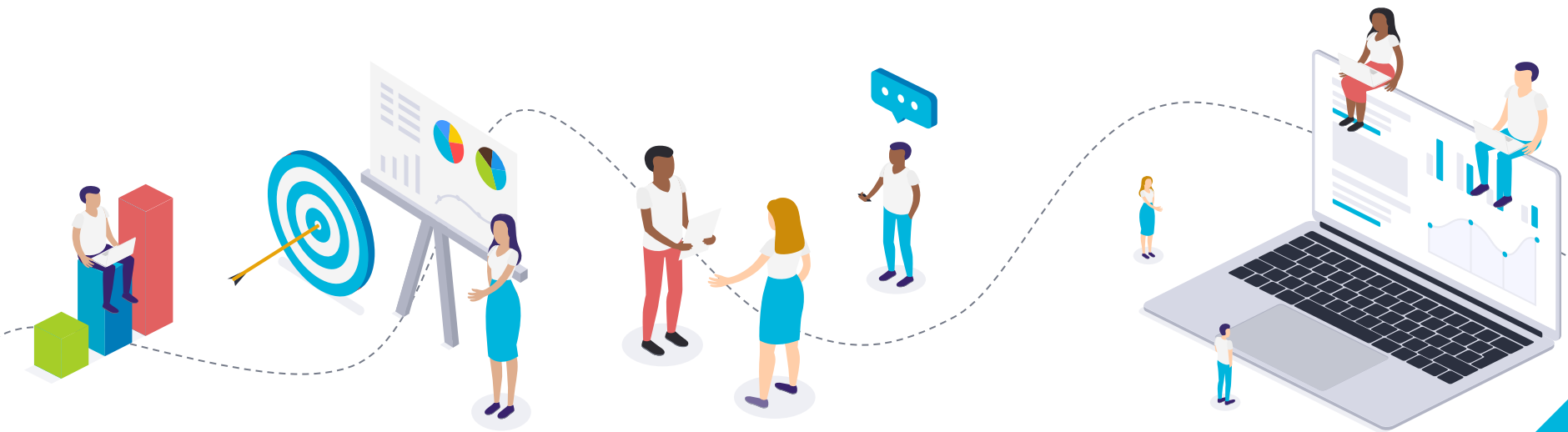


## Программа профессиональной переподготовки «Технологии искусственного интеллекта, визуализации и анализа данных»



## Метрики оценки качества классификации

### 1) Доля правильных ответов (accuracy)

$$\text{accuracy}(\alpha, X) = \frac{1}{n} \sum_{i=1}^n [a(x_i) = y_i]$$

- Не рекомендуется использовать при несбалансированных выборках.

Предположим, выборка содержит 1000 объектов: 950 из класса 1, 50 из класса 2. Модель все имеющиеся объекты определяет в класс 1, но при этом  $\text{accuracy} = 0.95$ .

- Не учитывает цену ошибок.

Цена какой из ошибок выше: диагностировать заболевание у человека, у которого его нет, или не диагностировать заболевание у человека, у которого оно на самом деле есть?

[sklearn.metrics.accuracy\\_score – scikit-learn 1.0 documentation](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html)

## Метрики оценки качества классификации

Как соотносятся между собой результат работы алгоритма и истинный ответ удобно оценивать с помощью **матрицы ошибок**.

	$y = 1$	$y = -1$
$a(x) = 1$	True Positive (TP)	False Positive (FP)
$a(x) = -1$	False Negative (FN)	True Negative (TN)

False Positive (ложно положительный результат) и False Negative (ложно отрицательный результат) – это ошибки классификации. В статистике False Positive называют ошибкой первого рода, а False Negative – ошибкой второго рода.

[https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion\\_matrix.html](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html)

## Метрики оценки качества классификации

### 2) Точность (precision)

$$\text{precision}(\alpha, X) = \frac{TP}{TP + FP}$$

- Под *точностью* понимают долю объектов, определяемых классификатором как положительную, которая действительно является положительной.
- Повышение точности классификации позволяет сократить количество ложноположительных (FP) объектов.

[sklearn.metrics.precision\\_score](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_score.html) — [scikit-learn 1.0 documentation](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_score.html)

	$y = 1$	$y = -1$
$a(x) = 1$	True Positive (TP)	False Positive (FP)
$a(x) = -1$	False Negative (FN)	True Negative (TN)

## Метрики оценки качества классификации

### 3) Полнота (recall)

$$\text{recall}(\alpha, X) = \frac{TP}{TP + FN}$$

- *Полнота* показывает какая доля положительных объектов из всех положительных объектов была выделена классификатором.
- Увеличение значения полноты означает снижение количества ложноотрицательных (FN) объектов.

[https://scikit-learn.org/stable/modules/generated/sklearn.metrics.recall\\_score.html](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.recall_score.html)

	$y = 1$	$y = -1$
$a(x) = 1$	True Positive (TP)	False Positive (FP)
$a(x) = -1$	False Negative (FN)	True Negative (TN)

## Метрики оценки качества классификации

Необходимо построить модель, которая определяет, есть или нет определенное заболевание у пациента. При этом требуется, чтобы были выявлены как минимум 80% пациентов, которые действительно имеют данное заболевание. Тогда ставят задачу максимизации точности при условии  $recall(a, X) \geq 0.8$ . Следует особо обратить внимание на то, как точность и полнота работают в случае несбалансированных выборок.

Пусть рассматривается выборка со следующей матрицей ошибок:

	$y = 1$	$y = -1$
$a(x) = 1$	10	20
$a(x) = -1$	90	10000

$$accuracy(a, X) = 0.99, \quad precision(a, X) = 0.33, \quad recall(a, X) = 0.1.$$

То, что доля верных ответов равняется 0.99, ни о чем не говорит: алгоритм все равно делает 66% ложных срабатываний и выявляет только 10% положительных случаев. Благодаря использованию точности и полноты становится понятно, что алгоритм нужно улучшать.

## Метрики оценки качества классификации

- 4) F-мера – гармоническое среднее точности и полноты:

$$F = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

[https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1\\_score.html](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html)

## Метрики оценки качества классификации

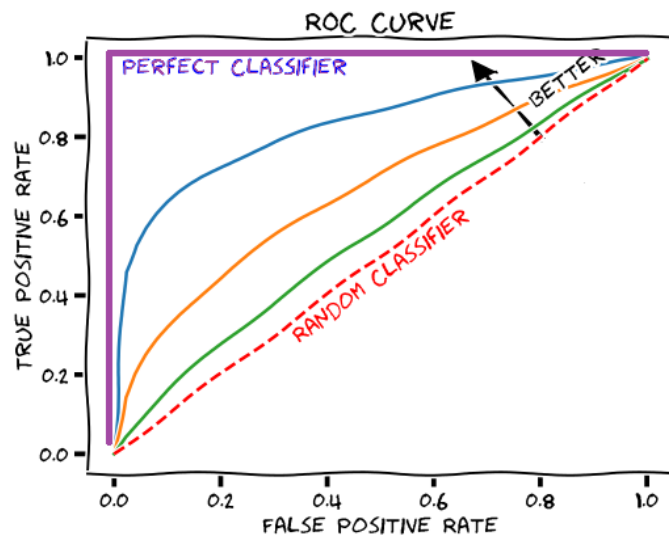
### 5) ROC-кривая

График характеристической кривой (receiver operator characteristic, ROC) – это широко используемый инструмент для отбора моделей классификации данных, основываясь на их качестве относительно долей ложно положительных и истинно положительных исходов, которые рассчитываются путем смещения порога решения классификатора.

ROC-кривая строится в координатах False Positive Rate (доля ложно положительных объектов) и True Positive Rate (доля истинно положительных объектов), и позволяет рассмотреть все пороговые значения для классификатора.

$$FPR = \frac{FP}{FP + TN}, \quad TPR = \frac{TP}{TP + FN}.$$

Идеальная ROC-кривая проходит через левый верхний угол и соответствует классификатору, который дает высокое значение полноты при низкой доле ложноположительных (FP) примеров.



[https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc\\_curve.html](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_curve.html)



## Метрики оценки качества классификации

### 6) AUC (area under the curve)

AUC принимает значения от 0 до 1, в предположении, что чем значение AUC ближе к 1, тем лучше качество модели.

В случае, когда классификатор не делает ошибок, т.е.  $FPR=0$ , а  $TPR=1$ , мы получим  $AUC=1$ . В противном случае, когда классификатор случайным образом определяет принадлежность объектов классам, AUC будет стремиться к 0.5.

[https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc\\_auc\\_score](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_auc_score)

Дополнительно: PR-кривая (изучить самостоятельно)

[https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision\\_recall\\_curve.html](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_recall_curve.html)

## Как оценивать многоклассовую классификацию?

В задачах многоклассовой классификации, как правило, стараются свести оценку качества к вычислению рассмотренных ранее метрик. Используются два подхода: *микро*- и *макро*-усреднение.

Рассмотрим на примере точности.

- При микро-усреднении TP и FP усредняются по всем классам, а затем вычисляется итоговая характеристика:

$$\text{precision}(\alpha, X) = \frac{\overline{TP}}{\overline{TP} + \overline{FN}}$$

- При макро-усреднении сначала вычисляется итоговая метрика для каждого класса, а затем результаты усредняются по всем классам:

$$\text{precision}(\alpha, X) = \frac{1}{K} \sum_{k=1}^K \text{precision}_k(\alpha, X), \quad \text{precision}_k(\alpha, X) = \frac{TP_k}{TP_k + FP_k}$$