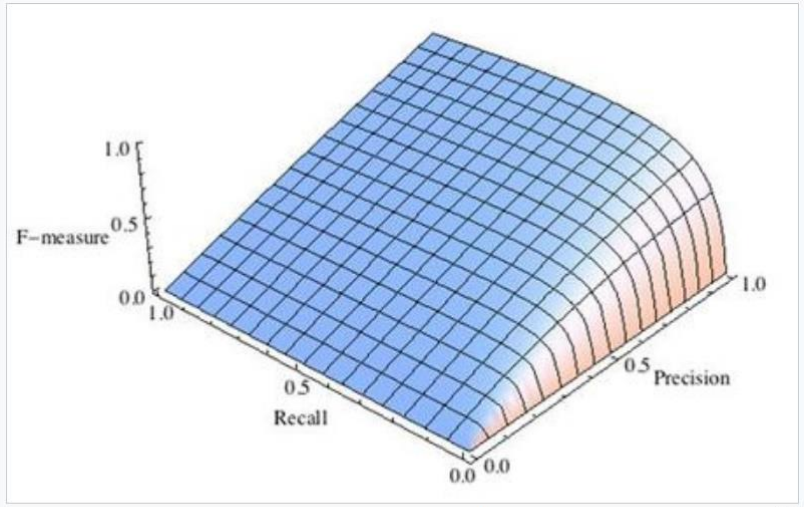


Рис.2 F-мера с приоритетом точности, $\beta^2 = \frac{1}{4}$

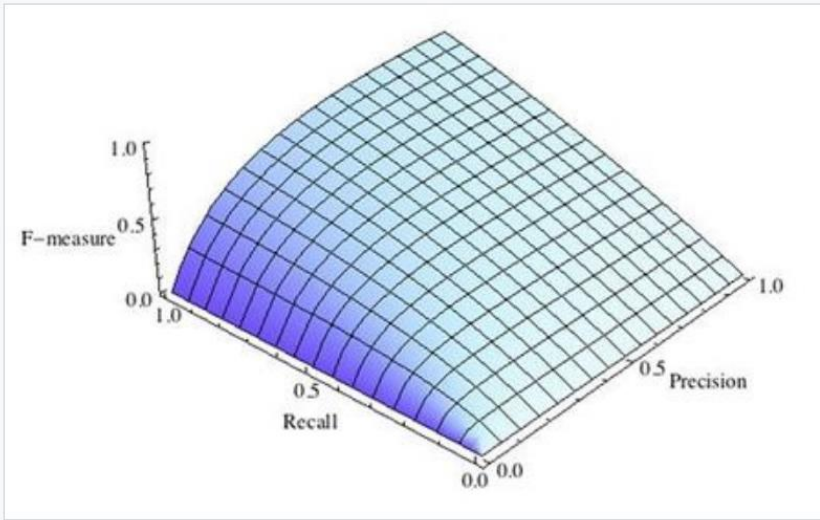
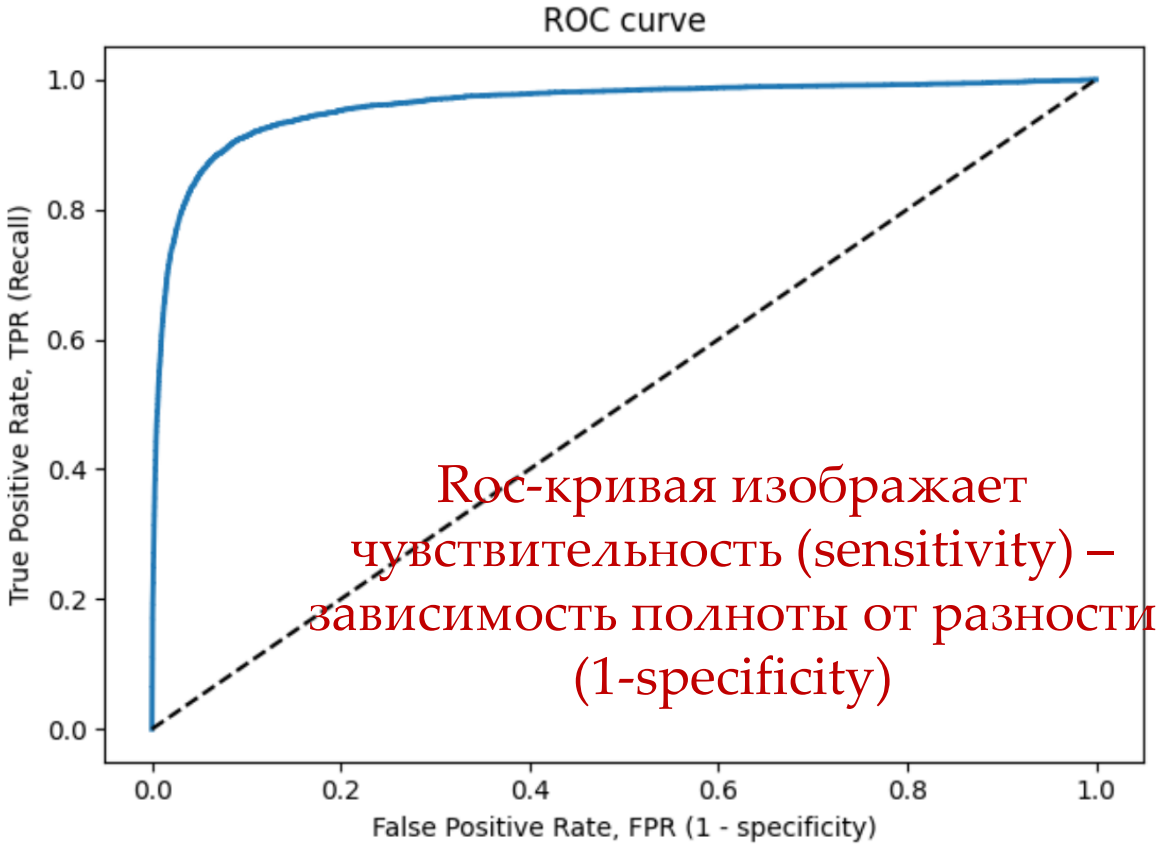


Рис.3 F-мера с приоритетом полноты, $\beta^2 = 2$

Метрики качества классификации

ROC-кривая (ROC-curve, Receiver Operating Characteristics curve**)**
– показывает долю истино полжительных решений (TPR) в сравнении с долей ложно положительных решений (FPR)



Roc-кривая изображает чувствительность (sensitivity) – зависимость полноты от разности (1-specificity)

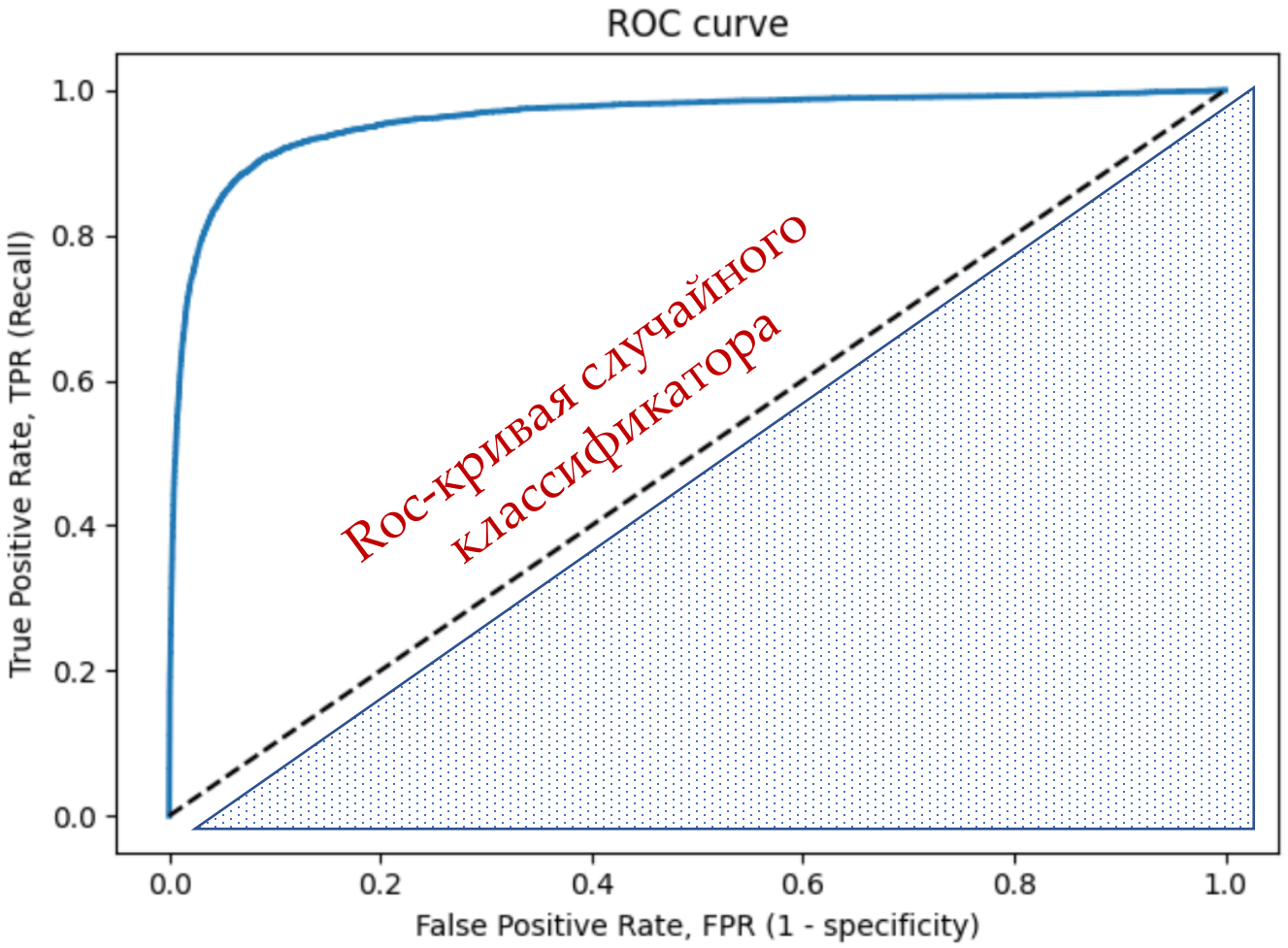
$$TPR = \frac{TP}{TP + FN} = \text{Recall}$$

$$FPR = \frac{FP}{FP + TN} = 1 - TNR$$

TNR – специфичность (specificity)

Метрики качества классификации

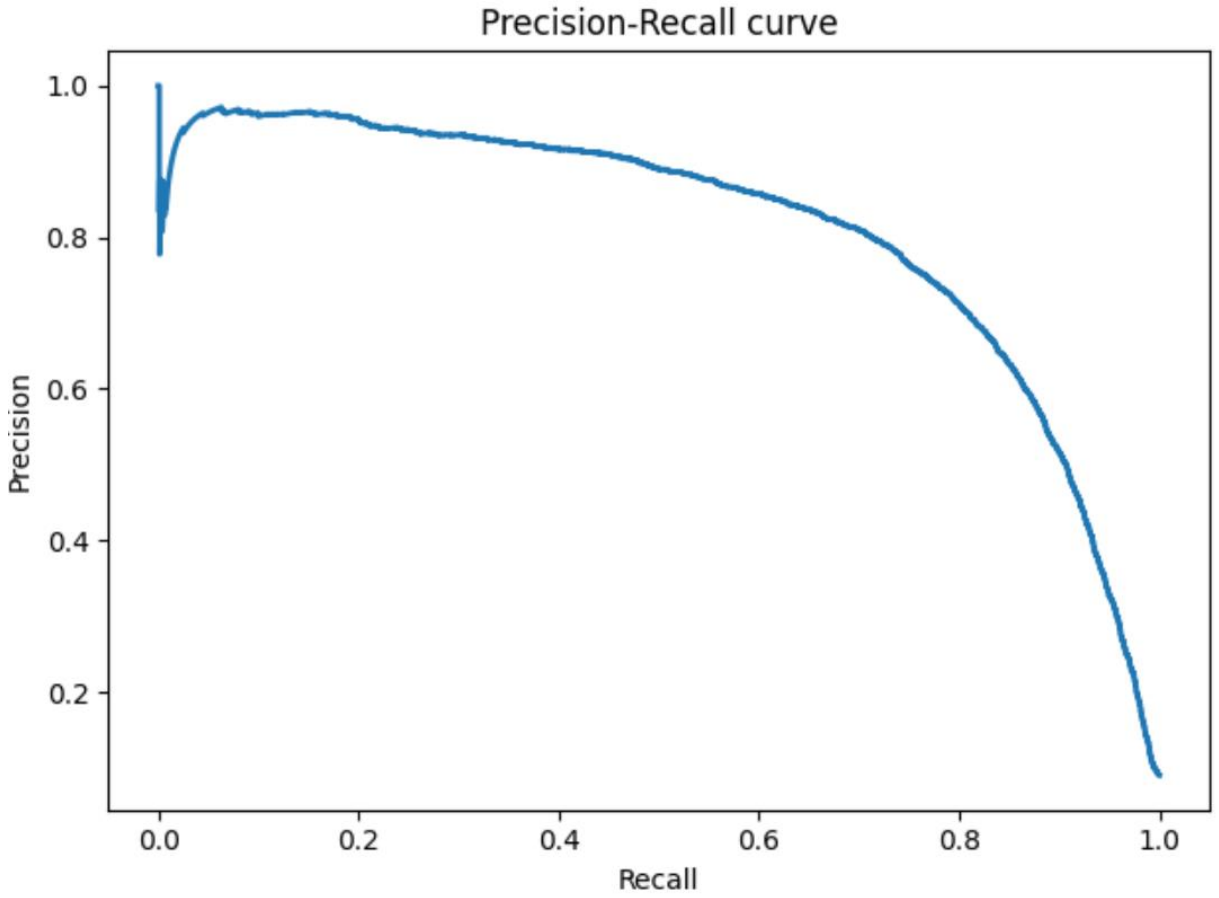
AUC-ROC (Area Under the Curve) – **площадь под ROC-кривой**



Безупречный классификатор будет иметь площадь под ROC-кривой (ROC-AUC), равную 1, тогда как чисто случайный классификатор - площадь 0.5.

Метрики качества классификации

PR-кривая (PR-curve, Precission-recall curve**) – показывает зависимость точности от полноты. Используется вместо ROC-кривой при несбалансированных классах**



Критерием качества семейства алгоритмов выступает **площадь под PR-кривой** (англ. **Area Under the Curve — AUC-PR**)