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

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 -> Y
- a(x) = y
- A – семейство моделей

Оценка качества

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

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:

- Random shuffle
- Bootstrap

Валидация моделей

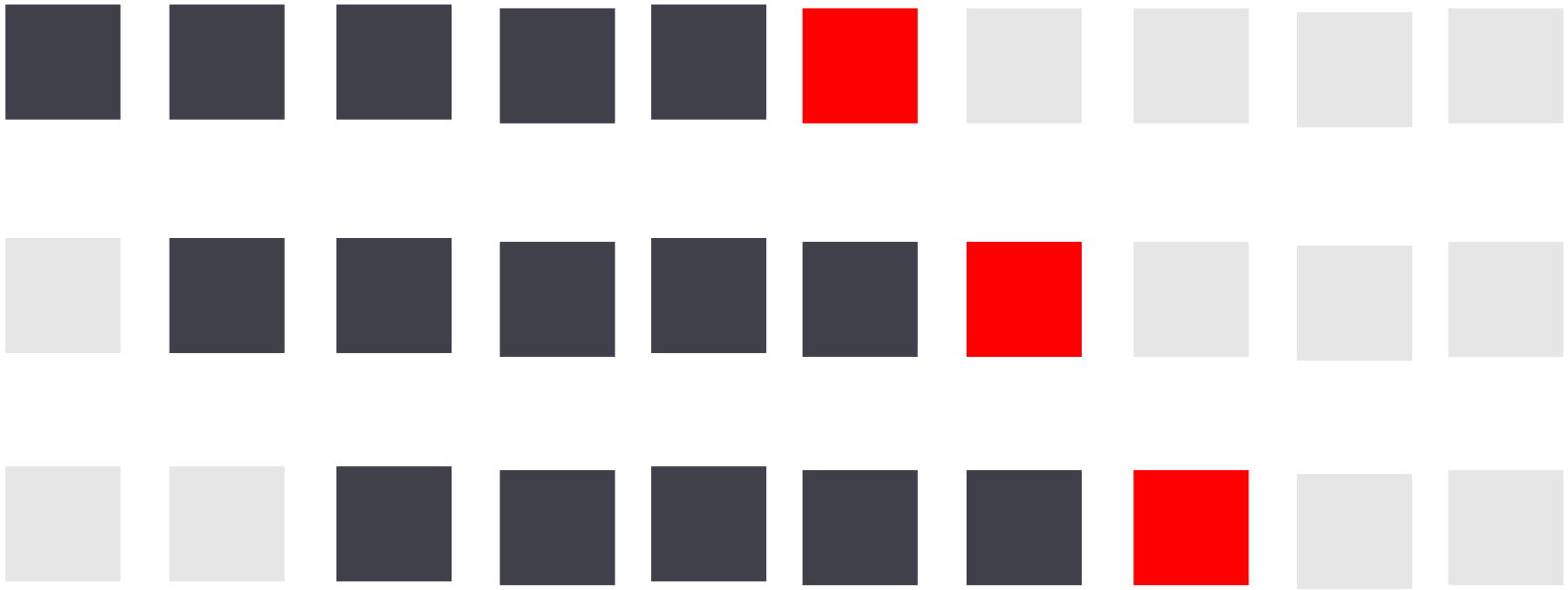
Особые случаи: временные ряды

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

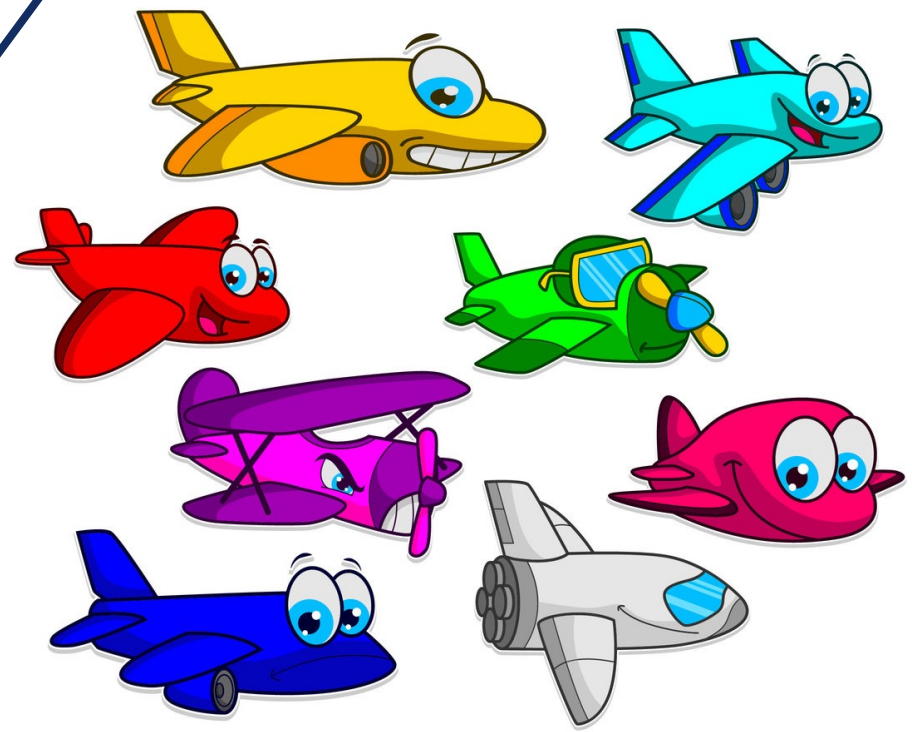
Метрики качества: классификация

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

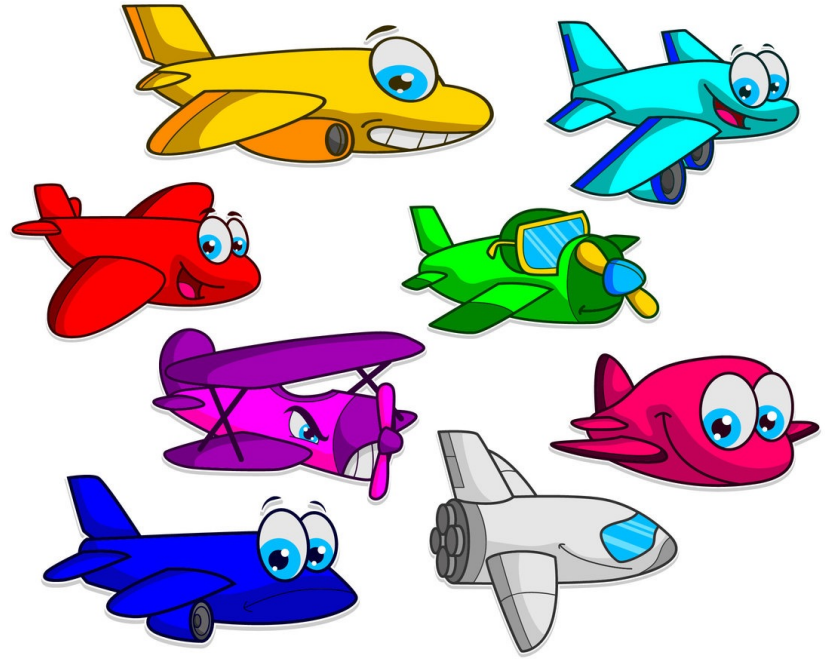
- Количество сбитых самолётов/количество выстрелов

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

Predicted Class \ Actual 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
Predicted Class	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

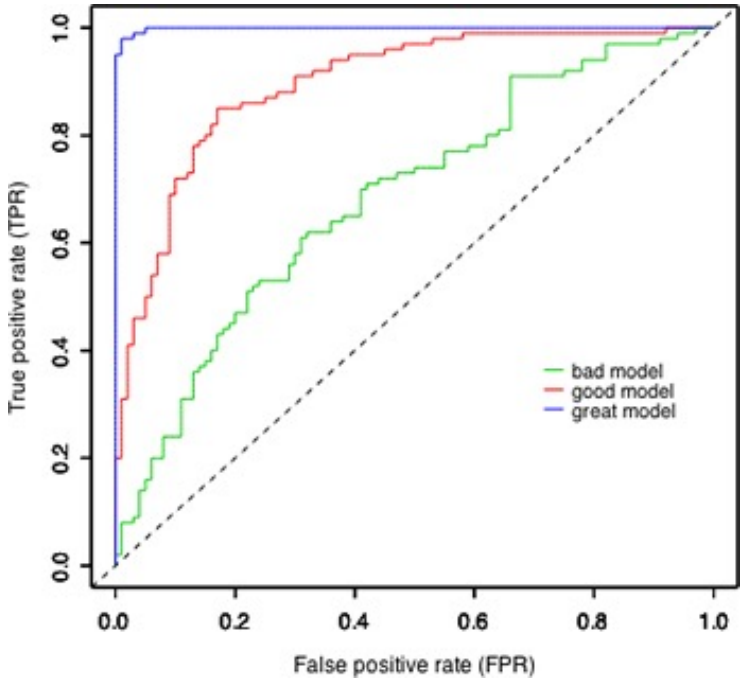
Метрики качества: классификация

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?

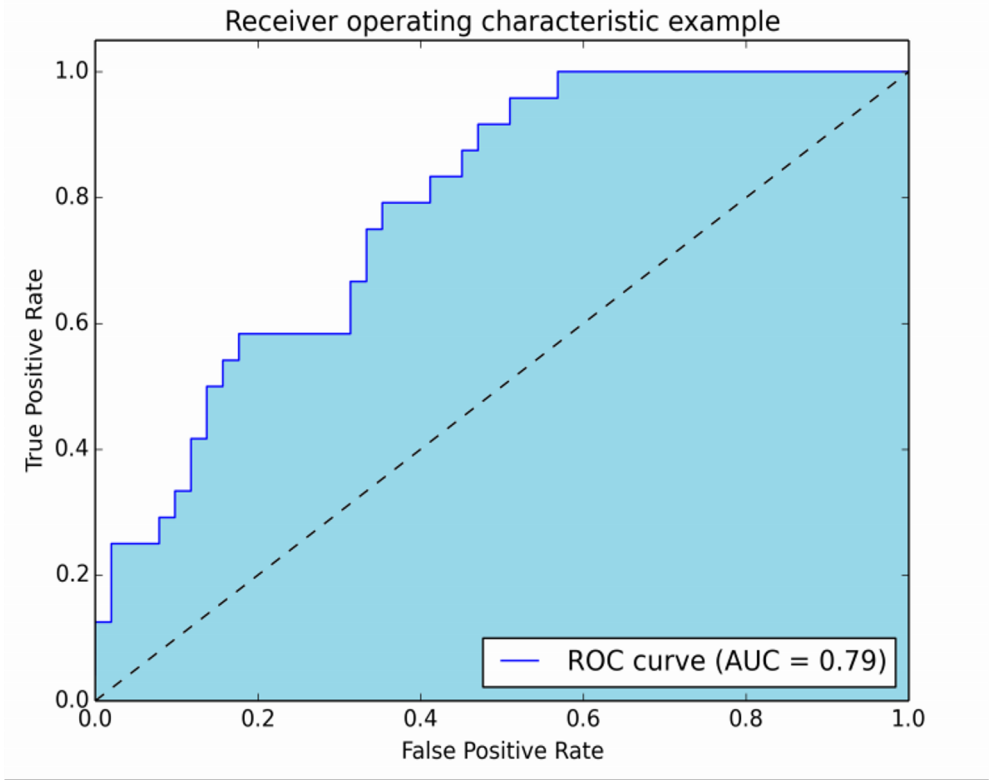
Измерить площадь под кривой – area under the curve!

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.

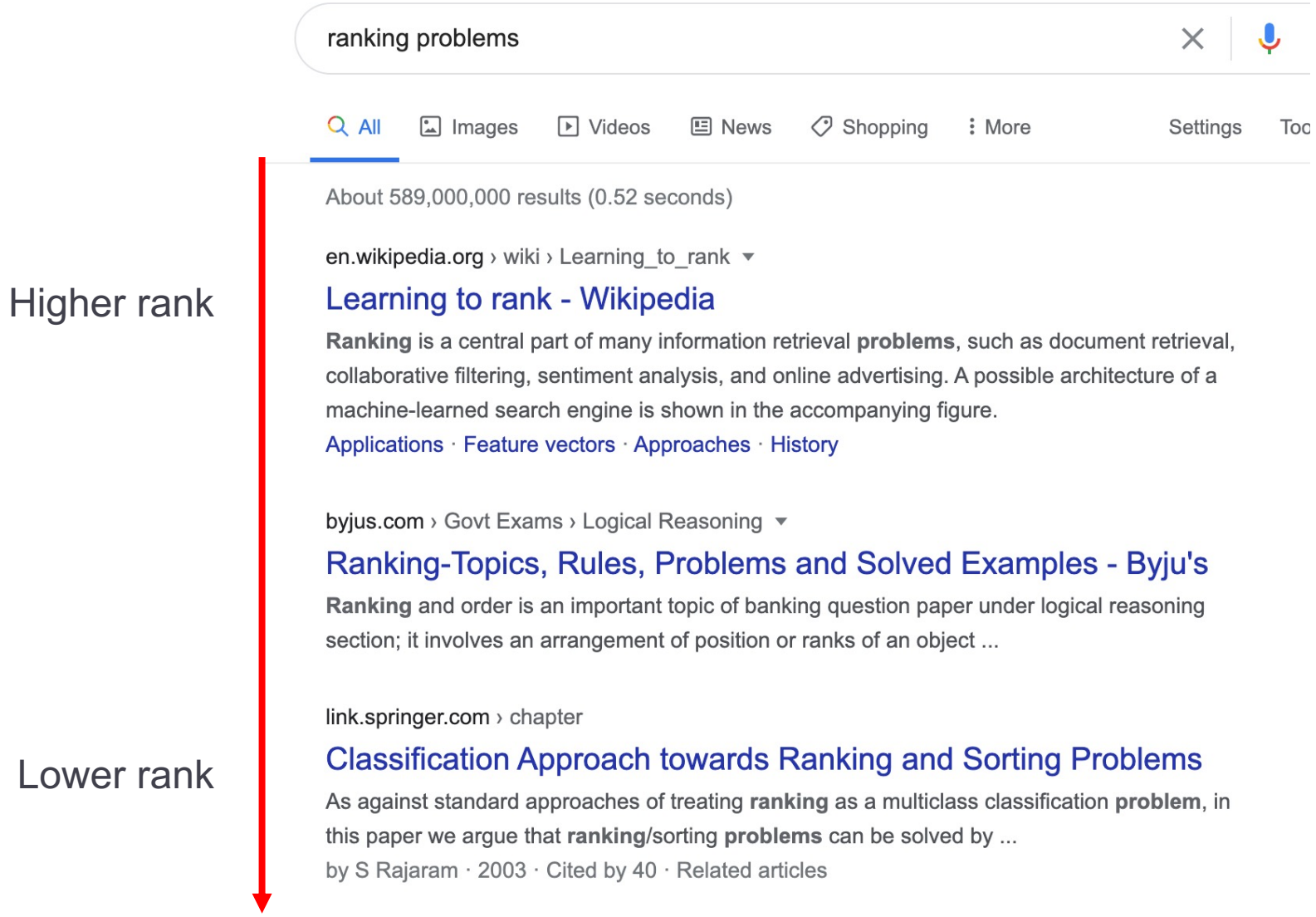
Tony-p
Слипоны
~~1 999 руб.~~ 1 590 руб.

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: 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: 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

Качество vs Сложность

Качество
vs
Сложность

Quality vs Complexity

Quality: чем меньше ошибка модели, тем лучше

Complexity: чем проще модель, тем стабильнее она работает

Качество
vs
Сложность

Quality vs Complexity

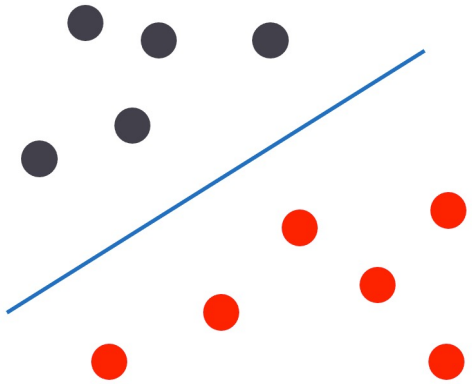
Quality: чем меньше ошибка модели, тем лучше

Complexity: чем проще модель, тем стабильнее она работает

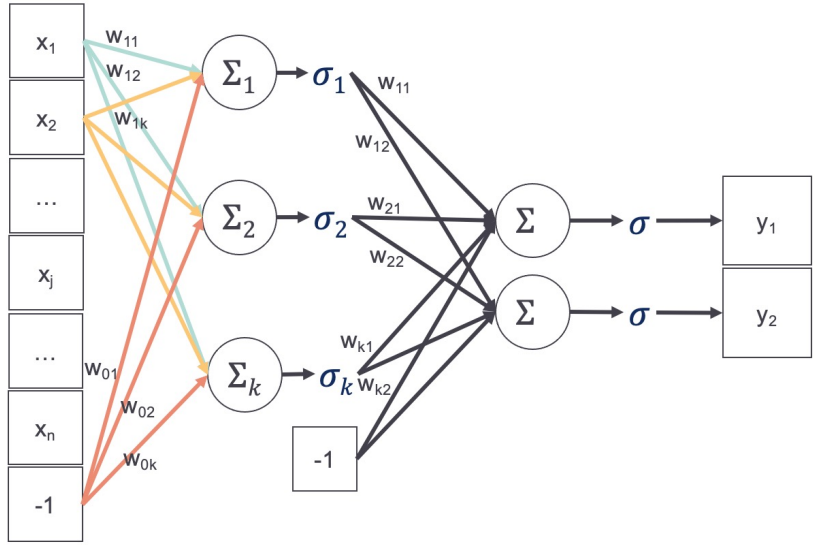
Часто, более сложные модели (или комбинации моделей) дают меньшую ошибку, на для использования в сервисе выбирают ближайший по качеству более простой аналог

Качество
vs
Сложность

Quality vs Complexity

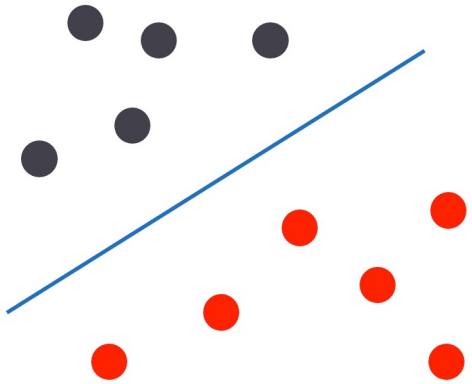


vs

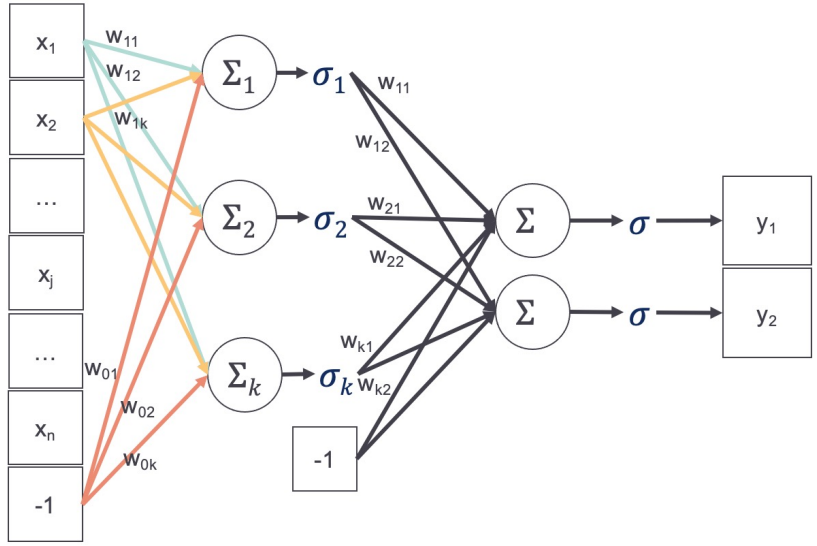


Качество
vs
Сложность

Quality vs Complexity



vs

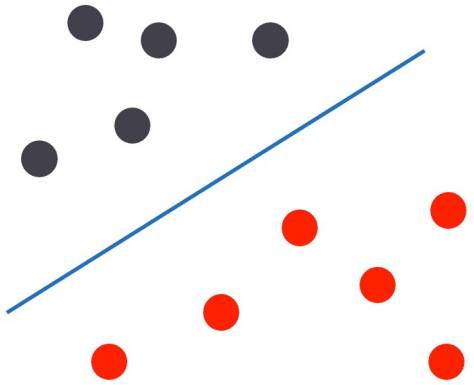


ROC AUC = 0,74

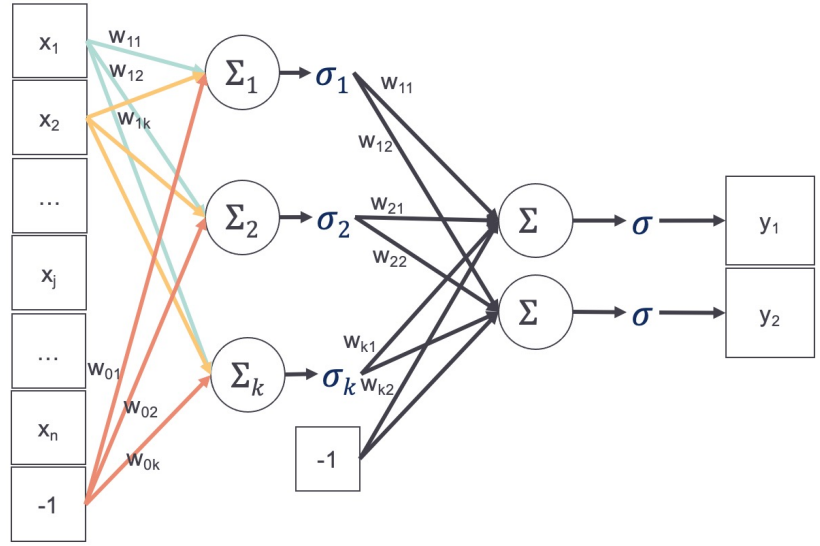
ROC AUC = 0,79

Качество
vs
Сложность

Quality vs Complexity



vs



ROC AUC = 0,74

ROC AUC = 0,79

- Связь качества модели и экономического эффекта: сколько нам стоит 0.05 ROC AUC?
- Готовы ли мы ради этого эффекта усложнить архитектуру для поддержки нейронных сетей?

Качество
vs
Сложность

Quality vs Complexity

Важно иметь несколько моделей-кандидатов разной сложности и понимать, какой прирост в качестве и эффекте дает усложнение модели

Качество
vs
Сложность

Модели-кандидаты

Важно иметь несколько моделей-кандидатов разной сложности и понимать, какой прирост в качестве и эффекте дает усложнение модели

Пример:

1. Constant model
2. Simple model with numeric features only
3. Complex model with numeric features only
4. Simple model with some feature engineering
5. Complex model with some feature engineering
6. Hybrid model

Качество
vs
Сложность

Constant model

1. Самый популярный класс в задаче классификации
2. Среднее или медиана (посчитанные по обучающей выборке!) в задаче регрессии
3. Last value (можно с учетом сезонности) в задаче прогнозирования
4. Most popular items для рекомендательной системы

Качество
vs
Сложность

Constant model

1. Самый популярный класс в задаче классификации
2. Среднее или медиана (посчитанные по обучающей выборке!) в задаче регрессии
3. Last value (можно с учетом сезонности) в задаче прогнозирования
4. Most popular items для рекомендательной системы

- для каждой задачи можно подобрать условно оптимальную константу
- это важный benchmark, позволяющий понять ценность решения

Качество
vs
Сложность

Constant model

В некоторых индустриях метрики качества даже устроены таким образом, чтобы оценивать относительный прирост качества модели.

Пример: задача прогнозирования оттока в телеком.

Метрика lift@k - во сколько раз ранжирование среди top k% абонентов согласно модели лучше случайного ранжирования?

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

Качество
vs
Сложность

Simple model

1. Регрессия по одному или нескольким признакам
2. Дерево решений небольшой глубины
3. Метод ближайших соседей по нескольким признакам
4. Rule-based (часто, это текущее production решение)

Качество
vs
Сложность

Модель другого типа

Часто, текущее production решение не является моделью машинного обучения

1. Rule-based system
2. Математическая модель (аналитическая формула)
3. Физическая модель

Их не вполне справедливо считать простыми, но это также хороший benchmark

Качество
vs
Сложность

Complex model

Дальнейшее снижение ошибки модели возможно за счет:

- feature engineering
- более сложный алгоритм с большим количеством параметров
- комбинации более сложного алгоритма и feature engineering

Качество
vs
Сложность

Complex model

Дальнейшее снижение ошибки модели возможно за счет:

- feature engineering
- более сложный алгоритм с большим количеством параметров
- комбинации более сложного алгоритма и feature engineering

Полезно проанализировать остатки модели, чтобы оценить наличие оставшегося сигнала в данных;

Имеем смысл смотреть на feature importance добавленных признаков, особенно если их сложно рассчитывать

Качество
vs
Сложность

Hybrid model

Альтернативный способ снижения ошибки – использование комбинации из нескольких подходов к решению задачи.

Подходов очень много, например:

- Стандартный stacking
- Content based + collaborative filtering recommender system
- Бинарная классификация + регрессия для одного из классов
- Физико-химическая модель + ml модель
- Термодинамическая модель + ml модель

и пр.

Качество
vs
Сложность

Quality vs Complexity

Модель	Precision@10% (cv mean)
Constant model	0.08
Physical model	0.71
Linear model (num features)	0.61
GB (feature engineering)	0.76
Physical model + GB on residuals	0.82
Ideal model	0.9

Полезно дать рекомендацию о том, какую модель вы считаете оптимальной.

Качество
vs
Сложность

Quality vs Complexity

Модель	Precision@10% (cv mean)
Constant model	0.08
Physical model	**0.71**
Linear model (num features)	0.61
GB (feature engineering)	0.76
Physical model + GB on residuals	**0.82**
Ideal model	0.9

Полезно дать рекомендацию о том, какую модель вы считаете оптимальной.

Дополнительные свойства

Валидация модели

Что нужно оценить?

- Качество модели

Дополнительные свойства:

- Экономический эффект
- Скорость устаревания модели
- Bias & fairness
- Интерпретация

Валидация модели

Качество модели

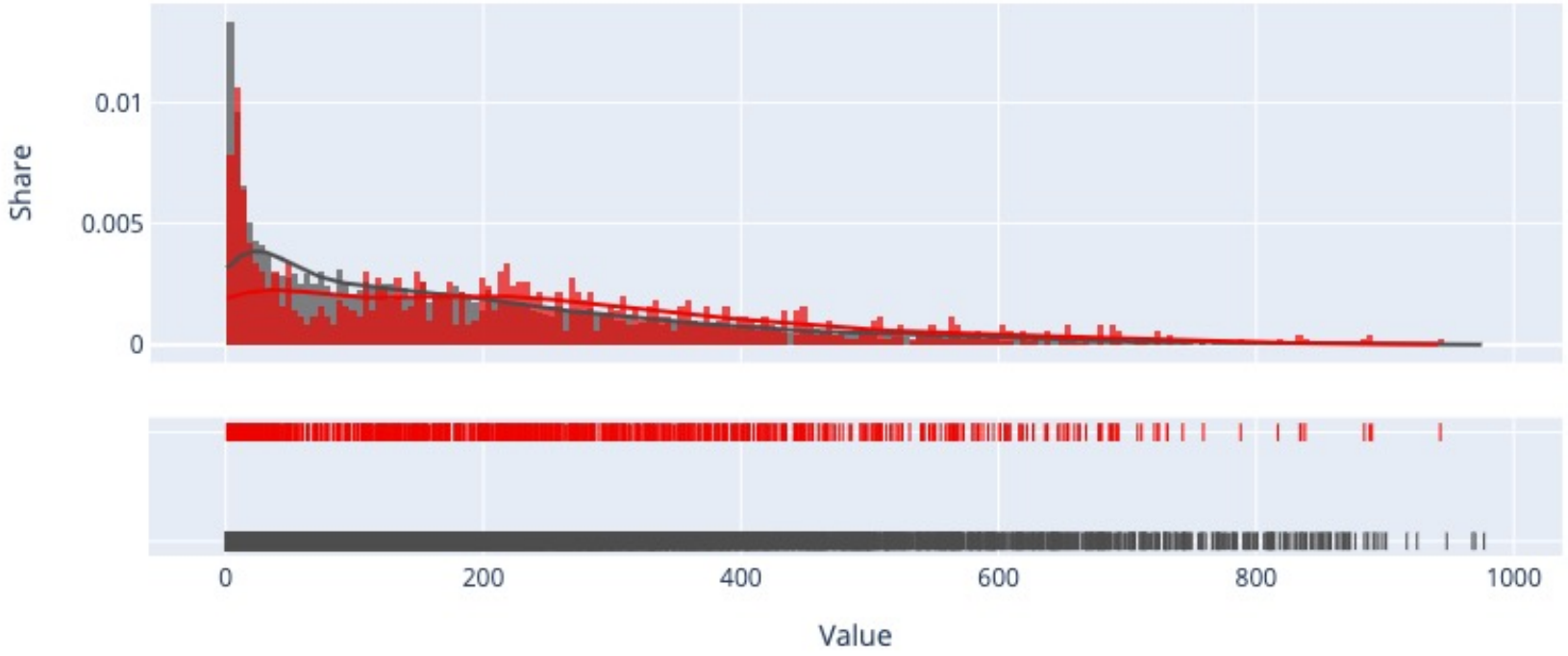
- Можно смотреть на несколько метрик, хотя оптимизируем всегда одну
- Интервальные оценки лучше точечных
- Cross-validation + hold-out test

Также, с помощью cross-validation можно оценить стабильность модели:

- меняется ли качество от фолда к фолду?
- меняется ли feature importance от фолда к фолду?

Валидация модели

Качество модели

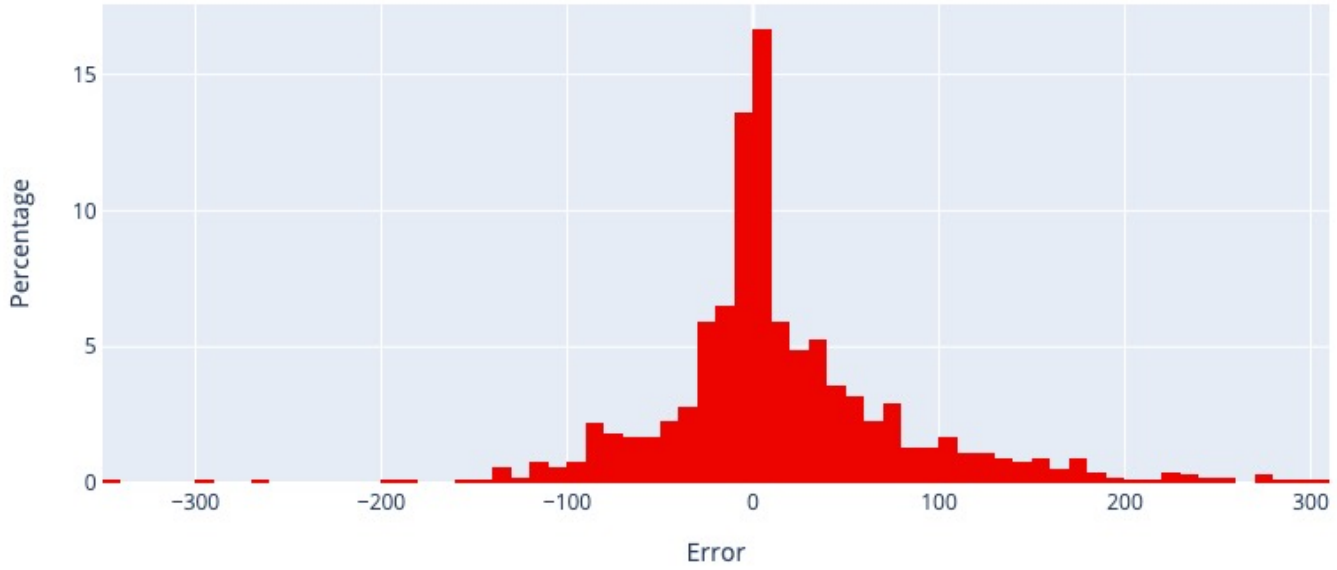


Дополнительно, имеет смысл сравнить:

- распределение target на обучении и отложенной выборке
- распределение model output на обучении и отложенной выборке

Валидация модели

Качество модели



Распределение ошибок поможет понять:

- склонна ли модель к недо/переоценке целевой функции
- остался ли сигнал в данных
- есть ли выбросы или сегменты с большей ошибкой

Валидация модели

Что ещё нужно оценить?

- Скорость устаревания модели
- Bias & fairness
- Интерпретация

Валидация модели

Скорость устаревания

Важная характеристика, на основе которой можно сделать вывод о необходимой частоте переобучения модели

Подход к оценке: (обучение, ошибка внутри ожидаемого интервала, ошибка за пределами интервала)

Быстрое устаревание:

Модель не устаревает:

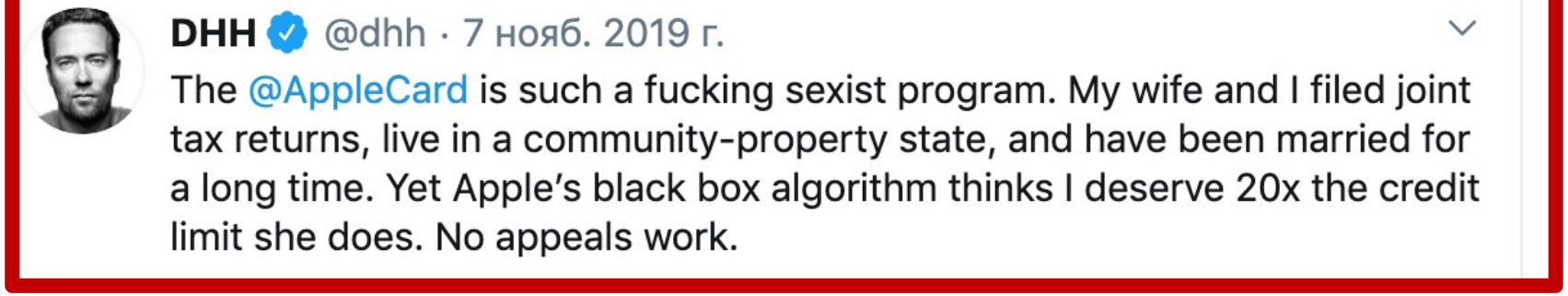
Среднее устаревание:

Валидация модели

Валидация модели и предотвращение ошибок

Валидация модели

Избежание предвзятости



Amazon scraps secret AI recruiting tool that showed bias against women

In effect, Amazon's system taught itself that male candidates were preferable. It penalized resumes that included the word "women's," as in "women's chess club captain." And it downgraded graduates of two all-women's colleges, according to people familiar with the matter. They did not specify the names of the schools.

https://www.technologyreview.com/2019/10/25/132184/a-biased-medical-algorithm-favored-white-people-for-healthcare-programs/

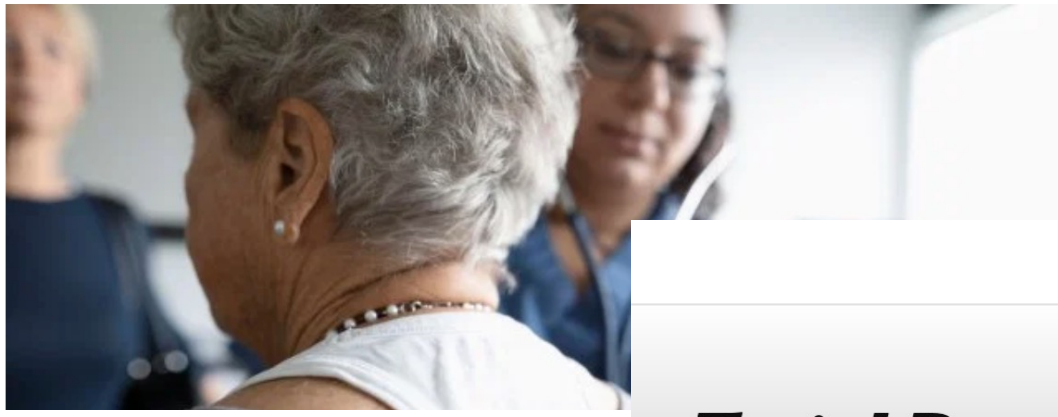
Валидация модели

Избежание предвзятости

MIT Technology Review

Artificial intelligence Oct 25

A biased medical algorithm favored white people for health-care programs



The New York Times

Facial Recognition Is Accurate, if You're a White Guy

https://www.technologyreview.com/2019/10/25/132184/a-biased-medical-algorithm-favored-white-people-for-healthcare-programs/
https://www.nytimes.com/2018/02/09/technology/facial-recognition-race-artificial-intelligence.html

Валидация модели

Доверие

More CEOs (84%) ‘agree’ that AI-based decisions need to be explainable than that AI is good for society (79%).

Mark J. Girouard, an employment attorney at Nilan Johnson Lewis, says one of his clients was vetting a company selling a resume screening tool, but didn’t want to make the decision until they knew what the algorithm was prioritizing in a person’s CV.

After an audit of the algorithm, the resume screening company found that the algorithm found two factors to be most indicative of job performance: their name was Jared, and whether they played high school lacrosse. Girouard’s client did not use the tool.

https://qz.com/1427621/companies-are-on-the-hook-if-their-hiring-algorithms-are-biased
https://hackernoon.com/dogs-wolves-data-science-and-why-machines-must-learn-like-humans-do-41c43bc7f982
https://www.pwc.com/mu/pwc-22nd-annual-global-ceo-survey-mu.pdf

Валидация модели

Регуляторные требования

Европа – GDPR

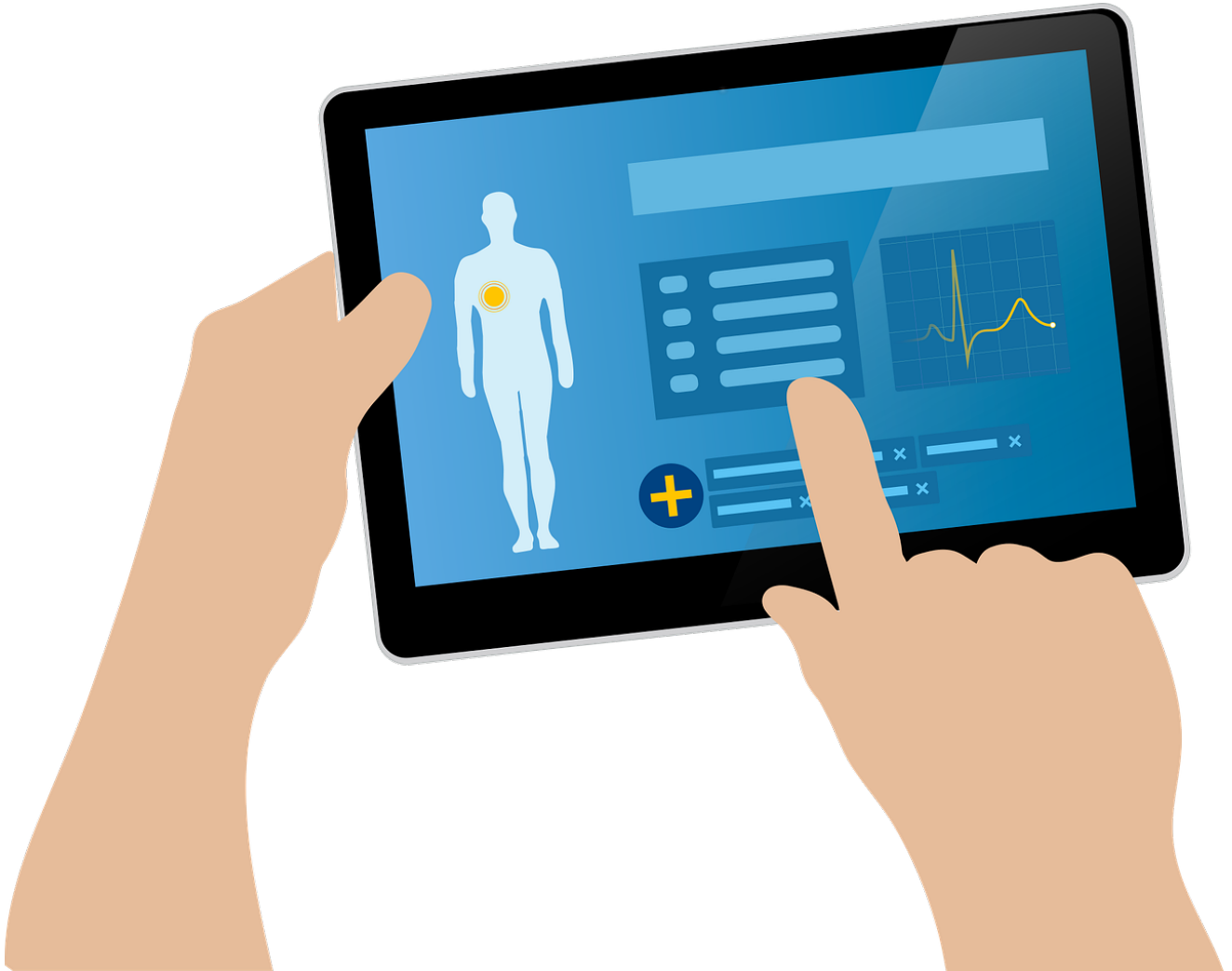
The right to access meaningful information about the logic involved, as well as the significance and the envisaged consequences of automated decision-making”

США - Equal Credit Opportunity Act

Statement of reasons for adverse action, must be specific and indicate the principal reason(s) for the adverse action

Валидация модели

Интерпретация модели



Валидация модели

Дополнительные свойства

Такие характеристики модели, как:

- калибровка
- качество в топе прогнозов
- ошибка в разрезе выбранных сегментов
- линейность по выбранным признакам

и пр.

Машинное обучение:
валидация моделей по историческим данным

Спасибо!

Эмели Драль