СПбГЭТУ «ЛЭТИ»
ПЕРВЫЙ ЭЛЕКТРОТЕХНИЧЕСКИЙ

Александр Калиниченко

ИНТЕЛЛЕКТУАЛЬНЫЕ ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ В МЕДИЦИНЕ

Модуль 2. Методы искусственного интеллекта

Тема 5. Метрики оценки эффективности моделей

МЕТРИКИ БИНАРНОЙ КЛАССИФИКАЦИИ

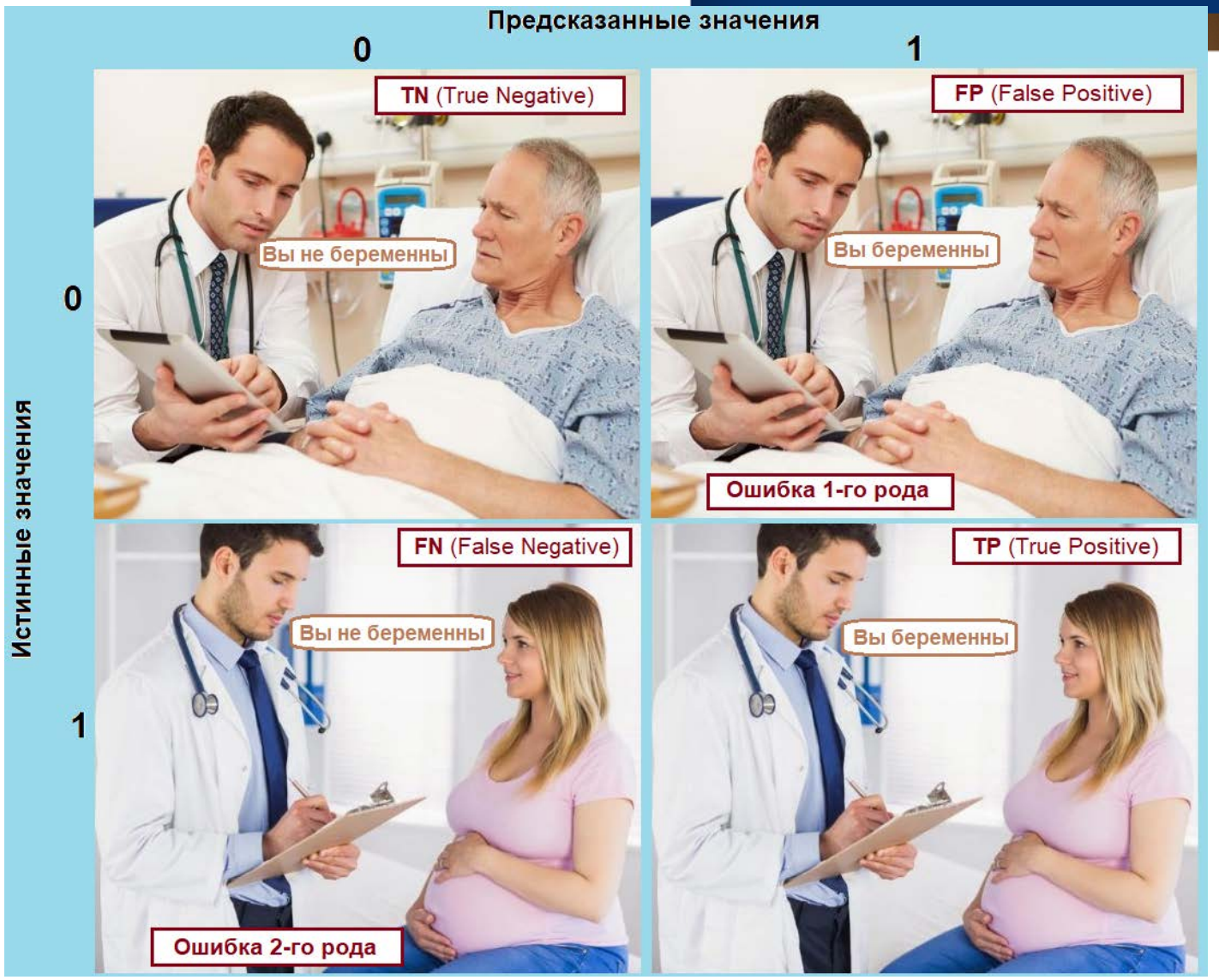
Обозначение	Название	Смысл
TP **True Positive**	Истинный положительный результат	Результат теста, который правильно указывает на наличие определенного условия или атрибута
TN **True Negative**	Истинный отрицательный результат	Результат теста, который правильно указывает на отсутствие определенного условия или атрибута
FP **False Positive**	Ложный положительный результат. Ошибка 1-го рода	Результат теста, который ошибочно указывает на наличие определенного условия или атрибута
FN **False Negative**	Ложный отрицательный результат. Ошибка 2-го рода	Результат теста, который ошибочно указывает на отсутствие определенного условия или атрибута

МАТРИЦА ОШИБОК

Confusion matrix (матрица ошибок, матрица несоответствий):

		Predicted: FALSE	Predicted: TRUE
Actual	FALSE	True Negative (TN)	False Positive (FP)
	TRUE	False Negative (FN)	True Positive (TP)

FALSE = 0
TRUE = 1

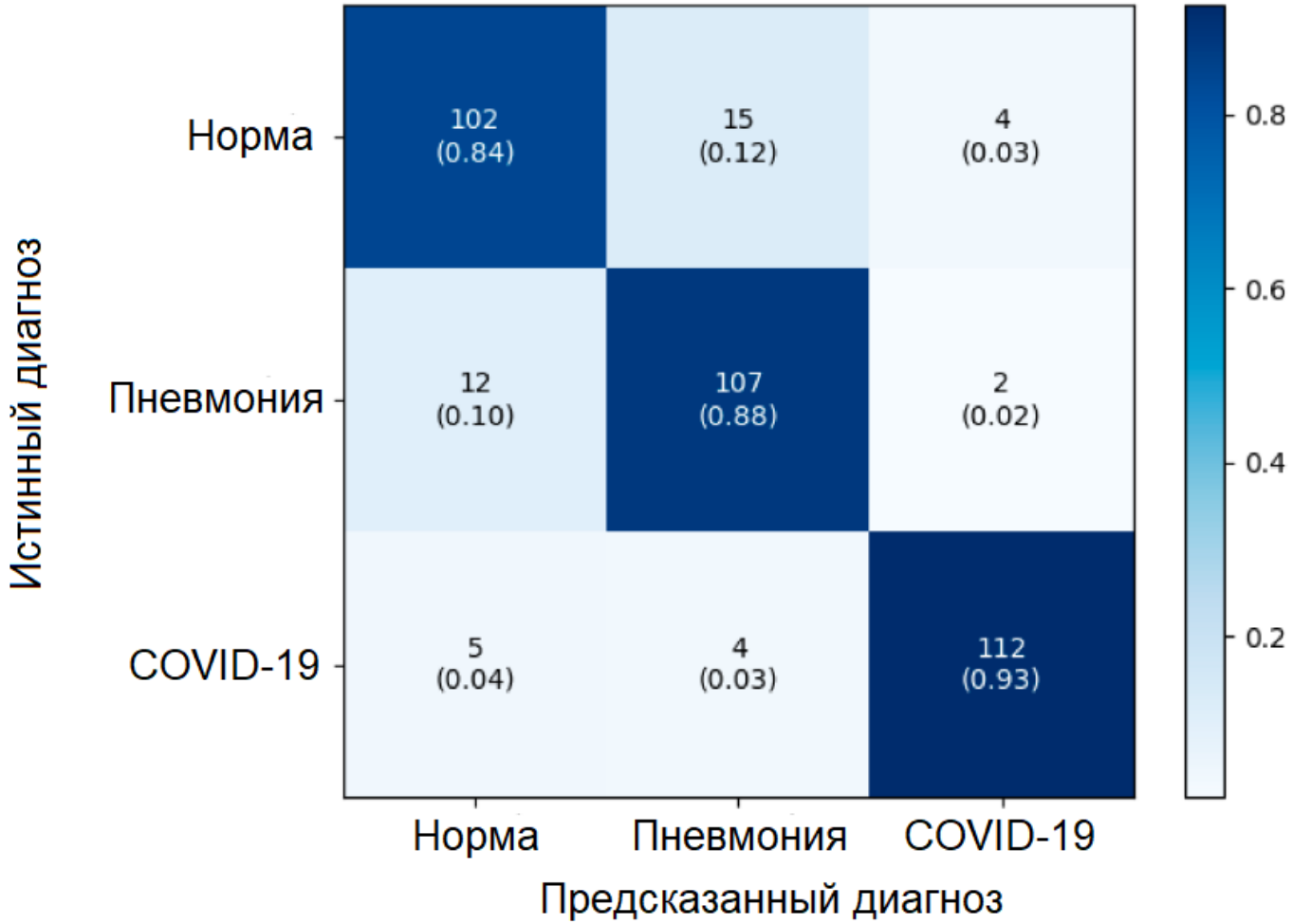


ПРИМЕР МАТРИЦЫ ОШИБОК

Всего пациентов **P+N = 8+4 = 12**		**Прогнозируемый диагноз**	
		Нераковые заболевания	Рак
Истинный диагноз	Нераковые заболевания N=4	TN=3	FP=1
	Рак P=8	FN=2	TP=6

Данная матрица показывает, что у 2-х из 8-ми пациентов с раком система ошибочно определила отсутствие рака, а из 4 пациентов без рака она у 1-го предсказала, наличие рака

МАТРИЦА ОШИБОК ПРИ МНОГОКЛАССОВОЙ КЛАССИФИКАЦИИ



ТОЧНОСТЬ И ПОЛНОТА

Точность – доля объектов, названных классификатором положительными и при этом действительно являющихся положительными

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

Полнота – показывает, какую долю объектов положительного класса из всех объектов положительного класса нашел алгоритм

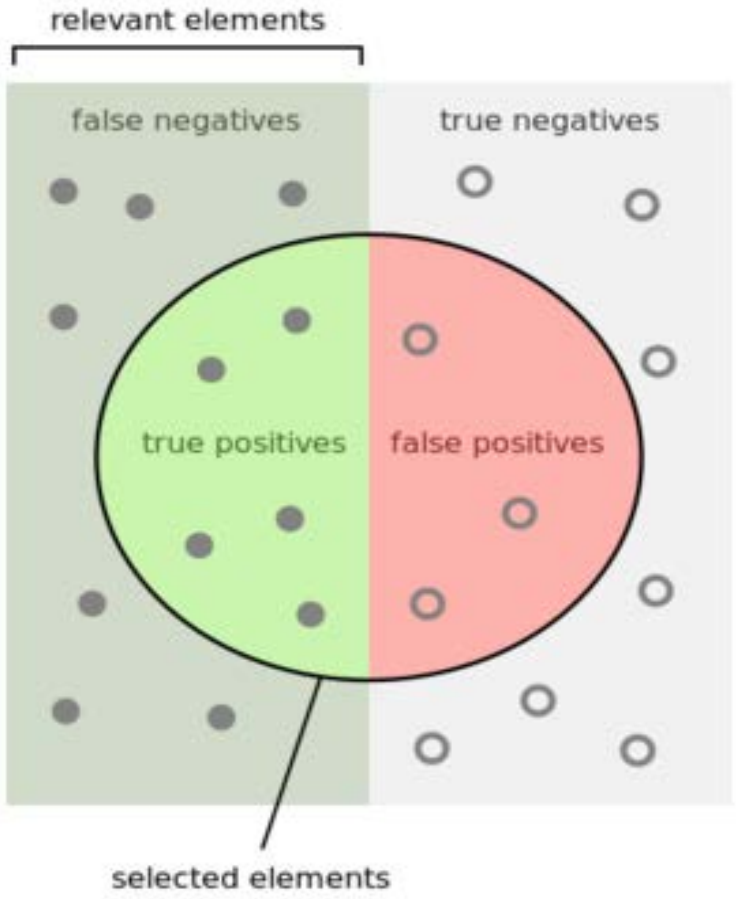
$$recall = \frac{TP}{TP + FN}$$

Интуитивно понятная, очевидная и почти неиспользуемая метрика - **доля правильных ответов алгоритма.**

Эта метрика бесполезна в задачах с неравными классами

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

ИЛЛЮСТРАЦИЯ ТОЧНОСТИ И ПОЛНОТЫ



How many selected items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{selected elements}}$$

How many relevant items are selected?

$$\text{Recall} = \frac{\text{true positives}}{\text{relevant elements}}$$

ИНТЕГРАЛЬНЫЕ ОЦЕНКИ

F-мера - среднее гармоническое precision и recall :

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

β определяет вес точности в метрике, и при β=1 это среднее гармоническое (с множителем 2, чтобы в случае precision = 1 и recall = 1 иметь F=1).

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

ФУНКЦИИ ПОТЕРЬ

- Среднеквадратическая ошибка (используется в задаче **регрессии**):

$$MSE = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - a_i)^2}$$

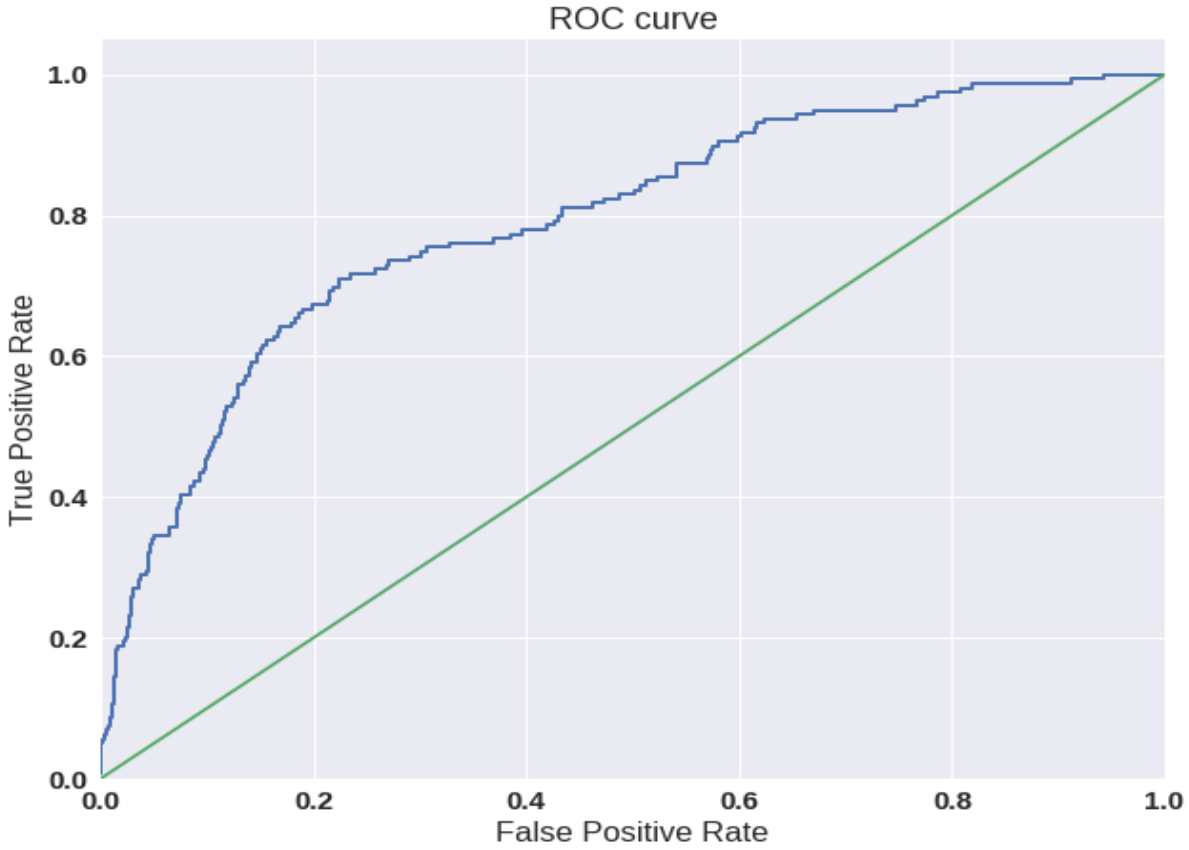
здесь a_i – это ответ алгоритма на i-ом объекте, y_i — истинная метка класса на i-ом объекте, а N – размер выборки.

Логистическая функция потерь:

$$\text{logloss} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(a_i) - (1 - y_i)\log(1 - a_i)] \rightarrow \min$$

Интуитивно можно представить минимизацию logloss как задачу максимизации accuracy путем штрафа за неверные предсказания.

ROC-КРИВАЯ И AUC



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

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

TPR – полнота (recall) - доля объектов положительного класса из всех объектов положительного класса

FPR - доля неверно предсказанных объектов negative класса

AUC (area under curve) – площадь под кривой
AUC=1 – безошибочная классификация
AUC=0.5 – случайный выбор

ЧУВСТВИТЕЛЬНОСТЬ И СПЕЦИФИЧНОСТЬ

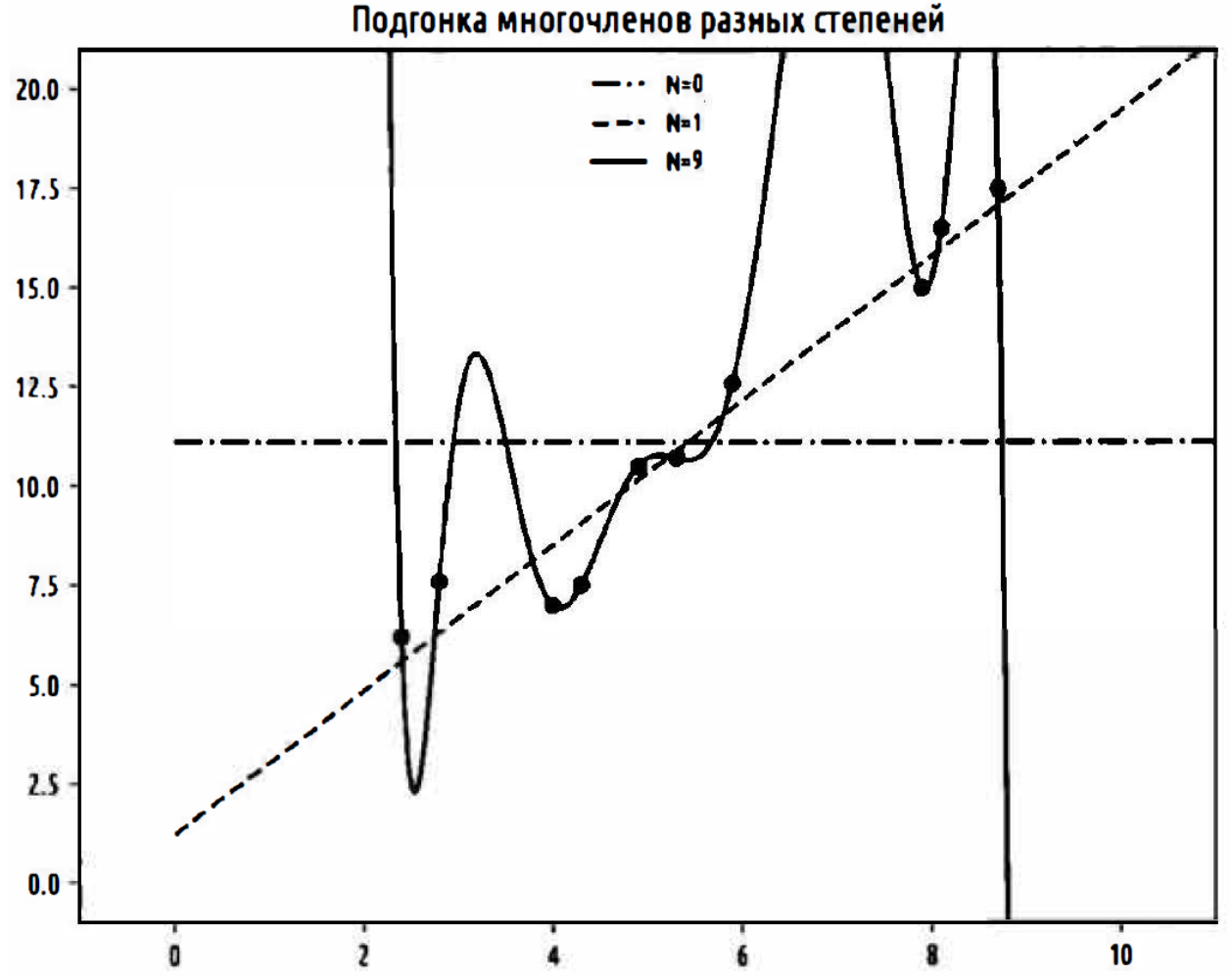
- **Чувствительность** – (истинно положительная пропорция) отражает долю положительных результатов, которые правильно идентифицированы как таковые. Иными словами, чувствительность диагностического теста показывает вероятность того, что больной субъект будет классифицирован именно как больной

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

- **Специфичность** – (истинно отрицательная пропорция) отражает долю отрицательных результатов, которые правильно идентифицированы как таковые, то есть вероятность того, что не больные субъекты будут классифицированы именно как не больные

$$\text{Specificity} = \frac{TN}{TN + FP}$$

НЕДООБУЧЕНИЕ И ПЕРЕОБУЧЕНИЕ



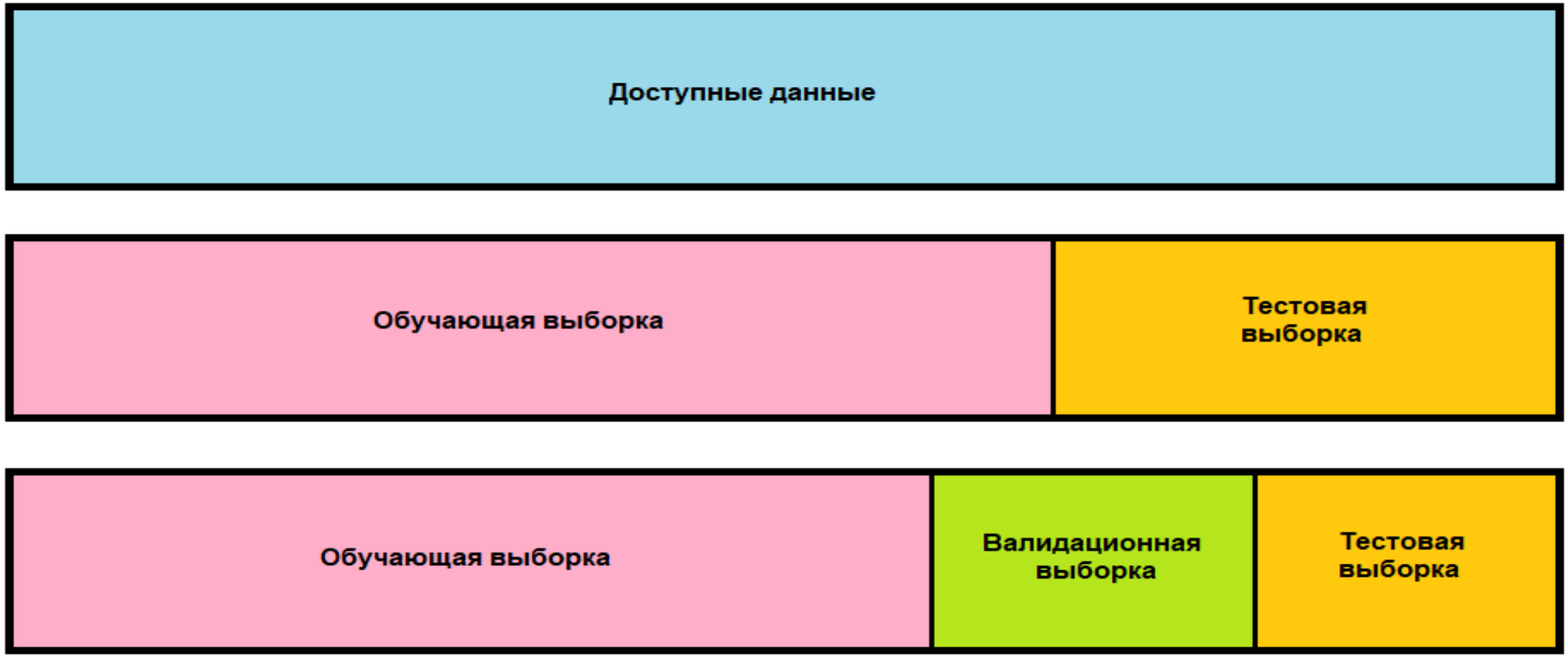
ОБУЧАЮЩАЯ И ТЕСТОВАЯ ВЫБОРКИ

Главная задача обучаемых алгоритмов – их способность *обобщаться*, то есть хорошо работать на новых данных. Поскольку на новых данных мы сразу не можем проверить качество построенной модели (нам ведь надо для них сделать прогноз, то есть истинных значений целевого признака мы для них не знаем), то надо пожертвовать небольшой порцией данных, чтобы на ней проверить качество модели.

Чаще всего это делается одним из двух способов:

- *отложенная выборка*
- *кросс-валидация*

ВЫБОРКИ ДАННЫХ В МАШИННОМ ОБУЧЕНИИ

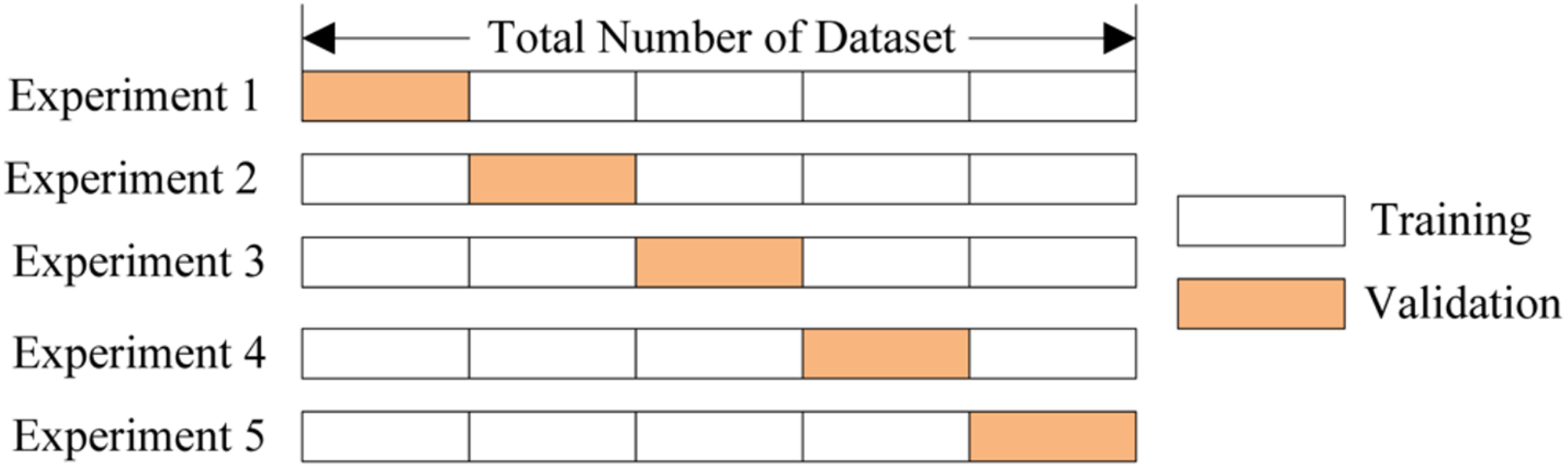


ОТЛОЖЕННАЯ ВЫБОРКА

Отложенная выборка *(held-out/hold-out set)*. При таком подходе мы оставляем какую-то долю обучающей выборки (как правило от 20% до 40%), обучаем модель на остальных данных (60-80% исходной выборки) и считаем некоторую метрику качества модели (например, самое простое – долю правильных ответов в задаче классификации) на отложенной выборке.

КРОСС-ВАЛИДАЦИЯ (1)

Кросс-валидация (cross-validation, на русский еще переводят как скользящий или перекрестный контроль). Самый частый случай – K-fold кросс-валидация:



КРОСС-ВАЛИДАЦИЯ (2)

- При кросс-валидации модель обучается K раз на разных (K−1) подвыборках исходной выборки, а проверяется на одной подвыборке (каждый раз на разной).
- Получаются K оценок качества модели, которые обычно усредняются, выдавая среднюю оценку качества классификации/регрессии на кросс-валидации.
- Кросс-валидация дает лучшую по сравнению с отложенной выборкой оценку качества модели на новых данных. Но кросс-валидация вычислительно дорогостоящая, если данных много.
- Кросс-валидация – очень важная техника в машинном обучении (применяемая также в статистике и эконометрике), с ее помощью выбираются гиперпараметры моделей, сравниваются модели между собой, оценивается полезность новых признаков в задаче и т.д.