# Машинное обучение:
## валидация моделей по историческим данным

МТС Тета
Эмели Драль

# Basics

1. ML basics & tools
2. **Валидация моделей по историческим данным**
3. Тестирование моделей в production

**Результат изучения:** знаете стандартные виды обучения, понимаете логику работы базовых алгоритмов, можете валидировать модели

# Валидация моделей по данным

1. Отложенная выборка и кросс-валидация
2. Метрики качества в задачах классификации, регрессии, ранжирования
3. Сложность и качество
4. Дополнительные свойства

# Валидация моделей

Валидация моделей

# 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) -> 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:

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

Валидация моделей

# Минимизация ошибок модели

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

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

Валидация моделей

# 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.

Валидация моделей

# 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 валиационной 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

## 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 тесте
7. При отсутствии существенных отличий в оценках на валидации и тесте считаем модель финальной
8. Можно перестроить модель на обучении + валидации

Валидация моделей

# 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?

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

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

# 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

Валидация моделей

# Кросс-валидация: k-fold

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


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

Валидация моделей

# Кросс-валидация: tk-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


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

Валидация моделей

# Стратегии кросс-валидация

Внутри k-fold возможны различные стратегии разбиения данных:

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

Альтернативная, но похожая стратегия:

- Bootstrap

Валидация моделей

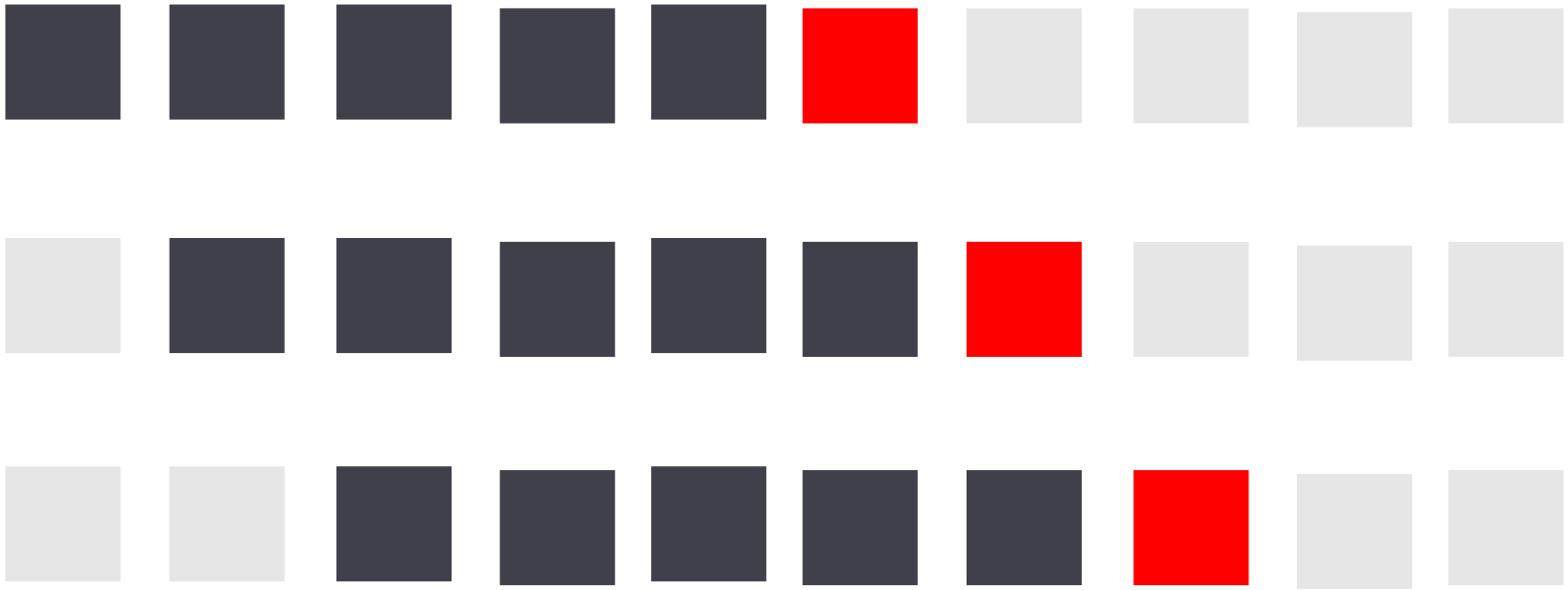
# Особые случаи: временные ряды

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

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 нетривиальный 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

Валидация моделей

# 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

# Метрики качества в задачах классификации

Метрики качества: классификация

# Метрики качества

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

Метрики качества: классификация

# Accuracy

Доля правильных ответов при классификации

Метрики качества: классификация

# Accuracy

Доля правильных ответов при классификации

target: 1 0 1 0 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

accuracy = 8/10 = 0.8

Метрики качества: классификация

# Метрики качества

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

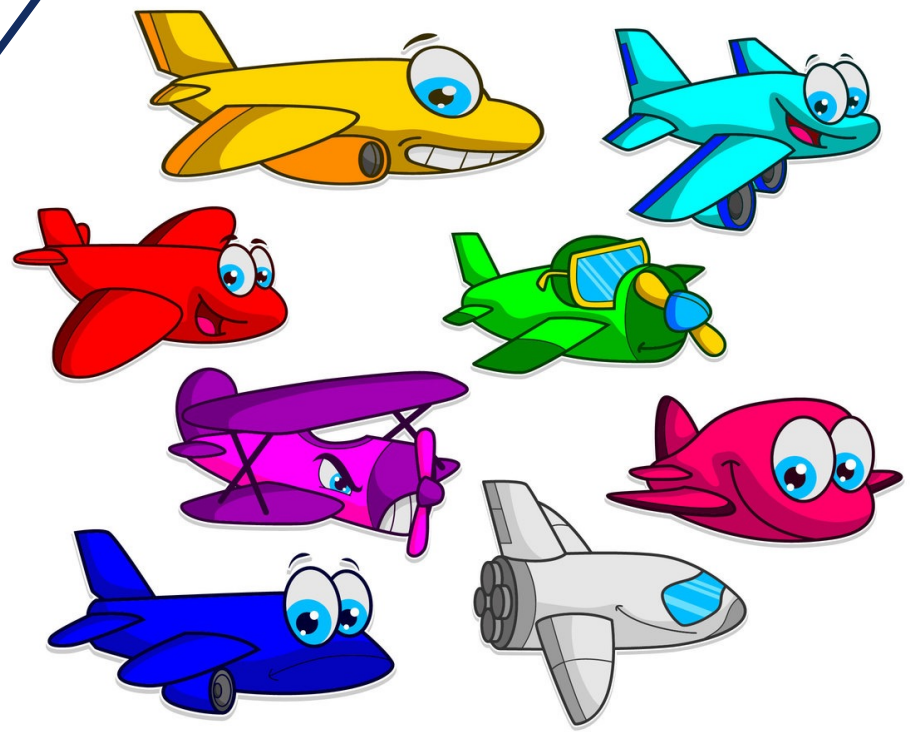
Метрики качества: классификация

# Precision & Recall

- Precision – точность
- Recall - полнота

# Метрики качества: классификация

## Сбитые самолёты



Источник изображения: https://www.freevector.com и https://www.angrybirds.com/

Метрики качества: классификация

# Precision

- Precision – точность выстрелов

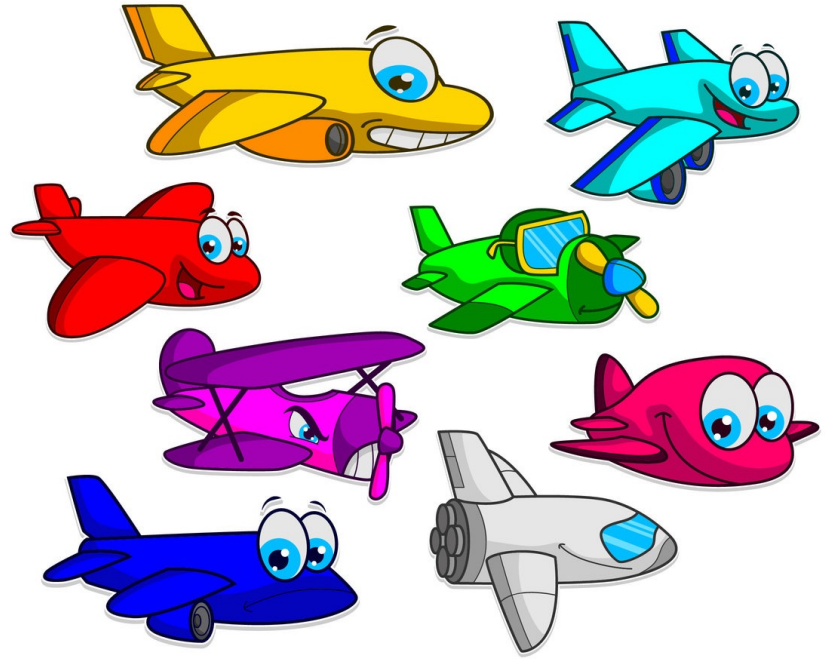
- Количество сбитых самолётов/количество выстрелов

Метрики качества: классификация

# Recall

- Recall – доля сбитых самолетов:

- Количество сбитых самолётов/общее количество самолётов



Метрики качества: классификация

# Считать вот так

| | | Actual Class: Yes | Actual Class: No |
|---|---|---|---|
| Predicted Class | Yes | TP | FP |
| | No | FN | TN |

Quality Metrics

Precision = TP / (TP + FP)

Recall = 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

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

Метрики качества: классификация

# Multiclass problem: macro-average

**Label 1**

| Predicted Class \ Actual Class | Yes | No |
|---|---|---|
| **Yes** | $TP_1$ | $FP_1$ |
| **No** | $FN_1$ | $TN_1$ |

$Precision_1 = TP_1/(TP_1 + FP_1)$

$Recall_1 = TP_1/(TP_1 + FN_1)$

**Label 2**

| Predicted Class \ Actual Class | Yes | No |
|---|---|---|
| **Yes** | $TP_2$ | $FP_2$ |
| **No** | $FN_2$ | $TN_2$ |

$Precision_2 = TP_2/(TP_2 + FP_2)$

$Recall_2 = TP_2/(TP_2 + FN_2)$

**Label 3**

| Predicted Class \ Actual Class | Yes | No |
|---|---|---|
| **Yes** | $TP_3$ | $FP_3$ |
| **No** | $FN_3$ | $TN_3$ |

$Precision_3 = TP_3/(TP_3 + FP_3)$

$Recall_3 = TP_3/(TP_3 + FN_3)$

Метрики качества: классификация

# Multiclass problem: macro-average

**Label 1**

| Predicted Class \ Actual Class | Yes | No |
|---|---|---|
| **Yes** | $TP_1$ | $FP_1$ |
| **No** | $FN_1$ | $TN_1$ |

**Label 2**

| Predicted Class \ Actual Class | Yes | No |
|---|---|---|
| **Yes** | $TP_2$ | $FP_2$ |
| **No** | $FN_2$ | $TN_2$ |

**Label 3**

| Predicted Class \ Actual Class | Yes | No |
|---|---|---|
| **Yes** | $TP_3$ | $FP_3$ |
| **No** | $FN_3$ | $TN_3$ |

$$Precision = \frac{Precision_1 + Precision_2 + Precision_3}{3}$$

$$Recall = \frac{Recall_1 + Recall_2 + Recall_3}{3}$$

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

Метрики качества: классификация

# Multiclass problem: micro-average

Actual Class

| Predicted Class | Label 1 | Label 2 | Label 3 |
|---|---|---|---|
| **Label 1** | $TP_1$ | $Err_{1->2}$ | $Err_{1->3}$ |
| **Label 2** | $Err_{2->1}$ | $TP_2$ | $Err_{2->3}$ |
| **Label 3** | $Err_{3->1}$ | $Err_{3->2}$ | $TP_3$ |

Multiclass errors:

$$FP_1 = Err_{1->2} + Err_{1->3}$$

$$FP_2 = Err_{2->1} + Err_{2->3}$$

$$FP_3 = Err_{3->1} + Err_{3->2}$$

$$FN_1 = Err_{2->1} + Err_{3->1}$$

$$FN_2 = Err_{1->2} + Err_{3->2}$$

$$FN_3 = Err_{1->3} + Err_{2->3}$$

Метрики качества: классификация

# Multiclass problem: micro-average

Actual Class

| Predicted Class | Label 1 | Label 2 | Label 3 |
|---|---|---|---|
| **Label 1** | $TP_1$ | $Err_{1->2}$ | $Err_{1->3}$ |
| **Label 2** | $Err_{2->1}$ | $TP_2$ | $Err_{2->3}$ |
| **Label 3** | $Err_{3->1}$ | $Err_{3->2}$ | $TP_3$ |

Micro-average:

$$Precision = \frac{TP_1 + TP_2 + TP_3}{TP_1 + TP_2 + TP_3 + FP_1 + FP_2 + FP_3}$$

$$Recall = \frac{TP_1 + TP_2 + TP_3}{TP_1 + TP_2 + TP_3 + FN_1 + FN_2 + FN_3}$$

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

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 | Actual Class: No |
|---|---|---|---|
| Predicted Class | Yes | TP | FP |
| | No | FN | TN |

Как считать:

1. Select Step Size
2. For each step calculate:
   - TRP = TP / (TP + FN)
   - FPR = FP / (FP + TN)
3. Plot the curve in TPR & FPR axes

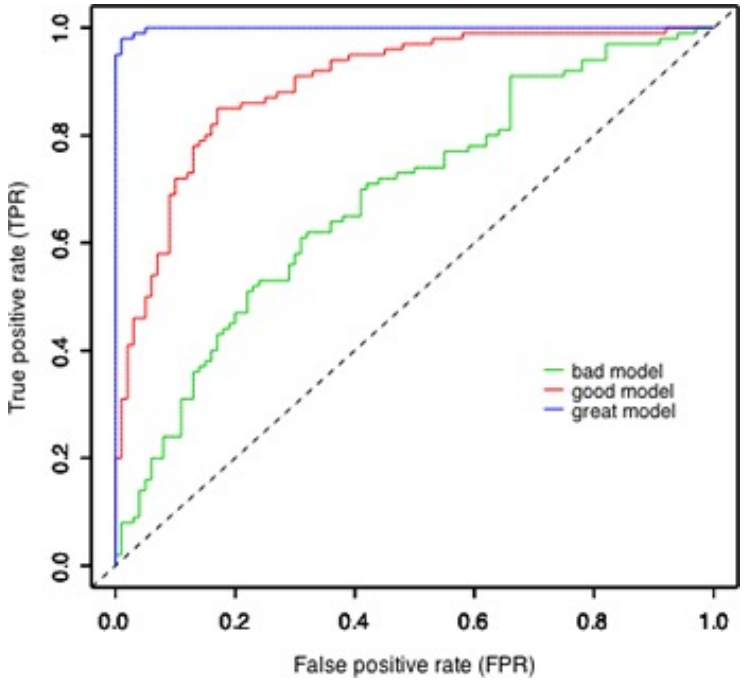
Метрики качества: классификация

# ROC curve

| | | Actual Class: Yes | Actual Class: No |
|---|---|---|---|
| Predicted Class | **Yes** | TP | FP |
| | **No** | FN | TN |

Как считать:

1. Select Step Size
2. For each step calculate:
   - TRP = TP / (TP + FN)
   - FPR = FP / (FP + TN)
3. Plot the curve in TPR & FPR axes



Метрики качества: классификация

# ROC curve

Как оценить кривую численно?

Метрики качества: классификация

# ROC curve

Как оценить кривую численно?

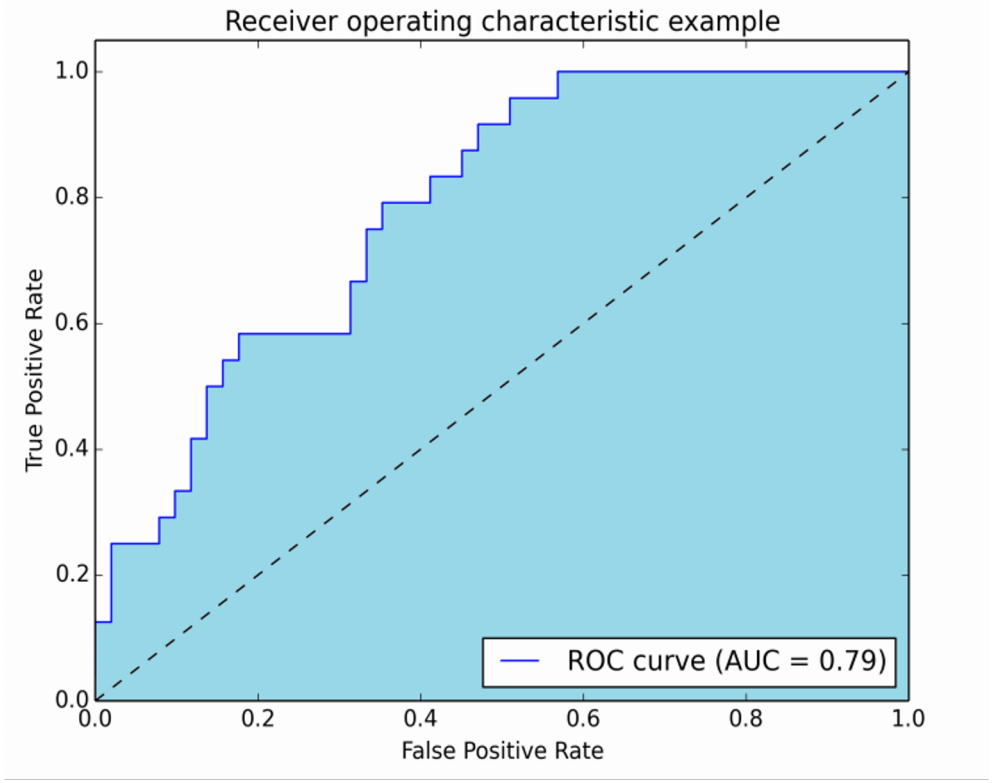
Измерить площадь под кривой – area under the curve!

Метрики качества: классификация

# ROC curve

Как оценить кривую численно?

Измерить площадь под кривой – area under the curve!



Метрики качества: классификация

# ROC curve

Что если классификация всё же не вероятностная?

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

Метрики качества: классификация

# Log loss

Логарифмическая ошибка

Хорошо оценивает вероятность

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

# 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

Метрики качества: регрессия

# Mean Absolute Error

- Отклонение прогноза от исходного значения
- Усредненное по всем наблюдениям

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

# Метрики качества: регрессия

## Root Mean Absolute Error

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

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Метрики качества: регрессия

# Mean Absolute Percentage Error

- Ошибка прогнозирования оценивается в процентах

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

Метрики качества: регрессия

# Symmetric Mean Absolute Percentage Error

- Ошибка оценивается в процентах
- Делается нормировка не только на факт, но и на прогноз

$$SMAPE = \frac{100\%}{n} \sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{(|y_i| + |\hat{y}_i|)/2}$$

Метрики качества: регрессия

# Symmetric Mean Absolute Percentage Error

Встречается 2 варианта расчета:

$$SMAPE = \frac{100\%}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{(|y_i| + |\hat{y}_i|)/2}$$

диапазон: 0 – 100%

$$SMAPE = \frac{100\%}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{(|y_i| + |\hat{y}_i|)}$$

диапазон: 0 – 200%

Метрики качества: регрессия

# Symmetric Mean Absolute Percentage Error

- По-разному штрафует за перепрогнозирование и недопрогнозирование

- Перепрогнозирование:
  $A_t = 100$, $F_t = 110$ ~ SMAPE = 4.76%

- Недопрогнозирование:
  $A_t = 100$, $F_t = 90$ ~ SMAPE = 5.26%

# Метрики качества в задачах ранжирования

Метрики качества: ранжирование

# Ранжирование

Чем задача ранжирования отличается от задачи регрессии?

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.

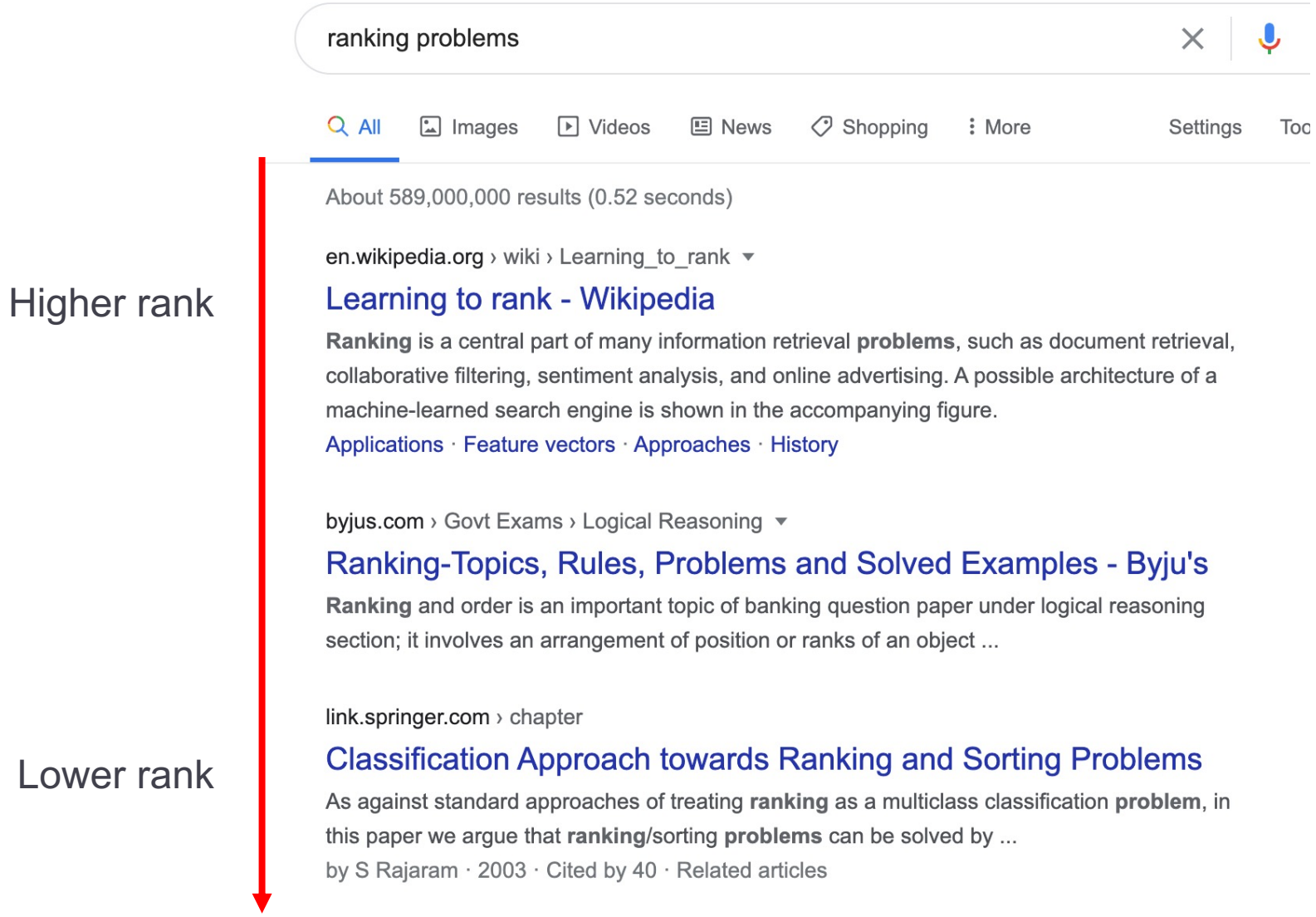
Tony-p
Слипоны
~~1 999 руб.~~ 1 590 руб.

Champion
Брюки спортивные
~~3 599 руб.~~ 1 970 руб.

Higher rank → Lower rank

# Метрики качества: ранжирование

# Ранжирование



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 релевантент, 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

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

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

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

Метрики качества: ранжирование

# 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 |

$$DCG_6 = 6.861$$

$$IDCG_6 = 7.141$$

$$nDCG_6 = 0.961$$

Метрики качества: ранжирование

# Precision@k

Какова точность модели ранжирования среди топ-k результатов?

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

Метрики качества: ранжирование

# Recall@k

Какова полнота модели ранжирования среди топ-k результатов?

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

# Метрики качества в прикладных задачах

Метрики качества

# Lift@k (телекоммуникации)

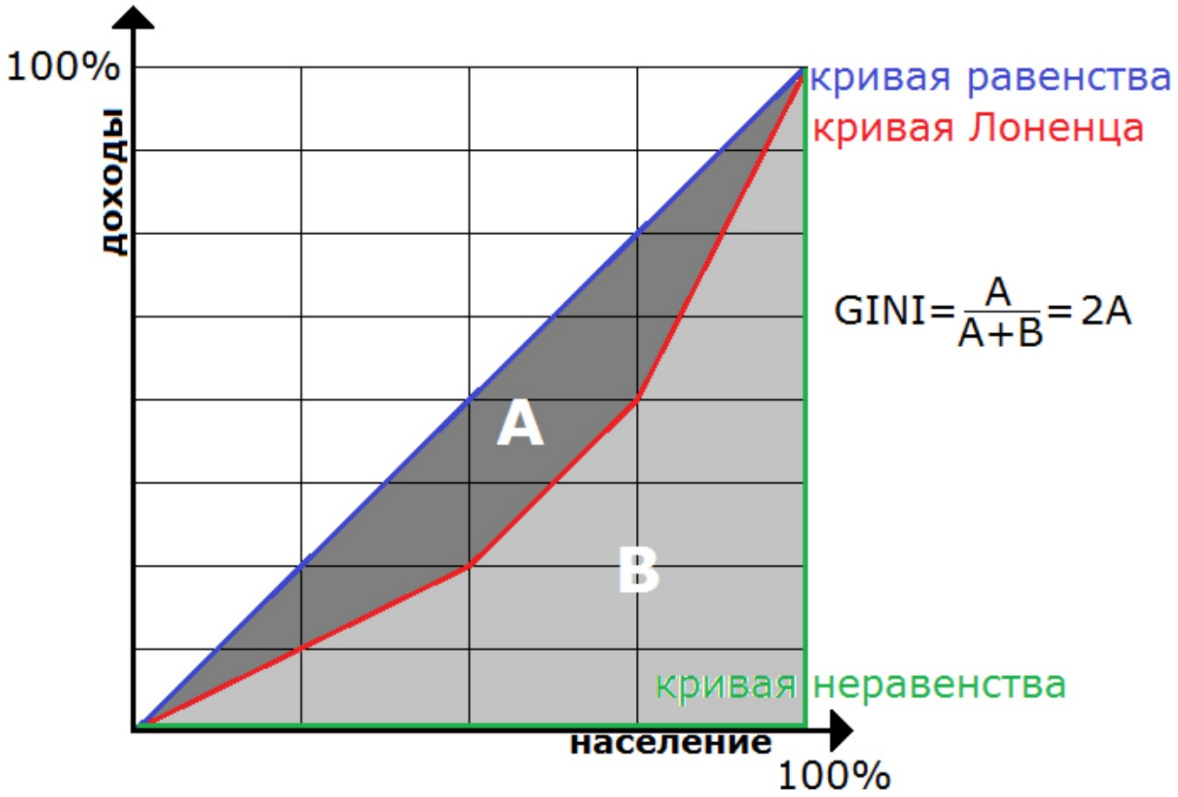
Насколько ранжирование в топ-к результатах лучше, чем случайное?

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

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

Метрики качества

# Gini (финансы)




Кривая Лоренца: доля всеобщего дохода на долю населения

$Gini_{model}$ = 2*ROCAUC - 1 $Gini_{index} = Gini_{model}/Gini_{ideal}$

Анализ малых данных:
https://dyakonov.org/2015/12/15/%D0%B7%D0%BD%D0%B0%D0%BA%D0%BE%D0%BC%D1%8C%D1%82%D0%B5%D1%81%D1%8C-%D0%B4%D0%B6%D0%B8%D0%BD%D0%B8/

Метрики качества

# Чувствительность и специфичность (медицина)

Клиническая чувствительность и специфичность

Чувствительность = число больных, выявленных тестом/истинное число больных

Специфичность  = число здоровых, выявленных тестом/истинное число здоровых

| | | Actual Class: Yes | Actual Class: No |
|---|---|---|---|
| Predicted Class | Yes | TP | FP |
| Predicted Class | No | FN | TN |

Как считать?

Чувствительность (TPR) = TP / (TP + FP) = TP/P

Специфичность (TNR)  = TN / (FP + NN) = TN/N

Метрики качества

# 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

# Особые случаи: офлайн 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

# Машинное обучение:
## валидация моделей по историческим данным

Спасибо!

Эмели Драль