Лекция 4.

Дерево решенияй. Лес. Случайный лес.

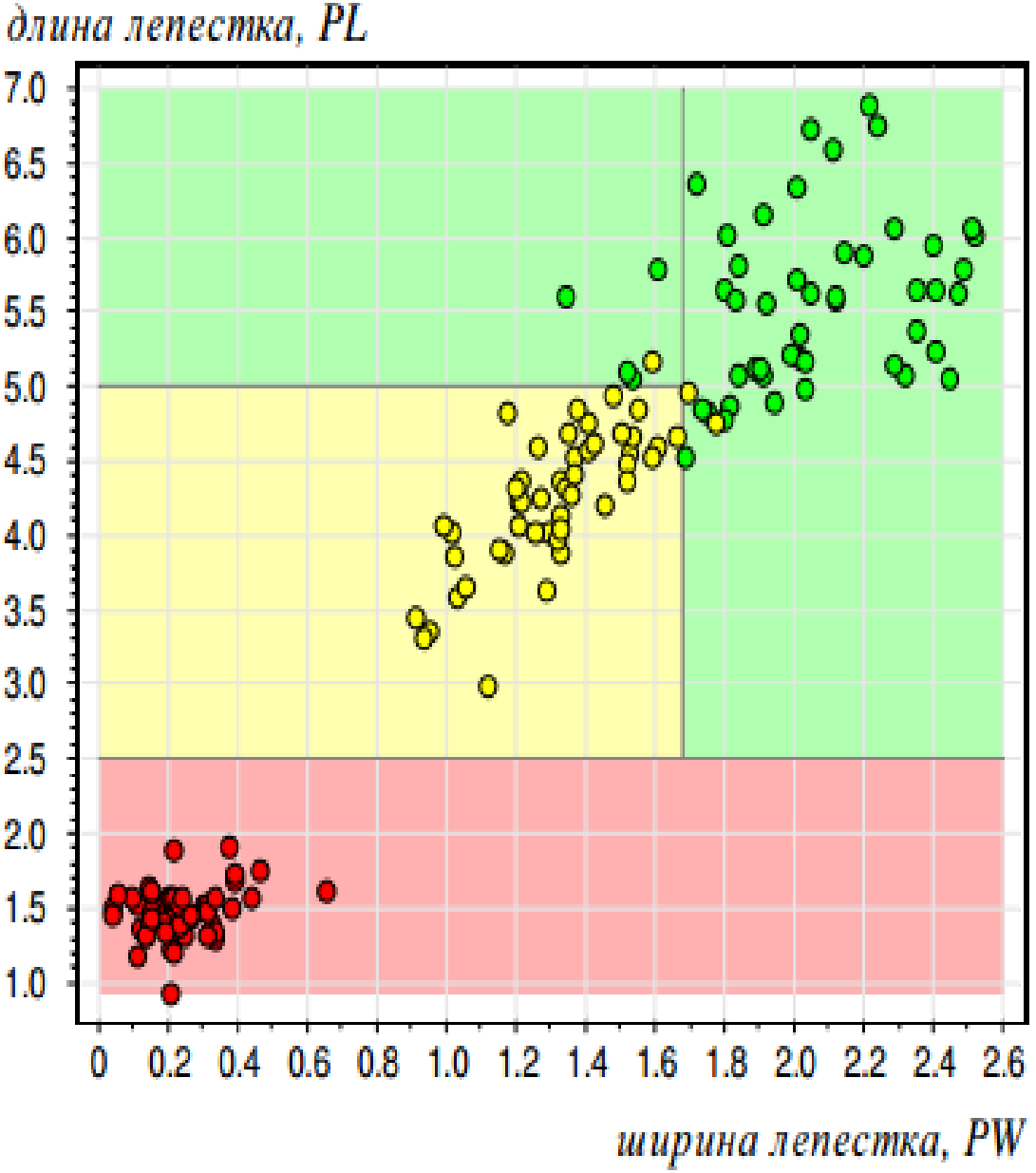
Павел Владимирович
Слипенчук
PavelMSTU@stego.su
ИУ-8

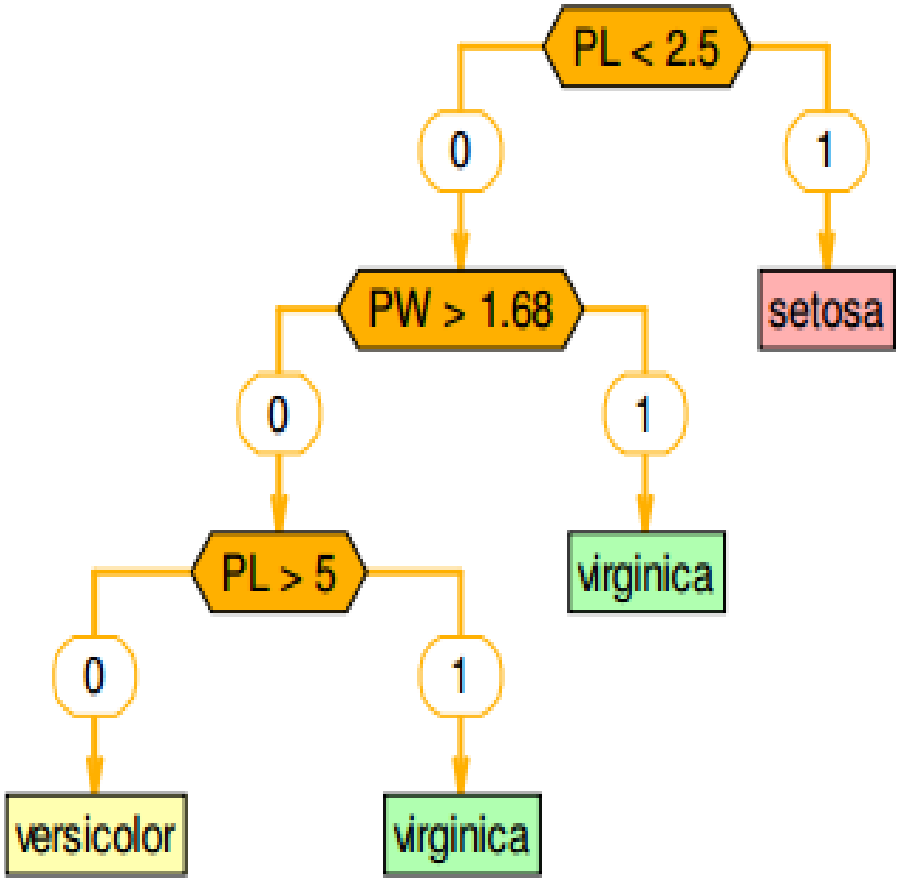
Москва
30.09.2016

Дерево решений

*http://scikit-learn.org/stable/tutorial/machine_learning_map/index.htmlv

Ирисы Фишера





Ирисы Фишера

Дерево решений – это совокупность булевых выражений.

setosa	$r_1(x) = [PL \leqslant 2.5]$
virginica	$r_2(x) = [PL > 2.5] \wedge [PW > 1.68]$
virginica	$r_3(x) = [PL > 5] \wedge [PW \leqslant 1.68]$
versicolor	$r_4(x) = [PL > 2.5] \wedge [PL \leqslant 5] \wedge [PW < 1.68]$

Случайный лес (Random forest)

Павлов
Юрий Леонидович
(р.1949)

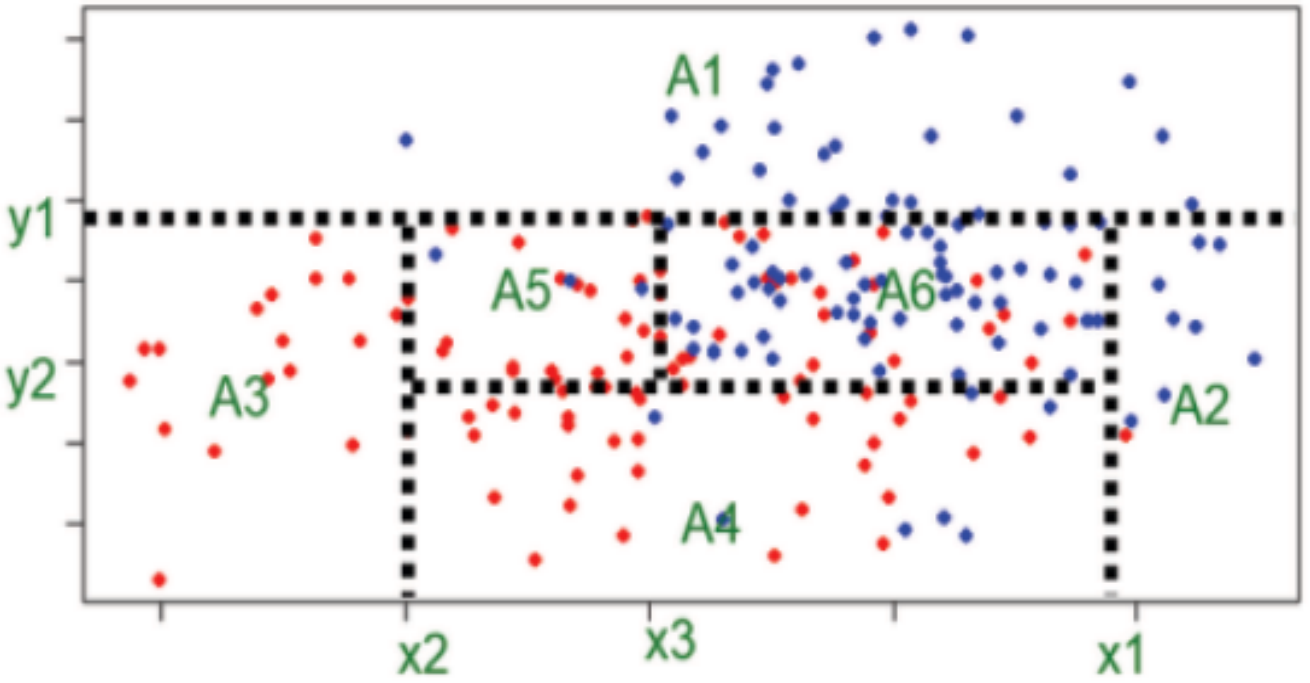
Лео Брейнман
(Leo Breiman)
(1928 — 2005)

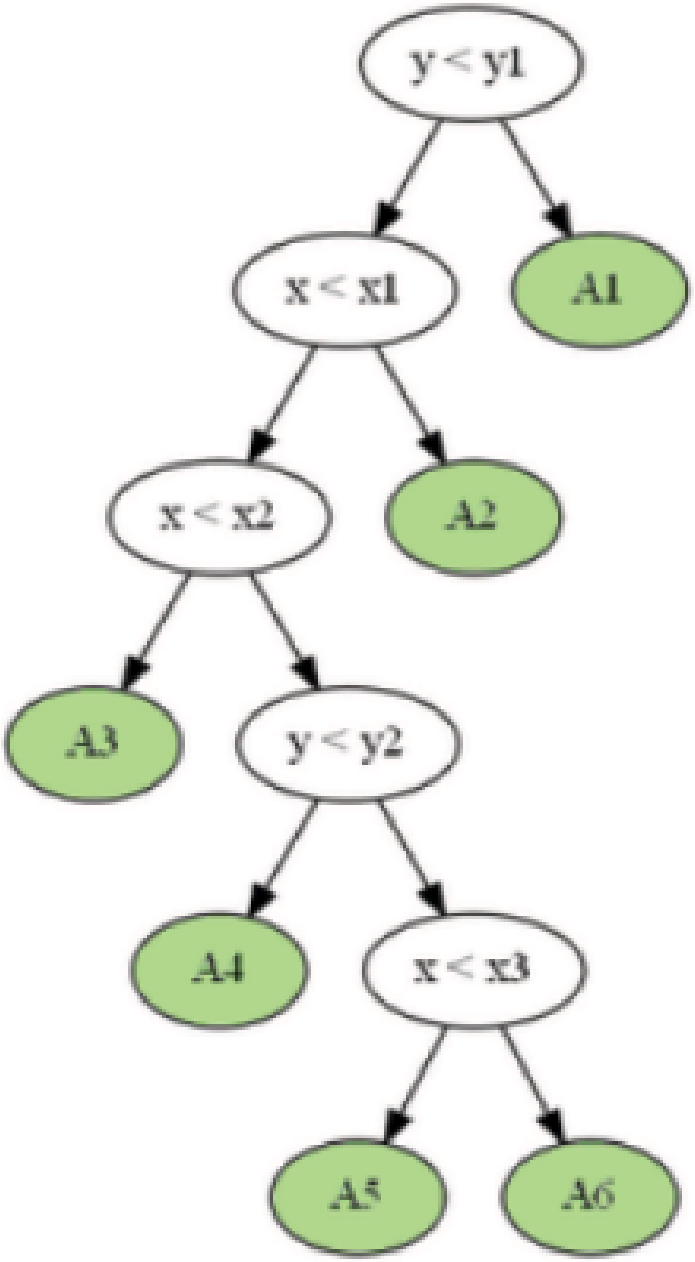
Случайный лес (Random forest)

Лес решений – это *ансамбль функции голосования,* каждый классификатор которого – дерево решений.

Случайный лес – это лес решений, получаемый в результате *случайного построения (?каких именно?)* множества различных деревьев решений

Построение дерева решений





- Каждому бину ставиться в соответствие класс
- Качество бина – **индекс Джини**.

Дерево решений. Термины

Вспоминаем теорию графов.

- Узел
- Корневой узел (=Корень)
- Внутренний узел
- Терминальный узел (=Лист)
- Бинарное дерево
- Предок
- Потомок
- Родитель
- Сын

Построение дерева решений

- Выбор бинов в дереве – ***случаен***.
- Выбор подмножества признаков из всех признаков – ***случаен***.
- Выбор подмножества выборки из всей обучающей выборки – ***случаен***.
- Как долго разбивать на бины? Когда следует остановить дерево решений?

Построение дерева решений

- Выбор бинов в дереве – ***случаен***.
- Выбор подмножества признаков из всех признаков – ***случаен***.
- Выбор подмножества выборки из всей обучающей выборки – ***случаен***.
- Варианты остановки:

  1) **индекс Джини** достаточно мал

  2) слишком далеко ушли от корня

  3) слишком мало элементов в бине

Построение дерева решений

- Нужно ли случайно строить само дерево решений?

  Плюсы / минусы?

Построение дерева решений

- Нужно ли случайно строить само дерево решений?
- **Плюсы**:

  1) не надо думать

  2) быстро

  3) при больших количествах деревьев, ансамбль «все поправит»
- **Минусы**:

  1) «совсем» не оптимально

Параметры случайного леса

- Критерий остановки
- Количество деревьев
- Как выбирать *признаки* для деревьев
- Как выбирать *обучающую выборку* для каждого дерева.

Усложняем ситуацию

- *Как можно улучшить случайный лес и сделать «случайный лес++»?*

Усложняем ситуацию

- Заменяем функцию голосования на что-то более сложное (например зависящее от мат.ожидания индекса Джини в каждом случае)
- Например: функцию голосования заменяем функцией суммирования. (Интервал: [-1, +1])
- Если индекс Джини близок к 0.5 – отказ от классификации
- Поиск аномалий – «неуверенный» ответ случайного леса.

Достоинства случайного леса

- Очень просто устроен
- Он *всегда* «хоть как-нибудь» да работает. (следствие «центральной эмпирической теоремы»)
- Легко программируется «с нуля»
- Скорость? Быстрый/долгий?

Достоинства случайного леса

- Очень просто устроен
- Он *всегда* «хоть как-нибудь» да работает. (следствие «центральной эмпирической теоремы»)
- Легко программируется «с нуля»
- Скорость. Условия дерева решений if-then-else просты и быстрые. С другой стороны для качественного случайного леса трубуется много деревьев.
- Если есть ПЛИС – мир прекрасен!)

Разминка

- Пусть все деревья ***сбалансированны*** и ***бинарны***. Пусть количество ***узлов*** в каждом дереве рано **M**. Пусть всего деревьев **N**.

  Какова сложность леса решений?

- Случайным образом из **N** элементов выборки берем a без возвращения и из **M** признаков берем b без возвращения. Сколько существует способов выбрать элементы и признаки

Вопросы на засыпку

- Хорошо или плохо справляется дерево решений с *категориальными признаками*?
- Почему в деревьях решений никогда не приводят различные сравнимые признаки к определенной шкале для адекватного вычисления *расстояния*?

«Центральная эмпирическая «теорема»» (о случайном лесе)

Если вы работаете в области, в которой совершенно некомпетентны и/или решаете задачу, про которую ничего не знаете – используйте случайный лес.

Этот классификатор – самый лучший!

- Выводы?

«Центральная эмпирическая «теорема»» (о случайном лесе)

- Случайный лес – решения многих задач на вполне приемлемом уровне
- Сравните свой классификатор со случайным лесом – вы поймете насколько ваш классификатор хорош
- Если ваш классификатор ***хуже*** случайного леса – много данных еще «не освоено»
- По настоящему интересные задачи – это те, которые не решаются ***случайным лесом***.

Вопросы на засыпку

- Хорошо или плохо справляется дерево решений с *категориальными признаками*?
- Почему в деревьях решений никогда не приводят различные сравнимые признаки к определенной шкале для адекватного вычисления *расстояния*?

Домашнее задание (Теория)

- **Простое**: прочитайте про «Oblivious Desision Tree» или «Yandex MatrixNet».
- **Сложное**: Распечатайте на бумаге и прочитайте с карандашом в руках статью:

  «СЛУЧАЙНЫЕ ЛЕСА: ОБЗОР»
  С. П. Чистяков
  (2013)

  * советую читать неторопясь и с бумаги «в несколько заходов».

Домашнее задание (Практика)

- **Простое**: освойте методы построения случайных лесов в scikit-learn: 1.11.2. Forests of randomized trees
- **Сложное**: На языке «чистого Си» или, на худой конец, на C++, C#, Scale, Go, … напишите свой класс (набор функций) построения случайного леса. Придумайте «случайный лес++».
- (!) во втором ДЗ будет зачача о RF.