

# Машинное обучение: оценка качества в машинном обучении

MADE academy  
Эмели Драль

# Базовые концепции машинного обучения

1. Виды обучения, виды задач, базовые концепции
2. Простые алгоритмы: логика построения и связь с математикой
3. Оценка качества в машинном обучении

# Метрики качества

1. Валидация моделей
2. Метрики качества в задачах классификации
3. Метрики качества в задачах регрессии
4. Метрики качества в задачах ранжирования

# Валидация моделей

# Валидация моделей

## Базовые концепты

Объекты и признаки:

- $x$  – объект
- $y$  – ответ
- $(f_1, f_2 \dots f_n)$  – признаки, описывающие объекты
- $F^{(l,n)}$  – матрица объект-признак
- $X$  – пространство объектов
- $Y$  – пространство ответов

Модель:

- $a: X \rightarrow Y$
- $a(x) = y$
- $A$  – семейство моделей

Оценка качества

- $Q(a, X)$  – ошибки модели  $a(x)$  на группе объектов  $X$

# Как построить модель?

1. Поставить задачу и подготовить набор данных  $X = (x_i, y_i)_{i=1,l}$
2. Выбрать семейство моделей  $A$
3. Минимизировать ошибки модели  $Q(a, X) \rightarrow$   
за счет этого получить конкретную модель  $a(x)$  из выбранного семейства  $A$

Валидация  
моделей

## Минимизация ошибок модели

С одной стороны, мы действительно строим конкретную модель  $a(x)$  из выбранного семейства  $A$  за счет минимизации  $Q(a, X)$ . Например, мы оцениваем такие параметры, как:

### Валидация моделей

1. Байесовский классификатор: параметры распределения из выбранного семейства для каждого из признаков
2. Дерево решений: структура дерева (последовательность выбранных порогов)

## Минимизация ошибок модели

С другой стороны, **не все параметры** модели поддаются оптимизации в процессе **обучения**.  
Например:

### Валидация моделей

1. Байесовский классификатор: семейство распределений для признаков
2. Дерево решений: критерий для оценки разбиения ( $H(j, t)$ ,  $G(j,t)$ , misclassification)
3. Метод ближайших соседей: количество соседей, метрика близости

# Виды параметров

Параметры модели делятся на 2 группы:

1. Гиперпараметры – параметры, значения которых фиксируются до обучения. Они определяют вид модели и процесс обучения.
2. Параметры – параметры, значения которых оцениваются в процессе обучения.

# Подбор параметров

Гиперпараметры и параметры оптимизируют по-разному:

1. Мы подбираем гиперпараметры с помощью отложенной (валидационной) выборки или процесса кросс-валидации
2. Мы оцениваем параметры в процессе обучения модели (часто, решая оптимизационную задачу)

## Валидация моделей

# Валидационная выборка

Данные делятся на 3 выборки:

- Обучающая выборка
- Валидационная выборка
- Тестовая выборка

Валидация  
моделей

# Валидационная выборка

Данные делятся на 3 выборки:

- Обучающая выборка
- Валидационная выборка
- Тестовая выборка

Обучение – для **построения** модели

Валидация – для **оценки качества** модели

Тест – для **проверки на переобучение\*** и наличие технических ошибок

\*переобучение под обучающую выборку или подбор параметров, оптимальный для фиксированной валидационной выборки

## Валидация моделей

# Валидационная выборка

Стратегии разбиения данных:

- последовательно во времени
- случайно
- случайно стратифицировано

Соотношения по размеру могут отличаться:

- 70/20/10
- 60/20/20
- 50/30/20

Важно, чтобы в обучающей выборке хватило данных для обучения. И чтобы оценки по валидации и тесту были достаточно надежны (интервальная оценка!)

## Валидация моделей

# Валидация моделей

## Валидационная выборка

Процесс валидации:

1. Фиксируем интересующие значения параметров
2. Строим модель на обучающей выборке
3. Оцениваем качество на валидации
4. Повторяем 1-3 с другими наборами параметров
5. Выбираем лучшую модель
6. Оцениваем её на тестовой выборке, исследуем разницу в качестве на валидации и teste
7. При отсутствии существенных отличий в оценках на валидации и teste считаем модель финальной
8. Можно перестроить модель на обучении + валидации

# Кросс-валидация (cross validation, cv)

Помните, мы опасались подобрать параметры, переобучившись под выбранную валидационную выборку?

## Валидация моделей

# Кросс-валидация

Помните, мы опасались подобрать параметры, переобучившись на выбранную валидационную выборку?

## Валидация моделей

Для того, чтобы избавиться от влияния конкретного разбиения на обучение и валидацию, давайте сделаем такое разбиение несколько раз!

# Кросс-валидация

Для того, чтобы избавиться от влияния конкретного разбиения на обучение и валидацию, давайте сделаем такое разбиение несколько раз!

1. Разбиваем данные на **k** частей



# Валидация моделей

## Кросс-валидация

Для того, чтобы избавиться от влияния конкретного разбиения на обучение и валидацию, давайте сделаем такое разбиение несколько раз!

1. Разбиваем данные на  $k$  частей



2.  $k-1$  часть объединяется в обучающую выборку,  
1 часть остается для оценка качества



# Кросс-валидация: k-fold

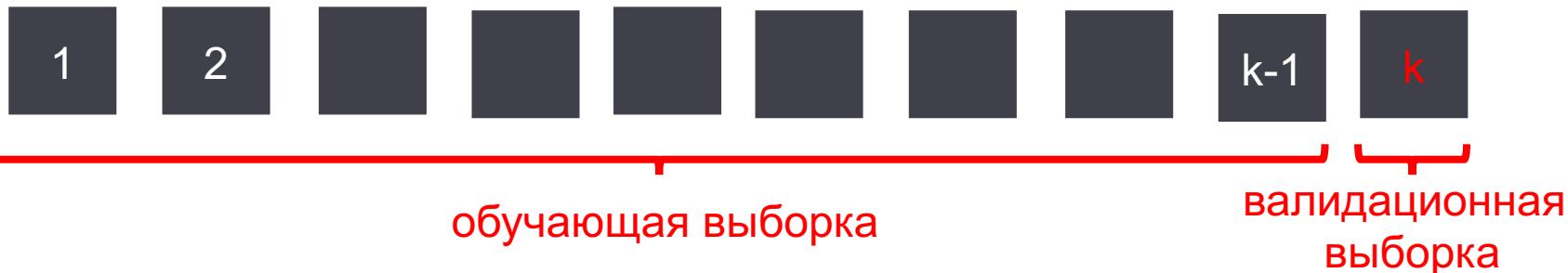
Для того, чтобы избавиться от влияния конкретного разбиения на обучение и валидацию, давайте сделаем такое разбиение несколько раз!

## Валидация моделей

1. Разбиваем данные на  $k$  частей



2.  $k-1$  часть объединяется в обучающую выборку,  
1 часть остается для оценка качества



3. Повторяем  $k$  раз так, чтобы каждая часть 1 раз  
стала валидационный выборкой

# Валидация моделей

## Кросс-валидация: $k$ -fold

Повторяем процесс разбиения данных на  $k$  частей  $t$  раз, для каждого разбиения производим  $k$ -fold cv

1. Разбиваем данные на  $k$  частей



2.  $k-1$  часть объединяется в обучающую выборку,  
 $1$  часть остается для оценка качества



обучающая выборка

валидационная  
выборка

3. Повторяем  $k$  раз так, чтобы каждая часть 1 раз  
стала валидационный выборкой

# Стратегии кросс-валидация

Внутри k-fold возможны различные стратегии разбиения данных:

- Random split
- Stratified split
- Leave-one-out (LOO)

Альтернативная, но похожая стратегия:

- Random shuffle
- Bootstrap

# Особые случаи: временные ряды

timeseries cross validation: moving window



Валидация  
моделей

# Особые случаи: временные ряды

timeseries cross validation: moving window with a fixed width



Валидация  
моделей

# Особые случаи: сессии

Классический вариант:

- Делим данные на выборки по id объекта, в данном случае по событиям или по сессиям

Валидация  
моделей

## Особые случаи: сессии

Классический вариант:

- Делим данные на выборки по id объекта, в данном случае по событиям или по сессиям

# Валидация моделей

Возможно, полезная правка для пользовательских сессий:

- Все события из одной сессии лежат в одной выборке
- Все сессии одного клиента лежат в одной выборке

# Валидация моделей

## Практические рекомендации

1. Предпочитайте **cv** фиксированной валидационной выборке
2. Не забывайте про **отложенный тест**, он поможет найти нетривиальные ошибки
3. На практике чаще всего ограничиваются **k-fold** ( $k = 5$  или  $10$ )
4. Выбирайте подходящую **стратегию cv**  
Контрольный вопрос: каковы недостатки выбранной стратегии **cv**, можно ли получить завышенную/заниженную оценку?
5. Помните про **особые случаи**

# Update: как построить модель?

1. Подготовить набор данных  $X = (x_i, y_i)_{i=1,l}$
2. Выбрать семейство моделей  $A$
3. Минимизировать ошибки модели  $Q(a, X)$ :
  - 3.1 выбрать **гиперпараметры** модели с помощью **кросс-валидации**
  - 3.2 зная гиперпараметры, подобрать **параметры** модели в результате **минимизации**  $Q(a, X)$  на всей обучающей выборке

## Валидация моделей

# Метрики качества в задачах классификации

# Метрики качества: классификация

## Метрики качества

- Accuracy
- Precision
- Recall
- F-measure
- ROC-AUC
- Log loss

# Accuracy

Доля правильных ответов при классификации

Метрики  
качества:  
классификация

# Accuracy

Доля правильных ответов при классификации

Метрики  
качества:  
классификация

target: 1 0 1 0 0 0 1 0 0

## Метрики качества: классификация

# Accuracy

Доля правильных ответов при классификации

target: 1 0 1 0 0 0 0 1 0 0

predicted: 0 0 1 0 0 0 0 1 1 0

## Метрики качества: классификация

# Accuracy

Доля правильных ответов при классификации

target: 1 0 1 0 0 0 0 1 0 0

predicted: 0 0 1 0 0 0 0 1 1 0

## Метрики качества: классификация

# Accuracy

Доля правильных ответов при классификации

target: 1 0 1 0 0 0 0 1 0 0

predicted: 0 0 1 0 0 0 0 1 1 0

$$\text{accuracy} = 8/10 = 0.8$$

# Метрики качества: классификация

## Метрики качества

- Accuracy
- Precision
- Recall
- F-measure
- ROC-AUC
- Log loss

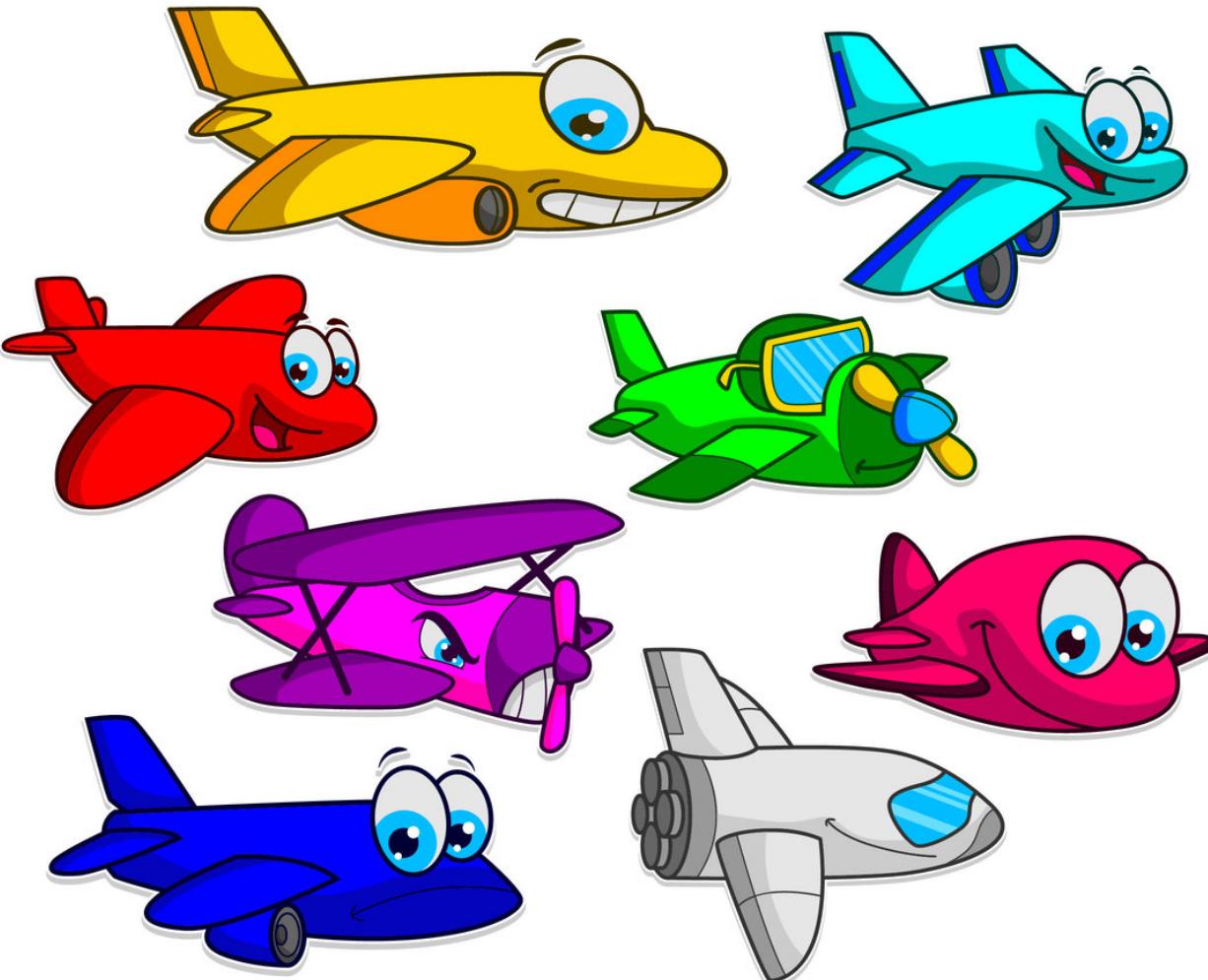
# Precision & Recall

- Precision – точность
- Recall - полнота

Метрики  
качества:  
классификация

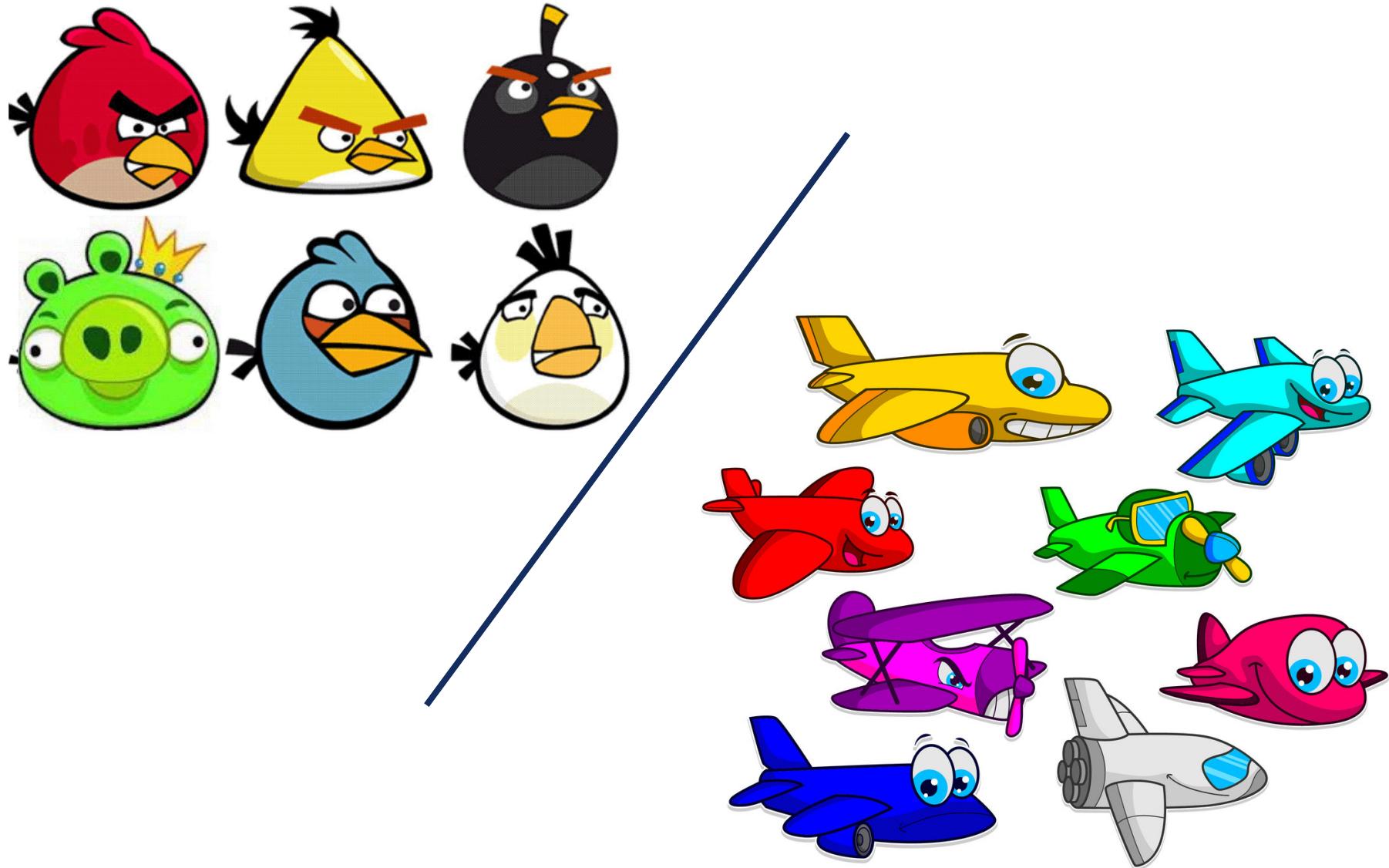
Метрики  
качества:  
классификация

# Сбитые самолёты



Метрики  
качества:  
классификация

## Сбитые самолёты



## Метрики качества: классификация

### Precision

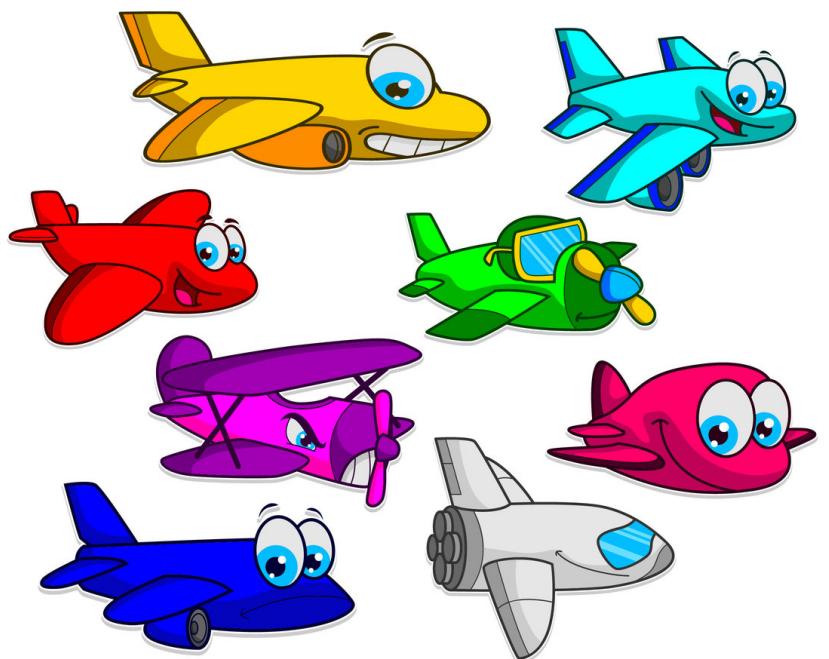
- Precision – точность выстрелов
- Количество сбитых самолётов/количество выстрелов



## Метрики качества: классификация

### Recall

- Recall – доля сбитых самолетов:
- Количество сбитых самолётов/общее количество самолётов



Метрики  
качества:  
классификация

## Считать вот так

		Actual Class	
		Yes	No
Predicted Class	Yes	True Positive	False Positive
	No	False Negative	True Negative

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

## Метрики качества: классификация

### F-measure (F-score, F1)

- Среднее гармоническое между precision и recall:
- Значение F-measure ближе к меньшему из precision и recall

$$F1 = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

# Метрики качества: классификация

## Метрики качества

- Accuracy
- Precision
- Recall
- F-measure
- ROC-AUC
- Log loss

## Метрики качества: классификация

### ROC AUC

- Применяется для оценки вероятностной классификации и ранжирования
- «Качество» ранжирования объектов по вероятности принадлежности к целевому классу
- Доля правильно отранжированных пар
- Вероятность встретить объект целевого класса раньше, чем объект нецелевого класса

Метрики  
качества:  
классификация

## ROC curve

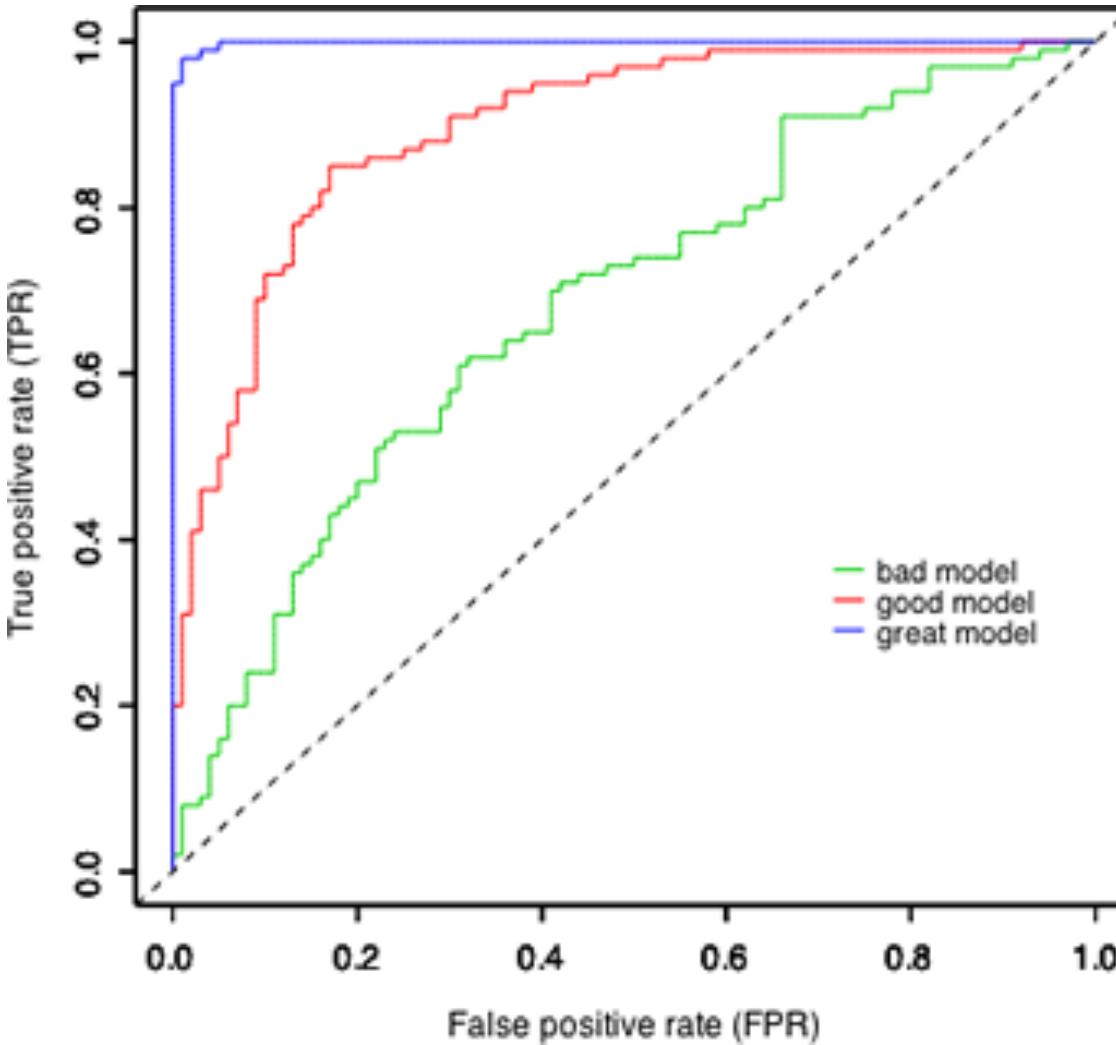
		Actual Class	
		Yes	No
Predicted Class	Yes	True Positive	False Positive
	No	False Negative	True Negative

$$TPR = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}}$$

$$FPR = \frac{\text{False positives}}{\text{False positives} + \text{True negatives}}.$$

Метрики  
качества:  
классификация

## ROC curve



# ROC curve

Как оценить кривую численно?

Метрики  
качества:  
классификация

# ROC curve

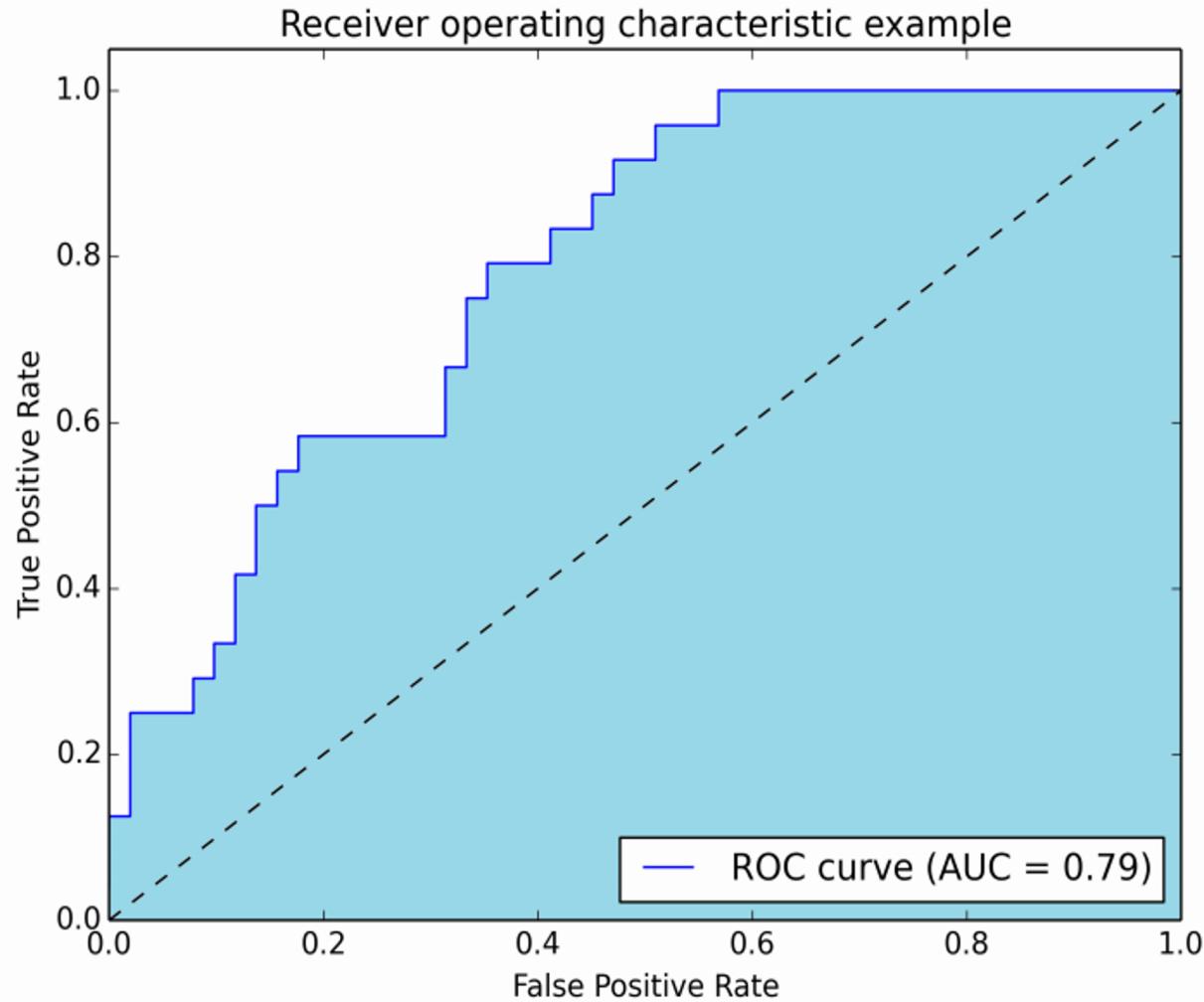
Как оценить кривую численно?

Измерить площадь под кривой – area under the curve!

Метрики  
качества:  
классификация

# Метрики качества: классификация

## ROC AUC



## ROC curve

Что если классификация всё же не вероятностная?

- Существуют способы адаптации ROC AUC для этого случая
- Однако пользоваться ими не рекомендуется

Метрики  
качества:  
классификация

# Log loss

Логарифмическая ошибка

Хорошо оценивает вероятность

Метрики  
качества:  
классификация

$$\text{LogLoss} = - \frac{1}{n} \sum_{i=1}^n [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$$

# Метрики качества: классификация

Пусть  $p_i = P(y_i = 1|x_i)$ , тогда  $1 - p_i = P(y_i = 0|x_i)$

# Метрики качества: классификация

Пусть  $p_i = P(y_i = 1|x_i)$ , тогда  $1 - p_i = P(y_i = 0|x_i)$

Теперь заметим, что выражение  $p_i^{y_i}(1 - p_i)^{1-y_i}$  -  
просто запись вероятности того класса, к которому  $x_i$   
фактически принадлежит

# Метрики качества: классификация

Пусть  $p_i = P(y_i = 1|x_i)$ , тогда  $1 - p_i = P(y_i = 0|x_i)$

Теперь заметим, что выражение  $p_i^{y_i}(1 - p_i)^{(1-y_i)}$  – просто запись вероятности того класса, к которому  $x_i$  фактически принадлежит

Произведение вероятностей фактических классов объектов из выборки – правдоподобие выборки:

$$\prod_{i=1}^n p_i^{y_i}(1 - p_i)^{(1-y_i)}$$

# Метрики качества: классификация

Пусть  $p_i = P(y_i = 1|x_i)$ , тогда  $1 - p_i = P(y_i = 0|x_i)$   
Теперь заметим, что выражение  $p_i^{y_i}(1 - p_i)^{(1-y_i)}$  –  
просто запись вероятности того класса, к которому  $x_i$   
фактически принадлежит

Произведение вероятностей фактических классов  
объектов из выборки – правдоподобие выборки:

$$\prod_{i=1}^n p_i^{y_i}(1 - p_i)^{(1-y_i)}$$

Если взять логарифм и умножить на  $-1$  – получим  $\log$  loss. Таким образом минимизация  $\log$  loss  
эквивалентна максимизации правдоподобия выборки!

# Метрики качества в задачах регрессии

# Метрики качества: регрессия

## Метрики качества

- ME
- MAE
- RMSE
- MAPE
- SMAPE

## Mean Absolute Error

- Отклонение прогноза от исходного значения
- Усредненное по всем наблюдениям

$$\text{MAE} = \frac{1}{n} \sum_{j=1}^n |y_j - \hat{y}_j|$$

Метрики  
качества:  
регрессия

# Метрики качества: регрессия

## Root Mean Absolute Error

- Корень из среднего квадратичного отклонения прогноза от исходного значения
- Сильнее штрафует за большие по модулю отклонения

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{j=1}^n (y_j - \hat{y}_j)^2}$$

## Mean Absolute Percentage Error

- Ошибка прогнозирования оценивается в процентах

Метрики  
качества:  
регрессия

$$M = \frac{100}{n} \sum_{t=1}^n \left| \frac{A_t - F_t}{A_t} \right|$$

## Метрики качества: регрессия

# Symmetric Mean Absolute Percentage Error

- Ошибка оценивается в процентах

$$\text{SMAPE} = \frac{100\%}{n} \sum_{t=1}^n \frac{|F_t - A_t|}{(|A_t| + |F_t|)/2}$$

## Метрики качества: регрессия

# Symmetric Mean Absolute Percentage Error

- Ошибка оценивается в процентах

$$\text{SMAPE} = \frac{100\%}{n} \sum_{t=1}^n \frac{|F_t - A_t|}{(|A_t| + |F_t|)/2}$$

$$\text{SMAPE} = \frac{100\%}{n} \sum_{t=1}^n \frac{|F_t - A_t|}{|A_t| + |F_t|}$$

## Метрики качества: регрессия

# Symmetric Mean Absolute Percentage Error

- По-разному штрафует за перепрогнозирование и недопрогнозирование
- Перепрогнозирование:  
 $A_t = 100, F_t = 110 \sim \text{SMAPE} = 4.76\%$
- Недопрогнозирование:  
 $A_t = 100, F_t = 90 \sim \text{SMAPE} = 5.26\%$

# Метрики качества в задачах ранжирования

# Ранжирование

Чем задача ранжирования отличается от задачи регрессии?

Метрики  
качества:  
ранжирование

Метрики  
качества:  
ранжирование

## Ранжирование

Чем задача ранжирования отличается от задачи регрессии?

Относительный порядок ответов модели интересует нас значительно больше, чем сами ответы модели.

## Метрики качества: ранжирование

# Ранжирование

Относительный порядок ответов модели интересует нас значительно больше, чем сами ответы модели.



Puma  
Ветровка  
3 490 руб.



Crocs  
Сланцы  
1 990 руб.



Tony-p  
Слипоны  
1 999 руб. 1 590 руб.



Champion  
Брюки спортивные  
3 599 руб. 1 970 руб.

Higher rank

Lower rank



# Метрики качества: ранжирование

Higher rank

Lower rank

# Ранжирование

The screenshot shows a search results page with the query "ranking problems" in the search bar. The results are categorized under "All". The first result is from Wikipedia about "Learning to rank - Wikipedia", which discusses ranking as a central part of information retrieval problems. The second result is from byjus.com about "Ranking-Topics, Rules, Problems and Solved Examples - Byju's", explaining ranking and order in banking question papers. The third result is from link.springer.com about "Classification Approach towards Ranking and Sorting Problems", discussing ranking as a multiclass classification problem. A red vertical arrow points downwards from the "Higher rank" label to the "Classification Approach" result, indicating its lower rank.

ranking problems

All Images Videos News Shopping More Settings Toc

About 589,000,000 results (0.52 seconds)

en.wikipedia.org › wiki › Learning\_to\_rank ▾

**Learning to rank - Wikipedia**

Ranking is a central part of many information retrieval **problems**, such as document retrieval, collaborative filtering, sentiment analysis, and online advertising. A possible architecture of a machine-learned search engine is shown in the accompanying figure.

[Applications](#) · [Feature vectors](#) · [Approaches](#) · [History](#)

byjus.com › Govt Exams › Logical Reasoning ▾

**Ranking-Topics, Rules, Problems and Solved Examples - Byju's**

Ranking and order is an important topic of banking question paper under logical reasoning section; it involves an arrangement of position or ranks of an object ...

link.springer.com › chapter

**Classification Approach towards Ranking and Sorting Problems**

As against standard approaches of treating **ranking** as a multiclass classification **problem**, in this paper we argue that **ranking/sorting problems** can be solved by ...

by S Rajaram · 2003 · Cited by 40 · Related articles

# Метрики качества: ранжирование

## Cumulative Gain

$$CG_p = \sum_{i=1}^p rel_i$$

кумулятивный выигрыш от ранжирования, где:

- рассматривается блок длиной  $p$
- $rel_i$  – оценка релевантности объекта на позиции  $i$

$rel$  зависит от задачи:

- бинарная функция (1 – релевантно, 0 - нет),
- числовая функция (стоимость товара, если он релевантен, 0 – если не релевантен)

Метрики  
качества:  
ранжирование

## Discounted Cumulative Gain (DCG)

Аналог CG, который позволяет **штрафовать** модель за то, что релевантные объекты находятся **дальше** от начала списка:

$$(1) DCG_p = \sum_{i=1}^p \frac{rel_i}{log_2(i + 1)}$$

$$(2) DCG_p = \sum_{i=1}^p \frac{2^{rel_i} - 1}{log_2(i + 1)}$$

## Метрики качества: ранжирование

### Normalized DCG

Нормализованная версия, которая позволяет:

- отнормировать оценку
- избавиться от влияния размера блока

$$IDCG_p = \sum_{i=1}^{|REL_p|} \frac{2^{rel_i} - 1}{\log_2(i + 1)}$$

$|REL_p|$  - список объектов, отранжированных по релевантности

$$nDCG_p = \frac{DCG_p}{IDCG_p}$$

## Normalized DCG (пример)

$i$	$rel_i$	$\log_2(i + 1)$	$rel_i/\log_2(i + 1)$
1	3	1	3
2	2	1.585	1.262
3	3	2	1.5
4	0	2.322	0
5	1	2.585	0.387
6	2	2.807	0.712

Метрики  
качества:  
ранжирование

$$DCG_6 = 6.861$$

$$IDCG_6 = 7.141$$

$$nDCG_6 = 0.961$$

# Precision@k

Какова точность модели ранжирования в среди топ-k результатов?

$$precision@k = \frac{tp@k}{tp@k + fp@k}$$

Метрики  
качества:  
ранжирование

# Recall@k

Какова полнота модели ранжирования в среди топ-k результатов?

$$recall@k = \frac{tp@k}{tp@k + fn@k}$$

Метрики  
качества:  
ранжирование

# Метрики качества: ранжирование

## Lift@k

Насколько ранжирование в топ-к результатах лучше, чем случайное?

$$lift@k = \frac{precision@k}{precision@all}$$

- при адекватном ранжировании метрика должна падать с ростом k
- однако для небольших k метрика будет нестабильной

# Метрики качества: ранжирование

## Кастомные метрики никто не отменял!

Учитывая особенности задачи, для которой строится модель ранжирования, имеет смысл разработать специализированную метрику:

1. Средняя позиция первого релевантного объекта
2. Доля блоков без релевантных объектов
3. Доля блоков без релевантных объектов в топ-3 и пр.

# Метрики качества: ранжирование

## Особые случаи: онлайн оценка алгоритмов ранжирования

Модели ранжирования сложно оценивать по историческим данным:

- релевантность может быть известна только для подмножества объектов
- модели ранжирования сложно сравнивать между собой (разная степень оцененности)
- нужно придумывать стратегии для оценки объектов, релевантность которых не известна

# Метрики качества

## Takeaways

1. Метрик качества много!
2. Важно выбрать метрику качества, подходящую не только математической, но и бизнес-задаче
3. Хорошо оценивать качество по нескольким метрикам
4. Важный вопрос для практического применения: как выбрать подходящую метрику?

# Машинное обучение: оценка качества в машинном обучении

Спасибо!  
Эмели Драль