НЕТОЛОГИЯ
групп

ЗАНЯТИЕ 2.1

ДЕРЕВЬЯ РЕШЕНИЙ. КЛАССИФИКАЦИЯ

Артур
Сапрыкин

ЦЕЛИ ЗАНЯТИЯ

В КОНЦЕ ЗАНЯТИЯ ВЫ НАУЧИТЕСЬ:

-применять **алгоритм** классификации, принятие решений которого **можно проинтерпретировать**

- **измерять качество** решений в задачах классификации

- **оценивать важность** фичей

-понимать **основу продвинутых алгоритмов**, таких как RandomForest, XGBoost, LGBM, etc..

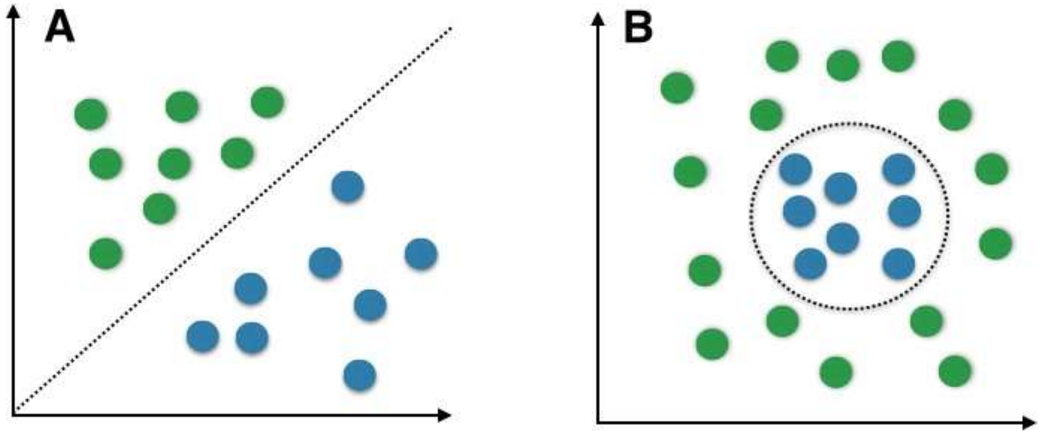
О ЧЁМ ПОГОВОРИМ И ЧТО СДЕЛАЕМ

1. Задача классификации: постановка и примеры
2. Дерево решений: как его построить?
   Обзор sklearn.tree.DecisionTreeClassifier
3. Достоинства и недостатки деревьев решений.
4. Визуализируем принятие решений и предсказания алгоритма; примем участие в соревновании Kaggle
5. Метрики качества в задачах классификации
6. Оценим решение Kaggle; классифицируем статьи Ведомостей: политика или финансы?

1. ЗАДАЧА КЛАССИФИКАЦИИ

ТИПЫ ЗАДАЧ

* **классификация**
* ранжирование
* регрессия
* кластеризация



ПРИМЕРЫ ЗАДАЧ КЛАССИФИКАЦИИ

Скоринг. Вернёт ли клиент кредит? (banking, insurance)

Отток. Перестанет ли пользоваться клиент услугами компании? Перестанет ли, если дать ему бонус? (marketing)

Intent recognition. О чём говорит пользователь в своём обращении? (может быть несколько intent’ов, может быть древовидная структура) (API.AI)

Image recognition. Что на картинке? (Google, FindFace)

ПОСТАНОВКА ЗАДАЧИ

Задача восстановления зависимости $y: X \to Y$, $|Y| < \infty$
по точкам *обучающей выборки* (x_i, y_i), $i = 1, \ldots, \ell$.

Дано: векторы $x_i = (x_i^1, \ldots, x_i^n)$ — объекты обучающей выборки,
$y_i = y(x_i)$ — классификации, ответы учителя, $i = 1, \ldots, \ell$:

$$\begin{pmatrix} x_1^1 & \ldots & x_1^n \\ \ldots & \ldots & \ldots \\ x_\ell^1 & \ldots & x_\ell^n \end{pmatrix} \xrightarrow{y^*} \begin{pmatrix} y_1 \\ \ldots \\ y_\ell \end{pmatrix}$$

Найти: функцию $a(x)$, способную классифицировать объекты
произвольной *тестовой выборки* $\tilde{x}_i = (\tilde{x}_i^1, \ldots, \tilde{x}_i^n)$, $i = 1, \ldots, k$:

$$\begin{pmatrix} \tilde{x}_1^1 & \ldots & \tilde{x}_1^n \\ \ldots & \ldots & \ldots \\ \tilde{x}_k^1 & \ldots & \tilde{x}_k^n \end{pmatrix} \xrightarrow{a?} \begin{pmatrix} a(\tilde{x}_1) \\ \ldots \\ a(\tilde{x}_k) \end{pmatrix}$$

** из материалов к курсу Воронцова на Coursera*

2. ПОСТРОЕНИЕ ДЕРЕВА РЕШЕНИЙ

ЦВЕТКИ ИРИСА: ЗАДАЧА

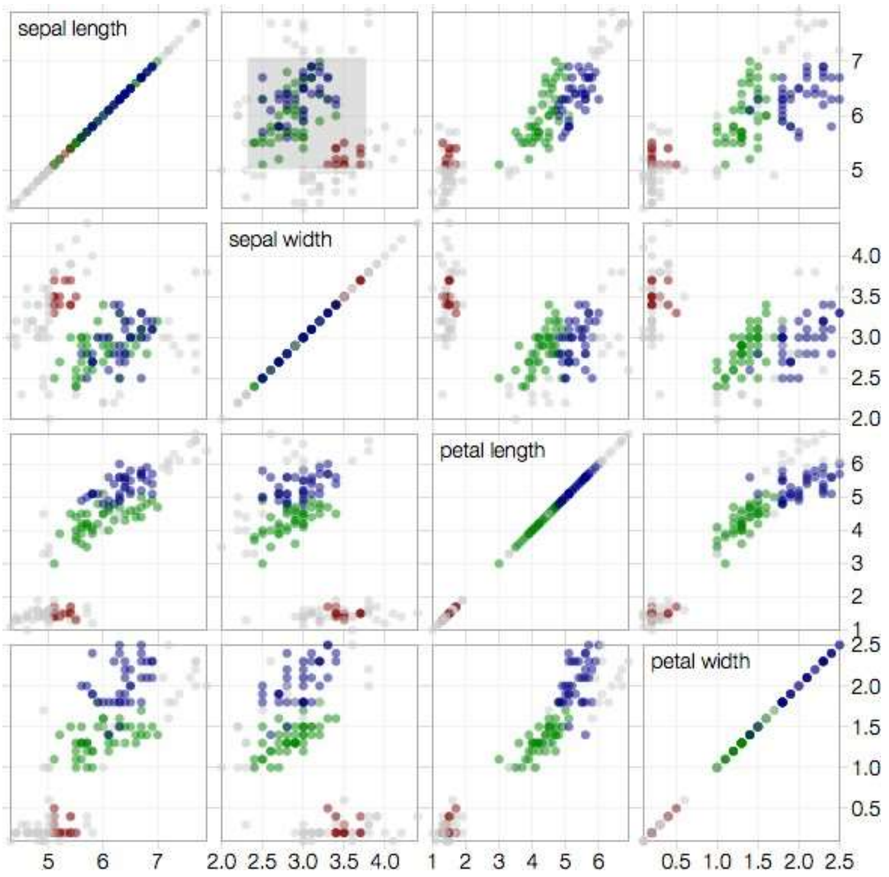
Ирис щетинистый
(*Iris setosa*)

Ирис разноцветный
(*Iris versicolor*)

Ирис виргинский
(*Iris virginica*)

** Ирисы Фишера, 1936г.*

ЦВЕТКИ ИРИСА: ДАННЫЕ



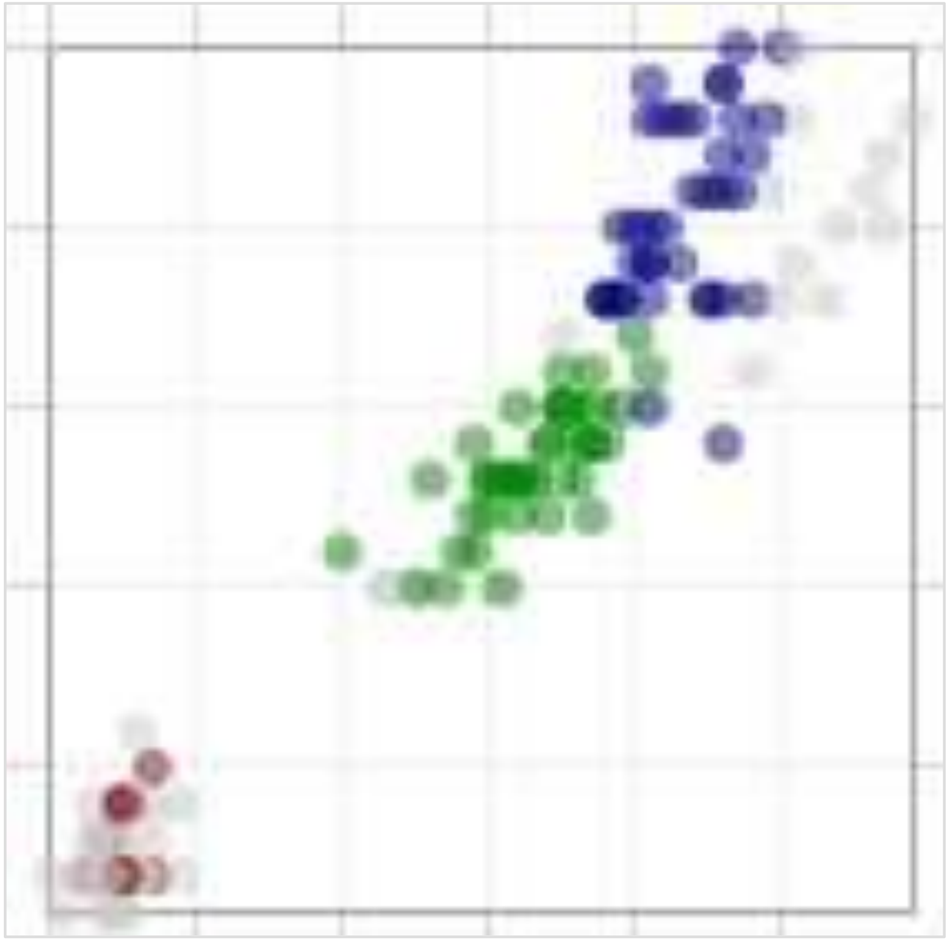
Дано:

- 3 вида цветков ириса
- 4 параметра: 2 длины и 2 ширины листа
- по 50 наборов значений на каждый вид

Найти:

- тип цветка по 4 параметрам

ЦВЕТКИ ИРИСА: ДАННЫЕ



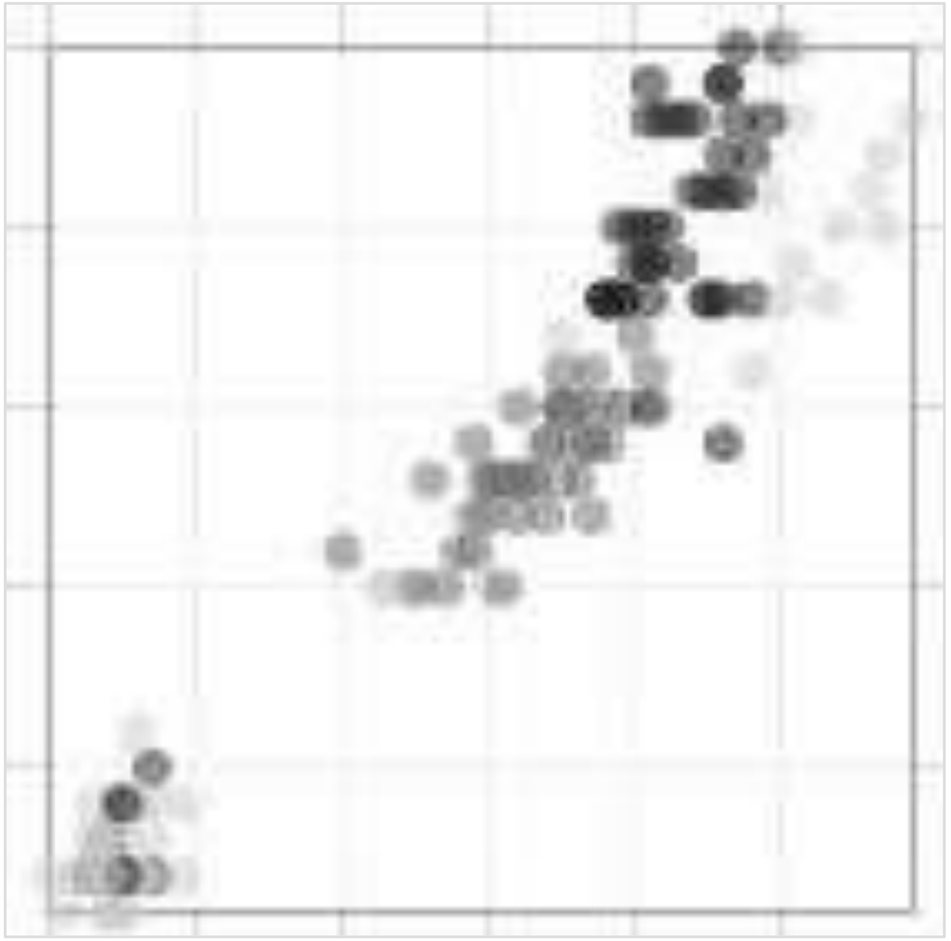
Дано:

- 3 вида цветков ириса
- 4 параметра: 2 длины и 2 ширины листа
- по 50 наборов значений на каждый вид

Найти:

- тип цветка по 4 параметрам

ЦВЕТКИ ИРИСА: ДАННЫЕ



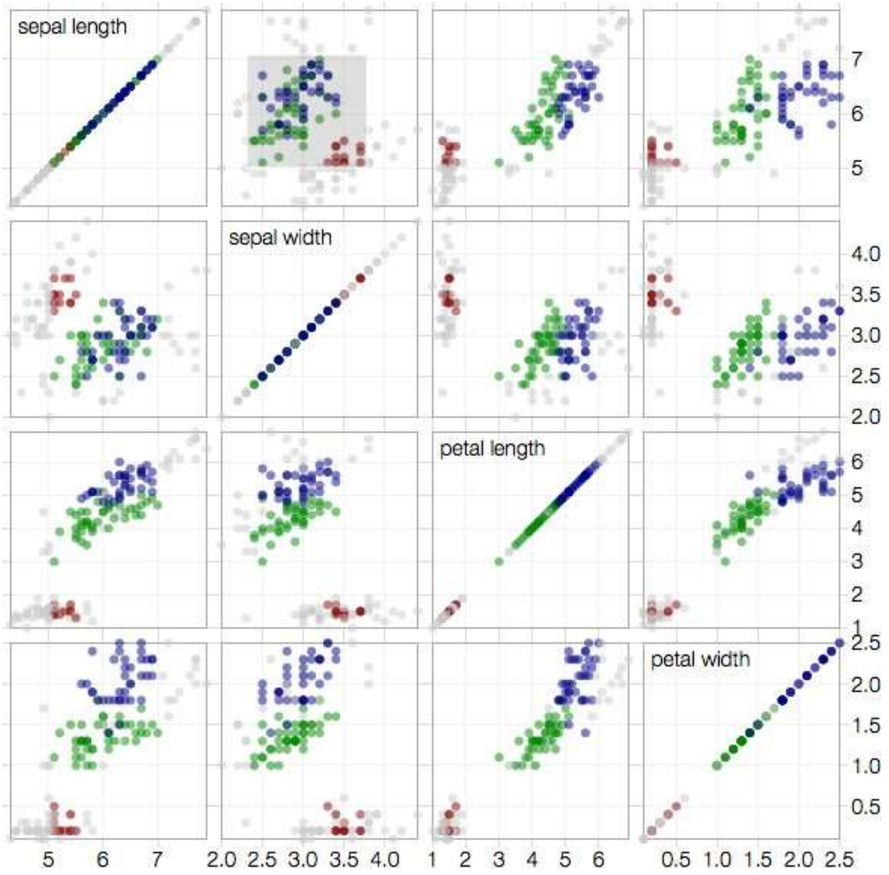
Дано:

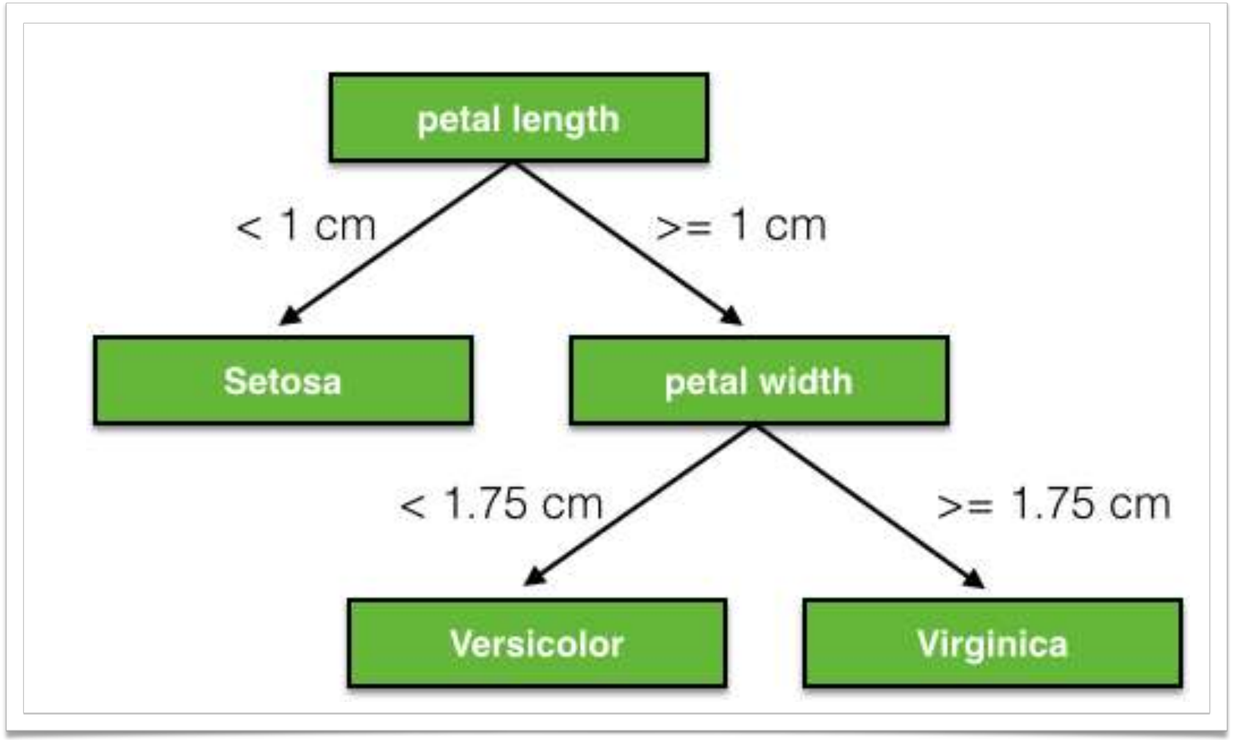
- 3 вида цветков ириса
- 4 параметра: 2 длины и 2 ширины листа
- по 50 наборов значений на каждый вид

Найти:

- тип цветка по 4 параметрам

ЦВЕТКИ ИРИСА: РЕШАЮЩЕЕ ДЕРЕВО

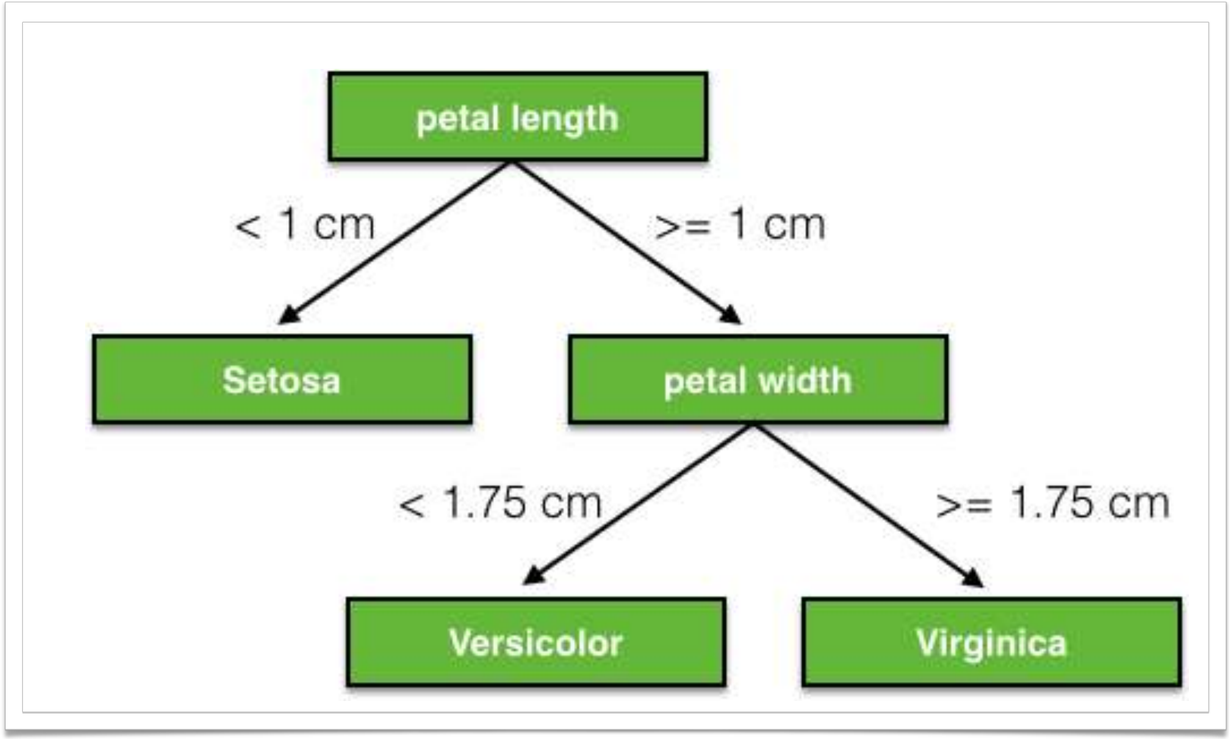




КАК ПОСТРОИТЬ ДЕРЕВО?

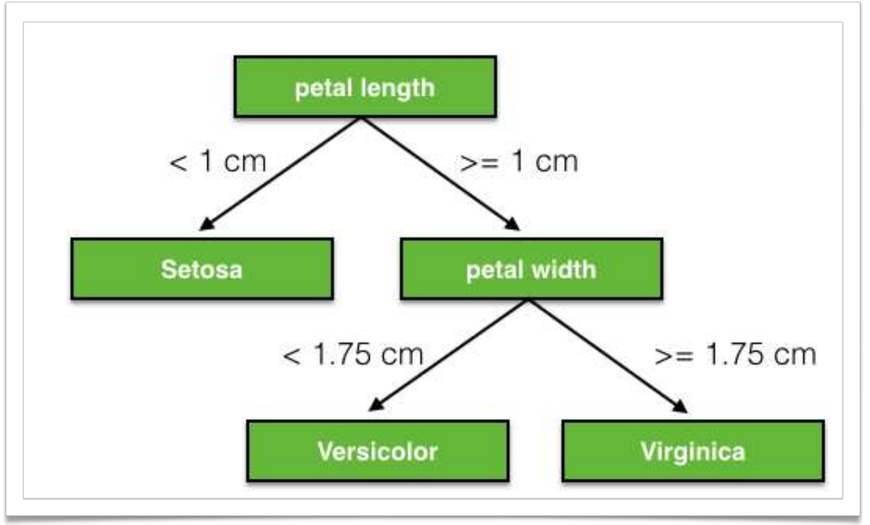
Определить:

- вид правила разбиения
- критерий информативности разбиения
- критерий останова
- метод стрижки
- обработка пропусков



ВИД ПРАВИЛА ДЛЯ РАЗБИЕНИЯ

- **одномерное**:
  сравнивается значение одной фичи вектора x
- **линейное**:
  сравнивается линейная комбинация фичей x
- **метрическое**:
  расстояние до точки признакового пространства



здесь используется одномерный предикат: сравнение идёт лишь по одной фиче из вектора признаков

ФУНКЦИОНАЛ КАЧЕСТВА РАЗБИЕНИЯ

Идея:

- взять признак
- отсортировать его по возрастанию
- в зависимости от целевой переменной установить порог разделения выборки на две, максимально снижая численно выражаемый разброс внутри каждой из 2 групп
- подобрать лучшее с точки зрения улучшения разбиение

Вопрос: а как измерить улучшение?

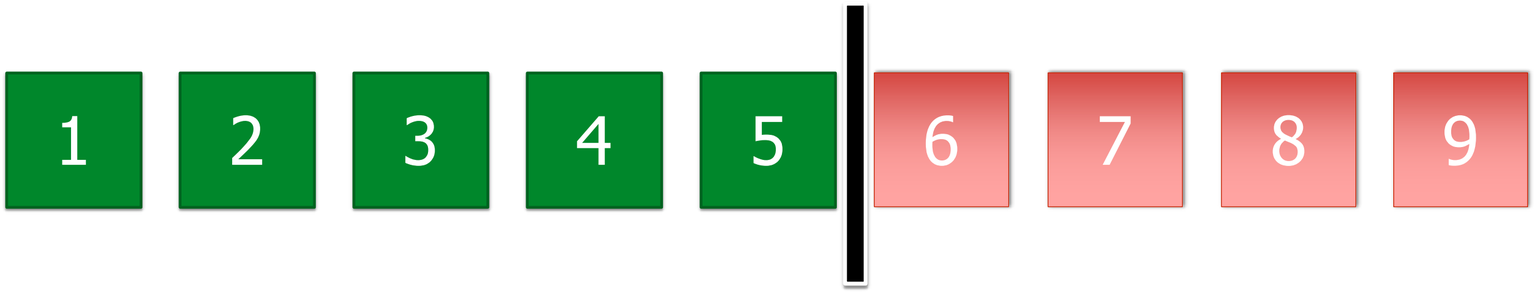
ИЗМЕРЕНИЕ ПОЭТАПНОГО УЛУЧШЕНИЯ

Есть 1 группа, в ней 2 класса

Пусть H(R) - «критерии информативности» группы,
больше разнообразия - больше H(R) - хуже для классификатора

Будем измерять улучшение разбиения по функционалу вида: $IG(R) = H(R) - q_{left}*H(R_{left}) - q_{right}*H(R_{right})$, где q_{left} и q_{right} - доли объектов, попавших в левый или правый класс соответственно

ИЗМЕРЕНИЕ ПОЭТАПНОГО УЛУЧШЕНИЯ

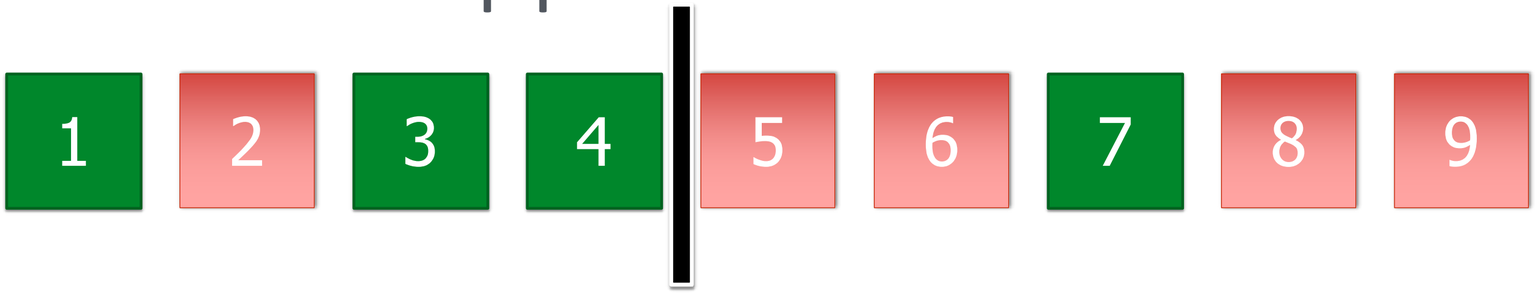


$$IG(R) = H(R) - q_{left}*H(R_{left}) - q_{right}*H(R_{right})$$

$$H(R) = x > 0$$
$$H(R_{left}) = 0$$
$$H(R_{right}) = 0$$

$$IG(R) = x - 5/9*0 - 4/9*0 = x > 0$$

КРИТЕРИЙ ДЖИНИ



$$H(R) = \sum_{k=1}^{K} p_k(1 - p_k)$$

K - количество классов

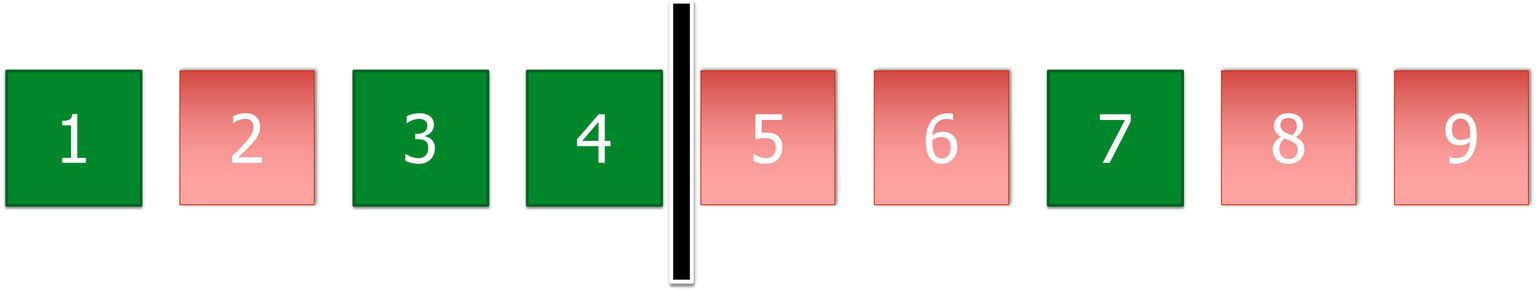
p_k - доля класса в выборке

$$IG(R) = H(R) - q_{left}*H(R_{left}) - q_{right}*H(R_{right})$$

$$H(R) = 4/9*(1-4/9) + 5/9*(1-5/9) = 0.494$$
$$H(R_{left}) = 3/4*(1-3/4) + 1/4*(1-1/4) = 0.375$$
$$H(R_{right}) = 1/5*(1-1/5) + 4/5*(1-4/5) = 0.32$$

$$IG(R) = 0.494 - 4/9*0.375 - 5/9*0.32 = \mathbf{0.15}$$

ЭНТРОПИЙНЫЙ КРИТЕРИЙ



$$H(R) = -\sum_{k=1}^{K} p_k \log p_k$$

K - количество классов

p_k - доля класса в выборке

$$IG(R) = H(R) - q_{left} * H(R_{left}) - q_{right} * H(R_{right})$$

$$H(R) = -4/9*\log2(4/9) - 5/9*\log2(5/9) = 0.991$$

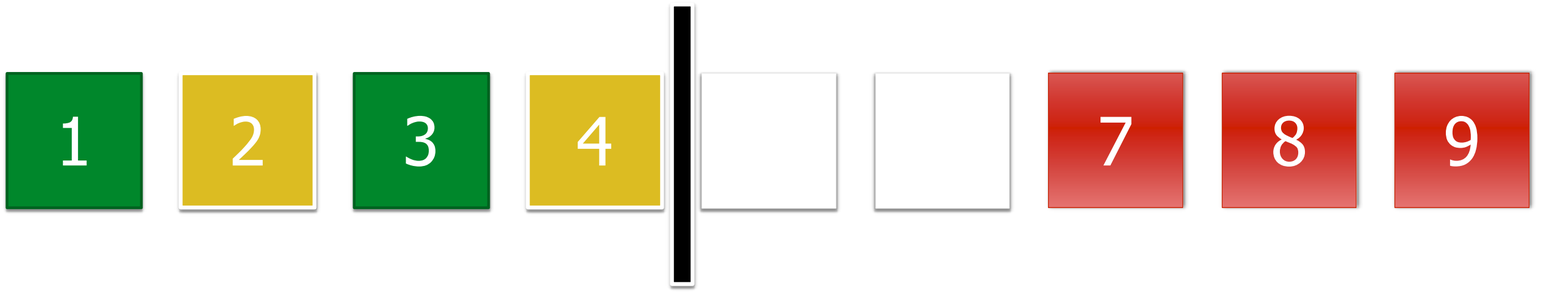
$$H(R_{left}) = -3/4*\log2(3/4) - 1/4*\log2(1/4) = 0.81$$

$$H(R_{right}) = -1/5*\log2(1/5) - 4/5*\log2(4/5) = 0.72$$

$$IG(R) = 0.991 - 4/9*0.811 - 5/9*0.722 = \mathbf{0.229}$$

КРИТЕРИЙ ДЖИНИ

$$IG(R) = ?$$



КРИТЕРИИ ОСТАНОВА

- Останов, когда в каждом листе объекты только одного класса
- Ограничение max глубины дерева
- Ограничение min число объектов в листьях
- Требование улучшения функционала качества при дроблении не менее, чем x или на x%

ПРОБЛЕМА ПРОПУСКОВ

- Выкинуть объекты с пропусками из обучающей (что на тестовой?)
- Замена на значения вне средние, медианные..
- Заменить на значения вне области значений фич
- Модифицировать алгоритм построения и работы дерева: включать элементы с пропусками в обе ветки дерева, но взвешивать качество разбиения по объёму пропусков

СТРИЖКА ДЕРЕВЬЕВ (PRUNING)

Стрижка из полностью построенного дерева убирает наименее информативные листья

Стрижка работает лучше раннего останова

Редко используется, т.к. деревья не используются самостоятельно, а в ансамблях она излишняя (там либо нужно переобучение, либо используется ограничение глубины)

В основе идея регуляризации: в функционале качестве поддерева линейно штрафуется количество листьев

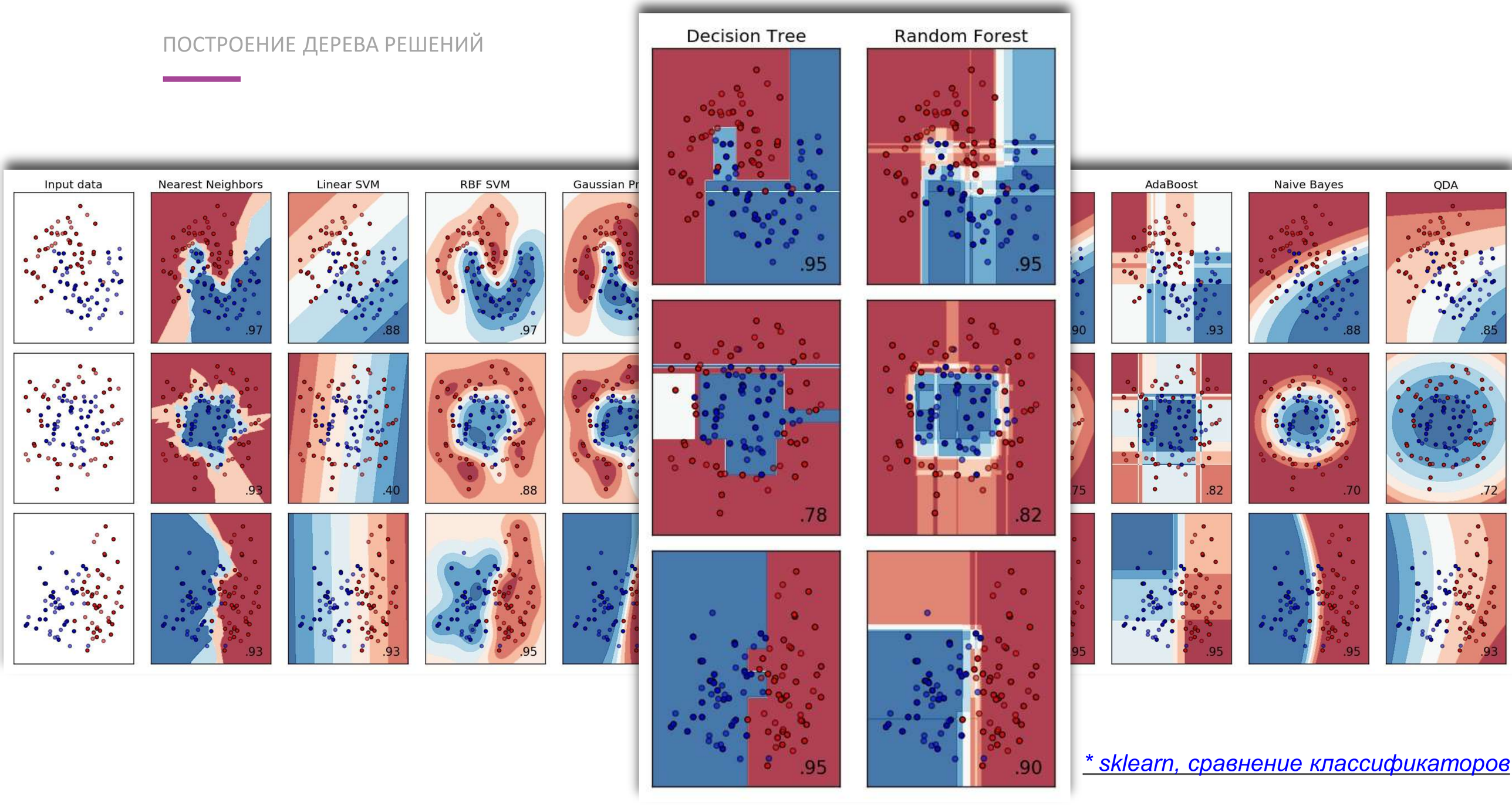
ПОПУЛЯРНЫЕ МЕТОДЫ ПОСТРОЕНИЯ

Деревья в силу дискретности не сводятся к оптимизации в аналитическом виде, поэтому все методы их построения являются эвристическими и жадными

Популярные методы отличаются ранее рассмотренными параметрами построения дерева

- ID3: энтропийный критерий, максимально жадный, требуется стрижка (1986)
- C4.5, C5.0: нормированный энтропийный критерий
- CART: критерий Джини - используется в sklearn (optimized)

ПОСТРОЕНИЕ ДЕРЕВА РЕШЕНИЙ



** sklearn, сравнение классификаторов*

РЕАЛИЗАЦИЯ В SKLEARN

sklearn.tree.DecisionTreeClassifier

* criterion='gini'
* splitter='best'
* max_depth=None
* min_samples_split=2
* min_samples_leaf=1
* min_weight_fraction_leaf=0.0
* max_features=None
* random_state=None
* max_leaf_nodes=None
* min_impurity_split=1e-07
* class_weight=None
* presort=False

Основные характеристики

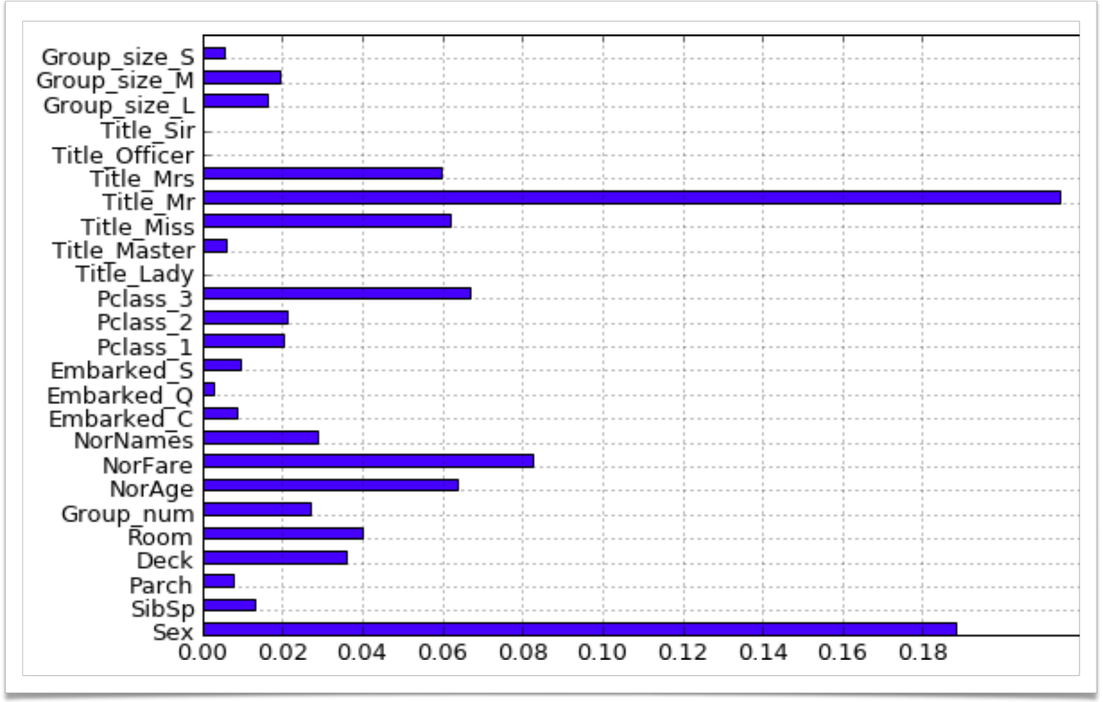
* 12 параметров
* Функционал качества: Джини / энтропия
* Реализованы различные простые критерии останова: кол-во объектов, улучшение качества..
* Не реализована стрижка дерева

Основные методы

* fit
* predict, predict_proba

РЕАЛИЗАЦИЯ В SKLEARN. БОНУС

Деревья могут оценивать важность фичей



Например, судя по решению, на выживаемость на Титанике сильнее всего влияли:

* наличие в обращении «Mr.»
* пол
* уровень дохода
* проживание в 3 классе
* возраст
* наличие в обращении «Mrs» / «Miss»

Решение на Kaggle

3. ДОСТОИНСТВА И НЕДОСТАТКИ ДЕРЕВЬЕВ РЕШЕНИЙ

ДОСТОИНСТВА

- Легко интерпретировать, визуализировать, «белый ящик»
- Простота подготовки данных: не требуется нормализация, dummy переменные, возможны пропуски
- Скорость работы

Official documentation view

НЕДОСТАТКИ

- Острая проблема переобучения
- Неустойчивость
- Не учитывает нелинейные зависимости или даже простые линейные, которые идут не по осям координат (f.e., представьте дерево для классификатора вида y>x)
- Чувствителен к несбалансированным классам
- Хорошо интерполирует, плохо экстраполирует

Official documentation view

ПРАКТИЧЕСКОЕ ЗАДАНИЕ 2

4. ОЦЕНКА КАЧЕСТВА КЛАССИФИКАЦИИ

МАТРИЦА ОШИБОК

confusion matrix	y = 1	y = 0
a = 1	True Positive	False Positive
a = 0	False Negative	True Negative

На тестовой выборке имеем:

* y - вектор истинных значений
* a - вектор предсказаний классификатора

Будем раскладывать все пары (предсказание, истина) по ячейкам матрицы ошибок

ACCURACY

confusion matrix	y = 1	y = 0
a = 1	0	0
a = 0	2	998

Accuracy, Доля верных ответов

(в просторечии точность, но не путать с точностью из ML!)

Простая метрика, но абсолютно непоказательна в задачах с несбалансированными классами

Пример:
определение качества теста на рак. Тест постоянно предсказывает отсутствие рака. Доля верных ответов: 99.8%

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

PRECISION

confusion matrix	y = 1	y = 0
a = 1	170	100
a = 0	30	700

Precision, Точность

отсутствие ложных срабатываний

Пример:
правильное распознавание намерения пользователя: лучше переспросить пользователя, чем сделать не то, что нужно

Точность = 170 / (170+100) = 0.629

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

RECALL

confusion matrix	y = 1	y = 0
a = 1	170	100
a = 0	30	700

Recall, Полнота

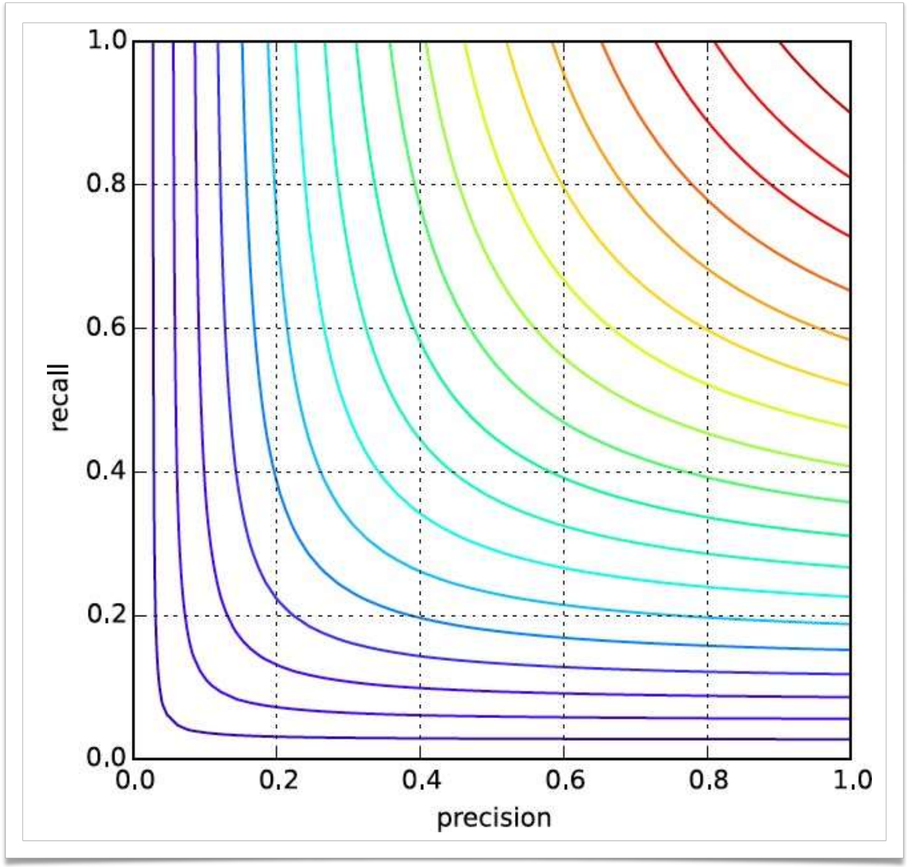
отсутствие ложных пропусков

Пример:
определение мошеннических действий в банке:
лучше найти лишнее и проверить, чем не найти

Точность = 170 / (170+30) = 0.85

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

F1 - МЕРА



** из материалов к курсу Воронцова на Coursera*

F1-мера

комбинация точности и полноты в одну метрику

Пример:
правильное распознавание намерения пользователя. Насколько мы уверены в том, что правильно поняли? Надо ли уточнить?

F1 = 2 * 0.629 * 0.85 / (0.629 + 0.85) = 0.723

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

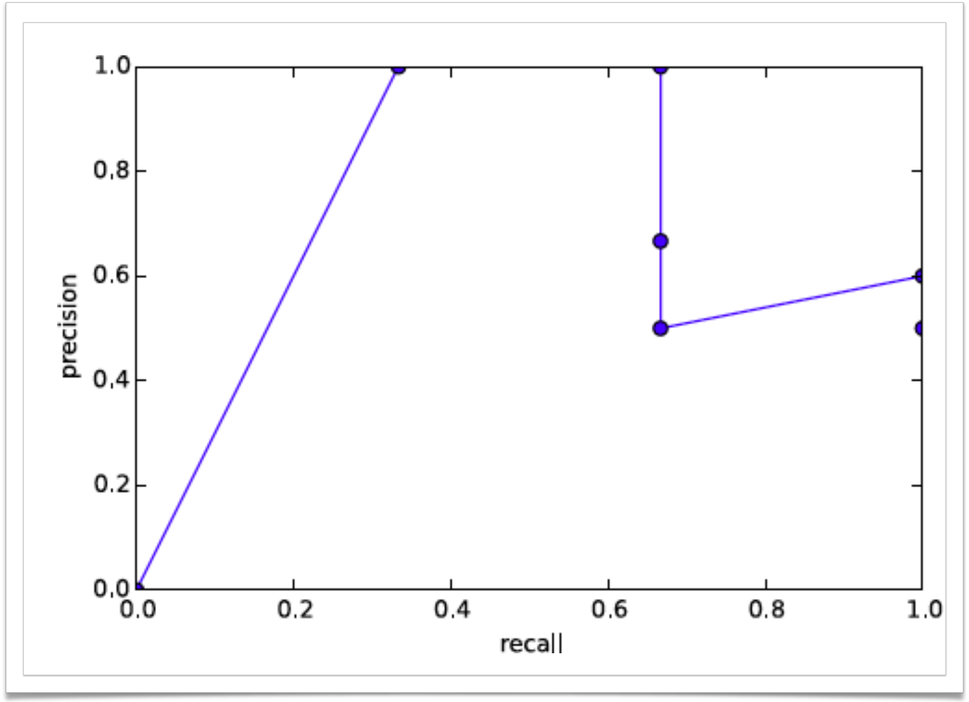
GOING DEEPER

Классификаторы обычно работают в 2 этапа:

* оценка вероятности принадлежности к классам a(x)
* выбор порога отсечения, при котором идёт распределение в тот или иной класс

Это 2 отдельные задачи, после получения оценки вероятности можно отсортировать объекты и в различные периоды времени использовать разные пороги

AUC-PRC



** из материалов к курсу Воронцова на Coursera*

Precision - Recall (PR curve)

мера качества разделения на классы

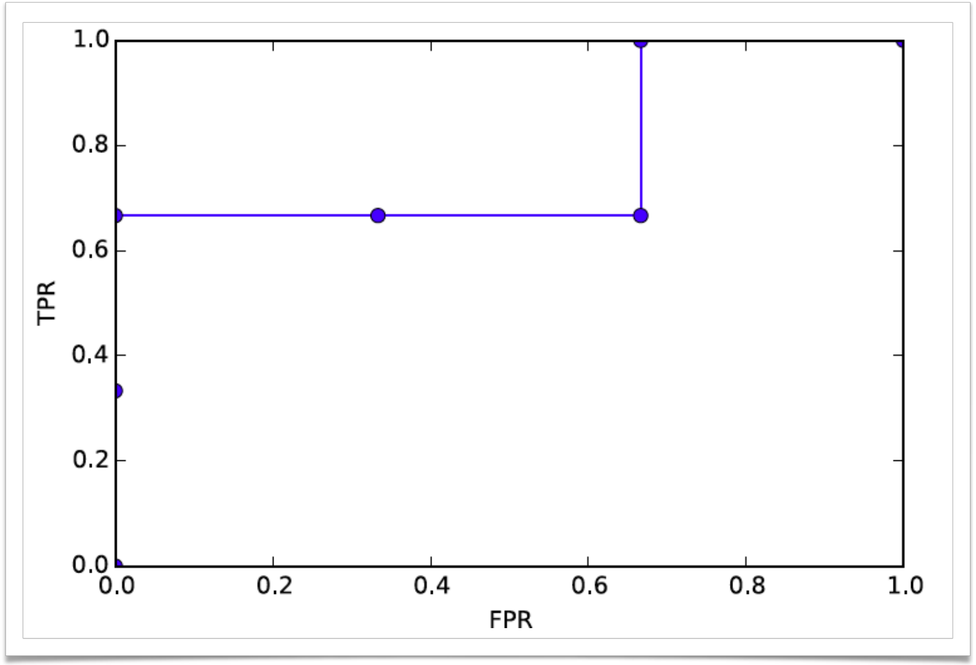
Построение:

1. Считаем вероятности принадлежности классу 1
2. Сортируем объекты по вероятности
3. Для каждого порога отсечения между объектами считаем precision и recall и последовательно наносим на график
4. Считаем площадь под кривой

AUC - PRC (area under curve: precision recall curve)

итоговая метрика. Больше - лучше

AUC-ROC



** из материалов к курсу Воронцова на Coursera*

Receiver Operating Characteristic curve

мера качества разделения на классы

Построение:

аналогично PR-кривой, но наносятся точки

FPR = FP / (FP + TN)
1-FPR - *специфичность алгоритма*

TPR = TP / (TP + FN)
TPR - *чувствительность алгоритма*

AUC - ROC

итоговая метрика. Больше - лучше

AUC-PRC **VS** AUC-ROC

AUC-PRC лучше использовать в несбалансированных классах

AUC-ROC можно использовать, когда алгоритм будет оцениваться на одних данных, но с разным соотношением классов

МНОГОКЛАССОВАЯ КЛАССИФИКАЦИЯ

Задача: $a(x) \in \{1, \ldots, K\}$
Сводится к K задачам отделения класса N от остальных

Как усреднить качество K задач?

МНОГОКЛАССОВАЯ КЛАССИФИКАЦИЯ

micro-averaging:
* вычислим confusion matrix для каждой задачи
* усредним по задачам
* вычислим итоговую метрику

классы делают вклад, пропорциональный размеру

macro-averaging:
* вычислим итоговую метрику для каждой задачи
* усредним по задачам

все классы делают равный вклад

РЕАЛИЗАЦИЯ В SKLEARN

sklearn.metrics

Some of these are restricted to the binary classification case:

`matthews_corrcoef` (y_true, y_pred[, ...])	Compute the Matthews correlation coefficient (MCC) for binary classes
`precision_recall_curve` (y_true, probas_pred)	Compute precision-recall pairs for different probability thresholds
`roc_curve` (y_true, y_score[, pos_label, ...])	Compute Receiver operating characteristic (ROC)

Others also work in the multiclass case:

`cohen_kappa_score` (y1, y2[, labels, weights])	Cohen's kappa: a statistic that measures inter-annotator agreement.
`confusion_matrix` (y_true, y_pred[, labels, ...])	Compute confusion matrix to evaluate the accuracy of a classification
`hinge_loss` (y_true, pred_decision[, labels, ...])	Average hinge loss (non-regularized)

Some also work in the multilabel case:

`accuracy_score` (y_true, y_pred[, normalize, ...])	Accuracy classification score.
`classification_report` (y_true, y_pred[, ...])	Build a text report showing the main classification metrics
`f1_score` (y_true, y_pred[, labels, ...])	Compute the F1 score, also known as balanced F-score or F-measure
`fbeta_score` (y_true, y_pred, beta[, labels, ...])	Compute the F-beta score
`hamming_loss` (y_true, y_pred[, labels, ...])	Compute the average Hamming loss.
`jaccard_similarity_score` (y_true, y_pred[, ...])	Jaccard similarity coefficient score
`log_loss` (y_true, y_pred[, eps, normalize, ...])	Log loss, aka logistic loss or cross-entropy loss.
`precision_recall_fscore_support` (y_true, y_pred)	Compute precision, recall, F-measure and support for each class
`precision_score` (y_true, y_pred[, labels, ...])	Compute the precision
`recall_score` (y_true, y_pred[, labels, ...])	Compute the recall
`zero_one_loss` (y_true, y_pred[, normalize, ...])	Zero-one classification loss.

And some work with binary and multilabel (but not multiclass) problems:

`average_precision_score` (y_true, y_score[, ...])	Compute average precision (AP) from prediction scores
`roc_auc_score` (y_true, y_score[, average, ...])	Compute Area Under the Curve (AUC) from prediction scores

Основные характеристики

* 19 функций
* Схожий интерфейс: функции от y, y_pred

Пример использования f1_score

```
>>> from sklearn.metrics import f1_score
>>> y_true = [0, 1, 2, 0, 1, 2]
>>> y_pred = [0, 2, 1, 0, 0, 1]
>>> f1_score(y_true, y_pred, average='macro')
0.26...
>>> f1_score(y_true, y_pred, average='micro')
0.33...
>>> f1_score(y_true, y_pred, average='weighted')
0.26...
>>> f1_score(y_true, y_pred, average=None)
array([ 0.8, 0. , 0. ])
```

ПРАКТИЧЕСКОЕ ЗАДАНИЕ 3

ЧТО МЫ СЕГОДНЯ УЗНАЛИ

ЧТО МЫ СЕГОДНЯ УЗНАЛИ

1. Деревья решений, объединённые в «лес», составляют одни из наиболее сильных алгоритмов. По одиночке же они являются слабыми, зато очень легко интерпретируемыми и визуализируемыми алгоритмами
2. Деревья позволяют оценивать важность признаков
3. Метрик качества много, они разные по смыслу, для своих задач надо выбирать подходящую

ПОЛЕЗНЫЕ МАТЕРИАЛЫ

1. Документация sklearn по деревьям
2. Open Data Science, habrahabr: Классификация, дерево решений и метод ближайших соседей
3. Лекция Евгения Соколова на ФКН ВШЭ по деревьям Конспект; ipynb тетрадка
4. Метрики sklearn
5. Метрики kaggle
6. Объяснение метрик на курсе Coursera от Соколова & Воронцова

НЕТОЛОГИЯ
групп

Спасибо за внимание!

Артур Сапрыкин