А.Крюков

Машинное обучение в астрофизике

Лекция 4

Деревья решений.

Деревья решений

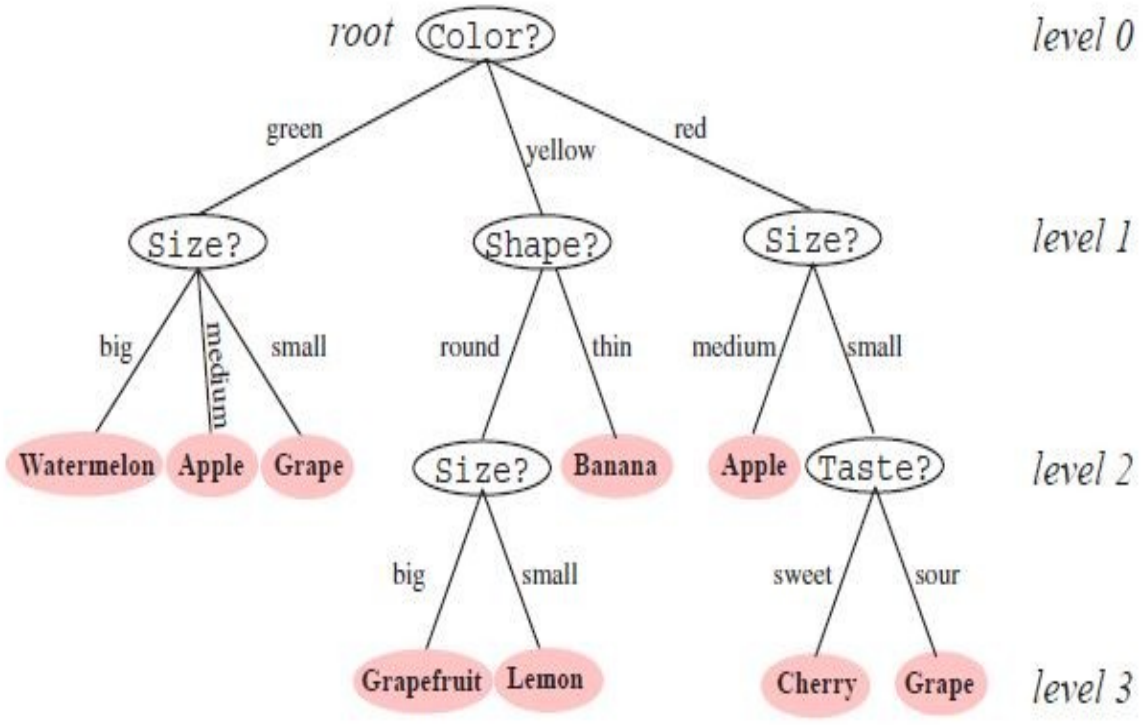
- Деревья решений — это способ классификации при помощи ответов на последовательность вопросов, которые зависят от ответов на предыдущие вопросы.
- Такая последовательность может быть представлена графами
  - Узлы графа — вопросы
  - Ребра — варианты ответов

Деревья решений

- Типы данных
  - Порядковые. Могут быть упорядочены
    - длина, температура;
    - задают вектор признаков.
  - Качественные. Нельзя упорядочить
    - цвет, национальность;
    - задают список атрибутов.

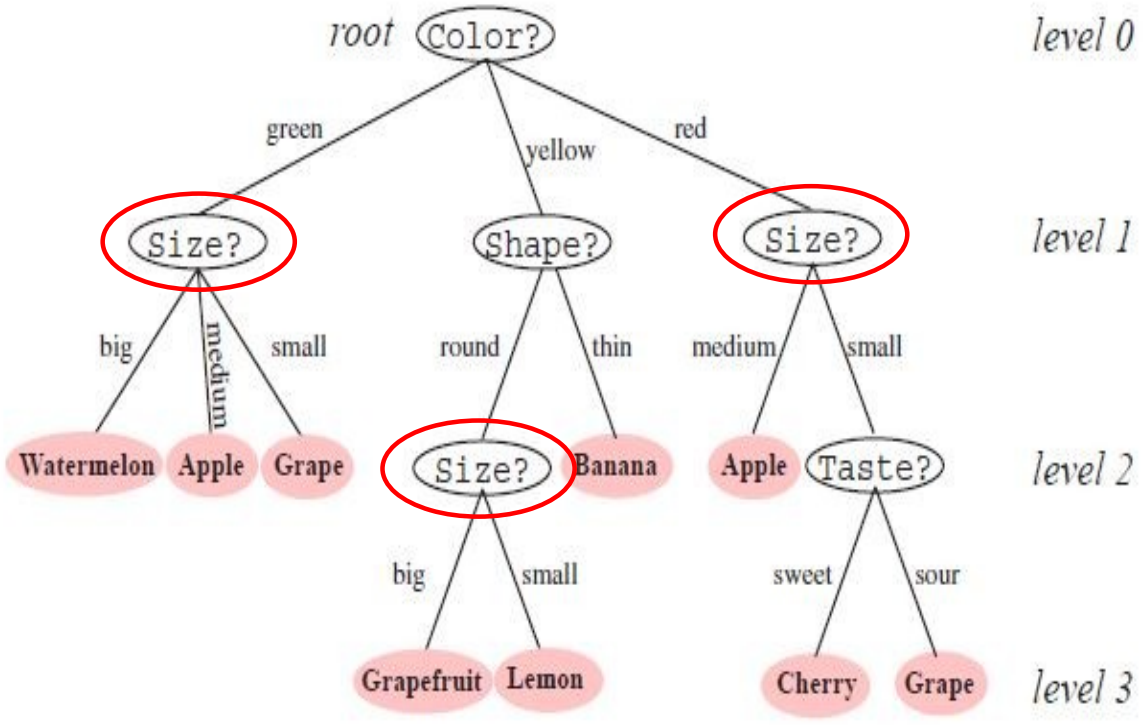
Деревья решений

- Пример: классификация фруктов.



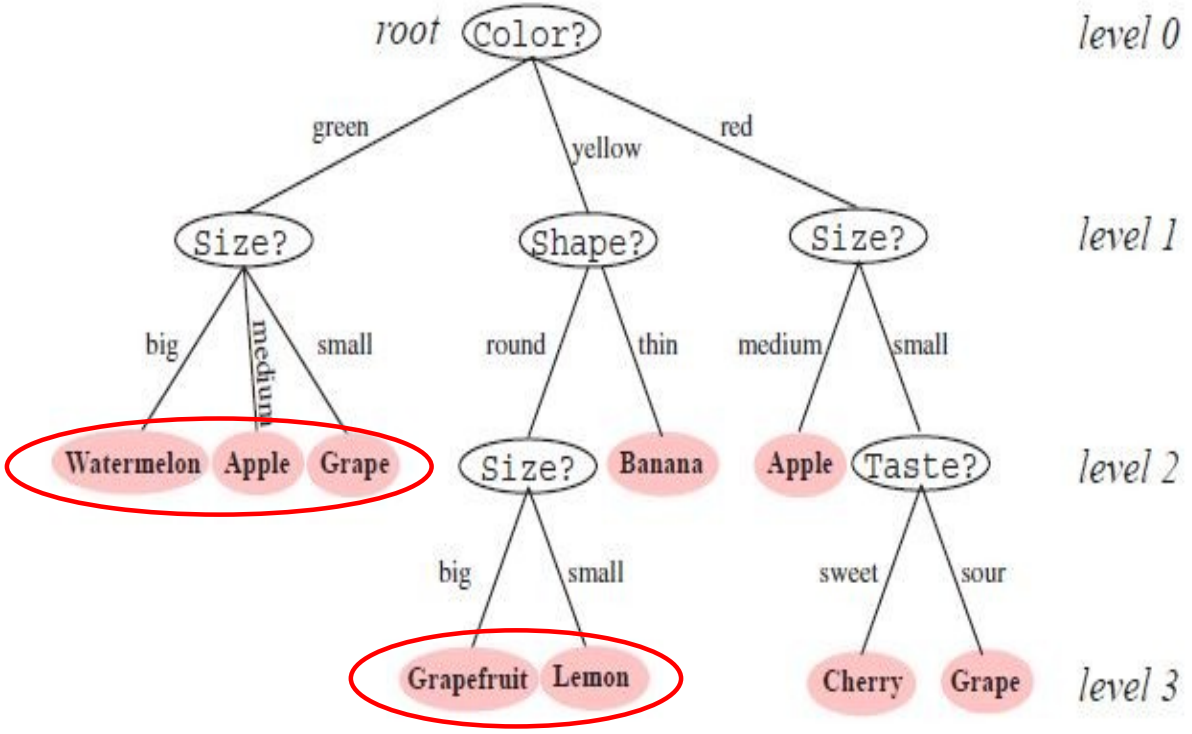
Деревья решений

- Пример: классификация фруктов.
- Один и тот же вопрос может быть в нескольких узлах.



Деревья решений

- Пример: классификация фруктов.
- Один и тот же вопрос может быть в нескольких узлах.
- Ответов на вопрос может быть разное количество.



Деревья решений

Пример: классификация фруктов.

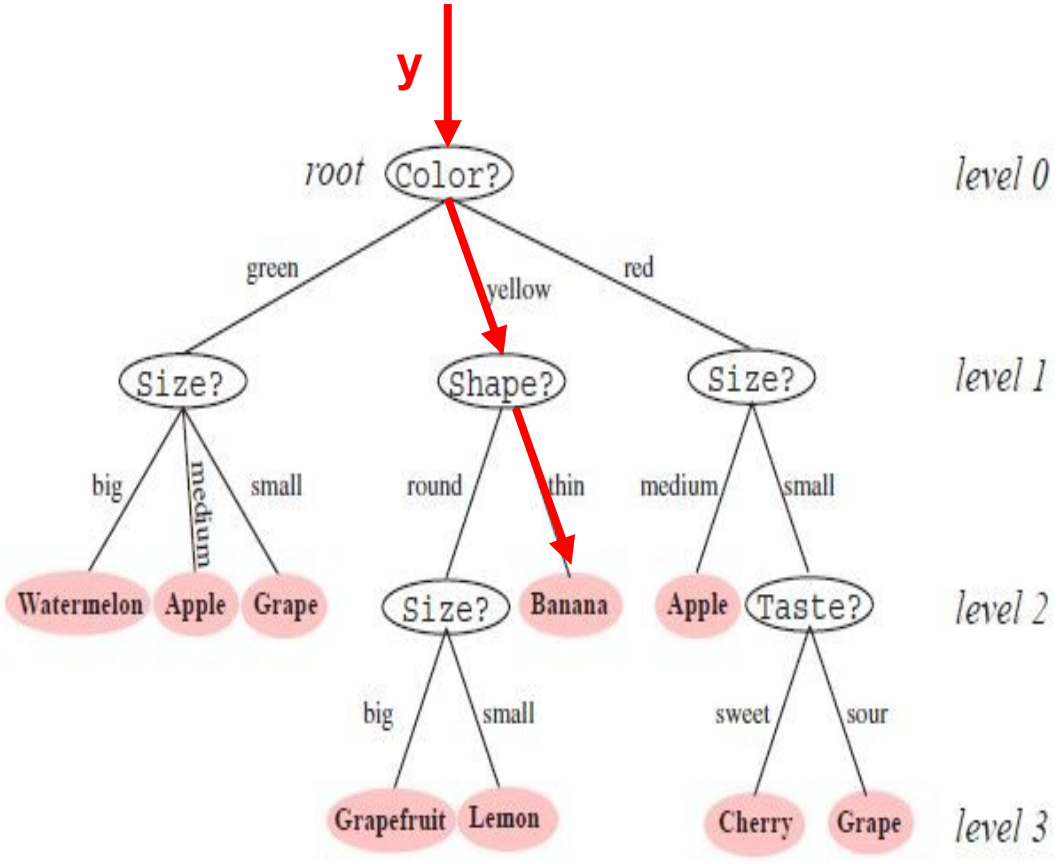
- Один и тот же вопрос может быть в нескольких узлах.
- Ответов на вопрос может быть разное количество.

Пусть y — набор признаков:

y={yellow, thin, sweet, medium}

Пройдя по дереву решений получаем ответ:

y => Banana



Деревья решений

- Видно, что при построении деревьев решений важно правильно разбить пространство параметров.
- Для небольших деревьев — это можно сделать с помощью экспертов.
- Качество классификации с помощью деревьев решений напрямую зависит от качества экспертных оценок.

Как строить деревья решений?

- CART (Classification and Regression Tree) — подход построения ДР

1) Разбиваем выборку на подмножества, которые определяются некоторым признаком. Формулируем классифицирующий вопрос.
2) Конечный ответ соответствует некоторой категории подмножества объектов.

Процесс построения ДР

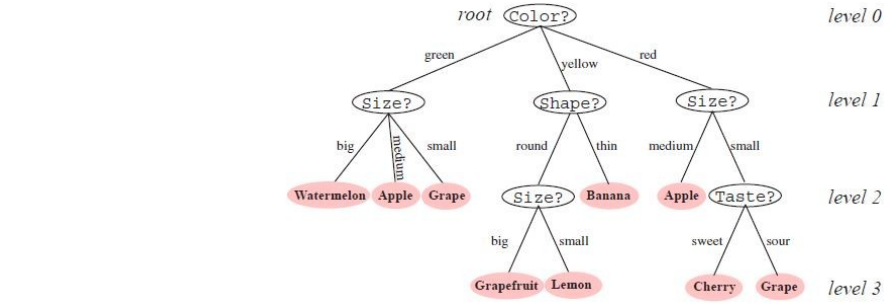
- Фактически, каждый узел ДР — это мини-классификатор
- В процессе построения ДР необходимо дать ответы на следующие вопросы:
  - какой признак будет рассмотрен в текущем узле?
  - число возможных подмножеств, на которые разбивает классификатор данного признака.
  - в каком случае данное подмножество будем считать окончательным и должно быть объявлено листом графа?
  - на каком уровне останавливаем дальнейшую классификацию?
  - как упрощать ДР, если оно становится слишком «ветвистым»?
  - что делать, если некоторые признаки в начальном векторе не определенны?

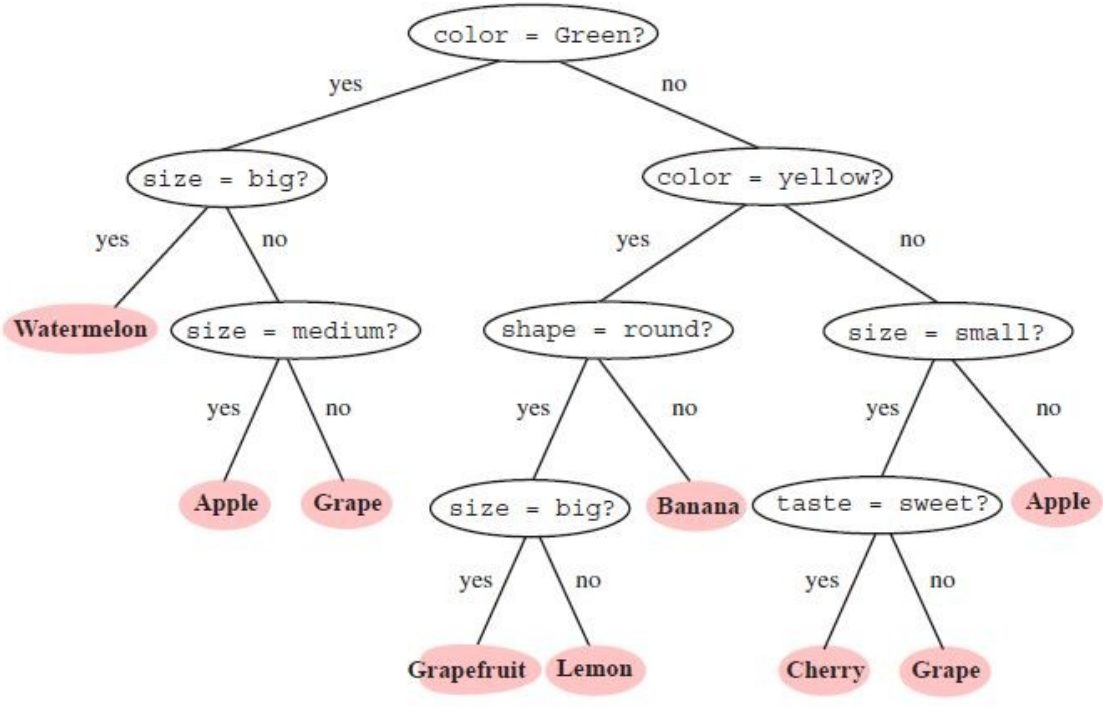
Порядок узлов в ДР

- Очевидно, что от порядка узлов (классифицируемых признаков) зависит качество ДР как классификатора.
  - Общий принцип — чем меньше и проще ДР, тем лучше.
  - Есть признаки более значимые, т. е. хорошо разделяющие множество объектов на подмножества, и менее значимые.
  - Как правило, значимые признаки надо использовать как можно раньше, чтобы в дальнейшем иметь дело с подмножествами малого объема.
  - Пример: Фрукты, признак — цвет. Так как яблоки могут быть очень разных цветов, то использование этого признака в узлах, близких к вершине дерева не будет отделять яблоки от других фруктов.

Число возможных подмножеств

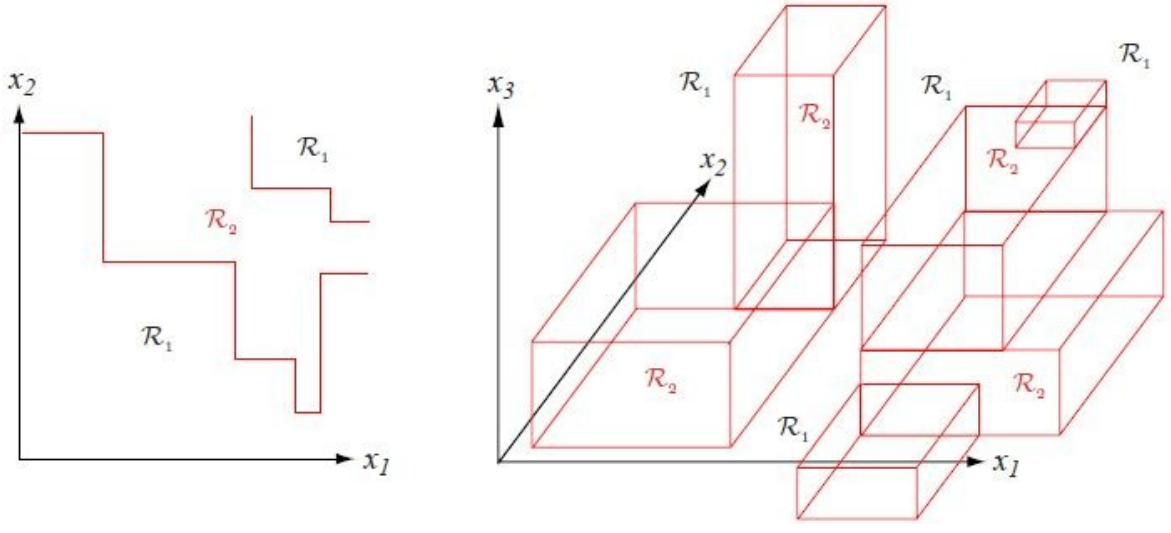
- Число возможных подмножеств, на которые будут расклассифицированы объекты, определяет степень ветвления ДР в данном узле.
  - Любое дерево может быть преобразовано в эквивалентное бинарное ДР.
  - Чем четче разделяются объекты по данному признаку, тем более простое ДР у нас будет.





Деревья решений

- В каждом узле фактически дается ответ на вопрос $x<x_0$?
- Таким образом, ДР рассекает пространство параметров гиперплоскостями на области в которых находятся объекты одного класса.

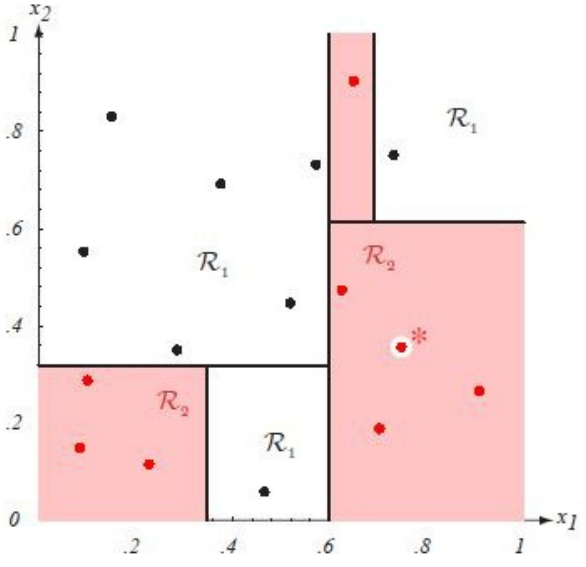


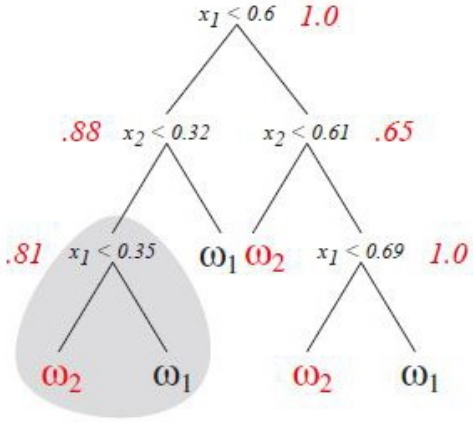
Деревья решений

- Пусть у нас есть два класса объектов ω_1 и ω_2 со значениями параметров x_1 и x_2.

ω_1 (black)	
x_1	x_2
.15	.83
.09	.55
.29	.35
.38	.70
.52	.48
.57	.73
.73	.75
.47	.06

ω_2 (red)	
x_1	x_2
.10	.29
.08	.15
.23	.16
.70	.19
.62	.47
.91	.27
.65	.90
.75	.36* (.32†)

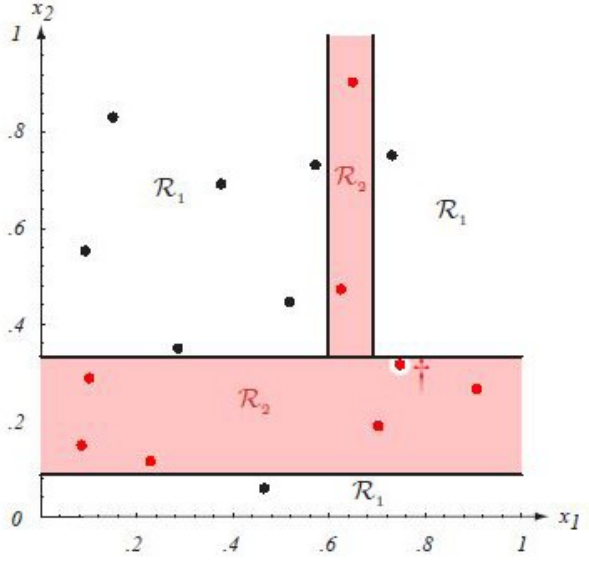


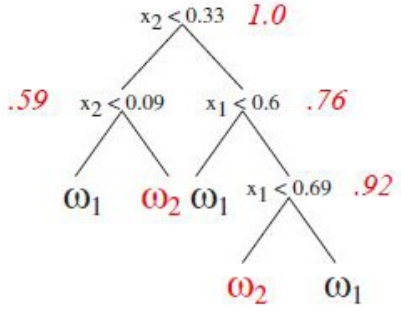


Деревья решений

- Небольшие изменения могут привести к существенной перестройке дерева.

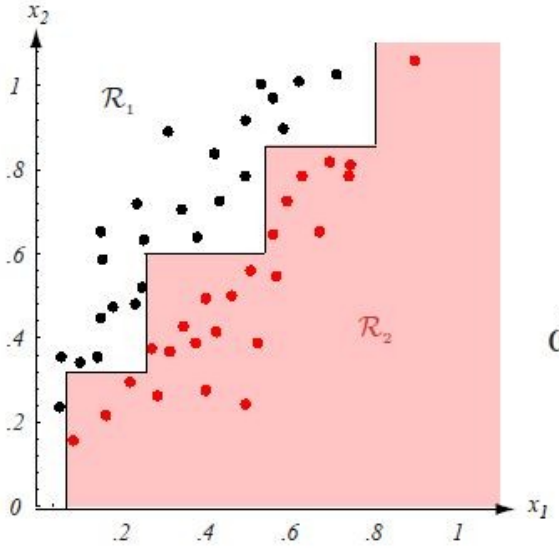
ω_1 (black)		ω_2 (red)	
x_1	x_2	x_1	x_2
.15	.83	.10	.29
.09	.55	.08	.15
.29	.35	.23	.16
.38	.70	.70	.19
.52	.48	.62	.47
.57	.73	.91	.27
.73	.75	.65	.90
.47	.06	.75	.36* (.32†)

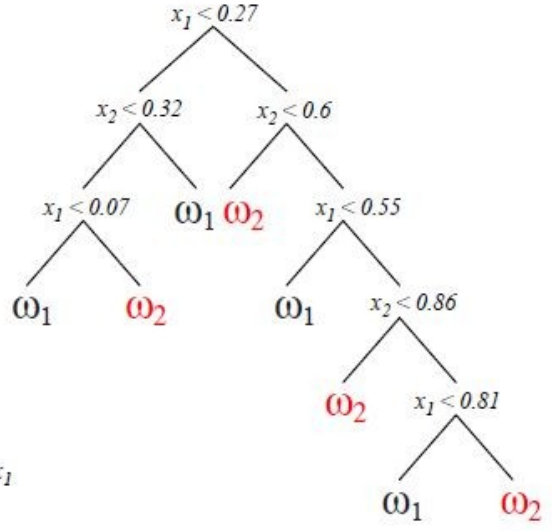




Деревья решений

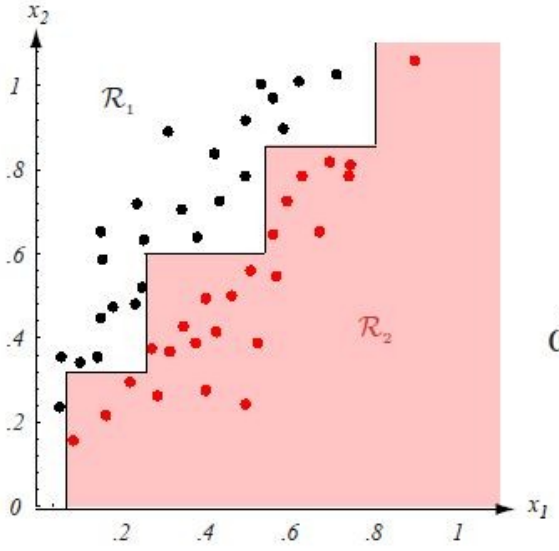
- Необходимо правильно выбирать правила расщепления в узлах деревьев.

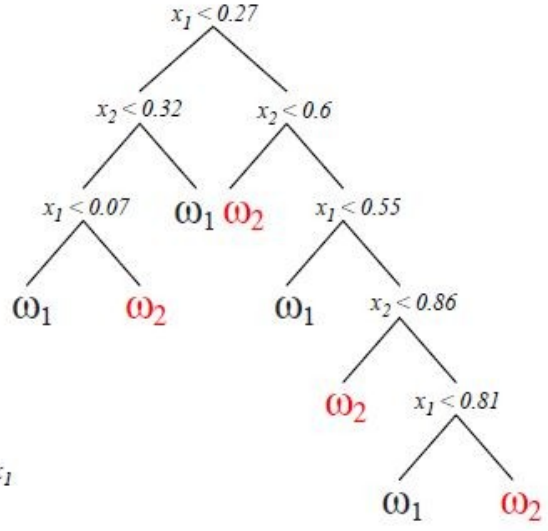


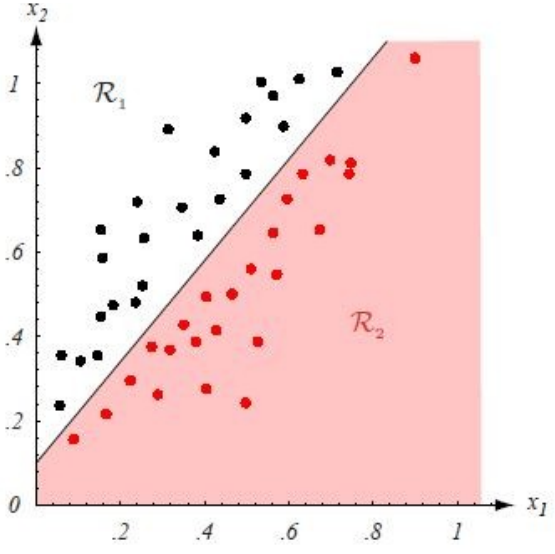


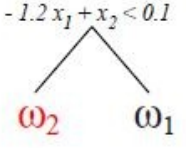
Деревья решений

- Необходимо правильно выбирать правила расщепления в узлах деревьев.



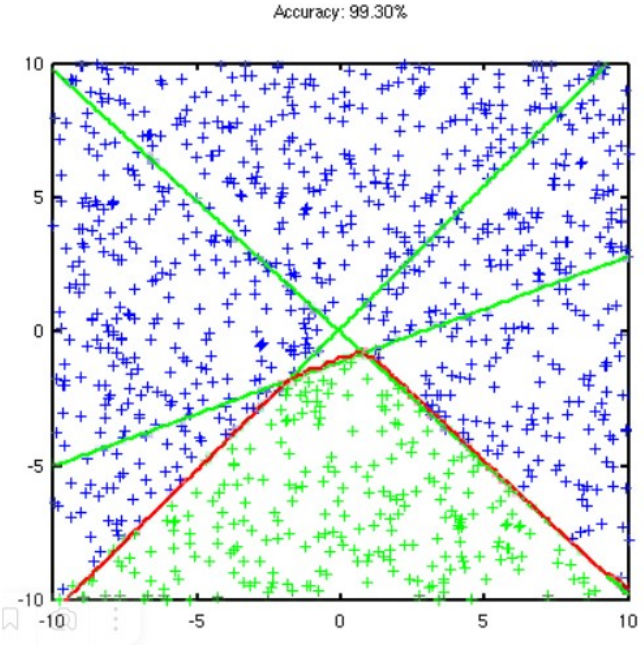


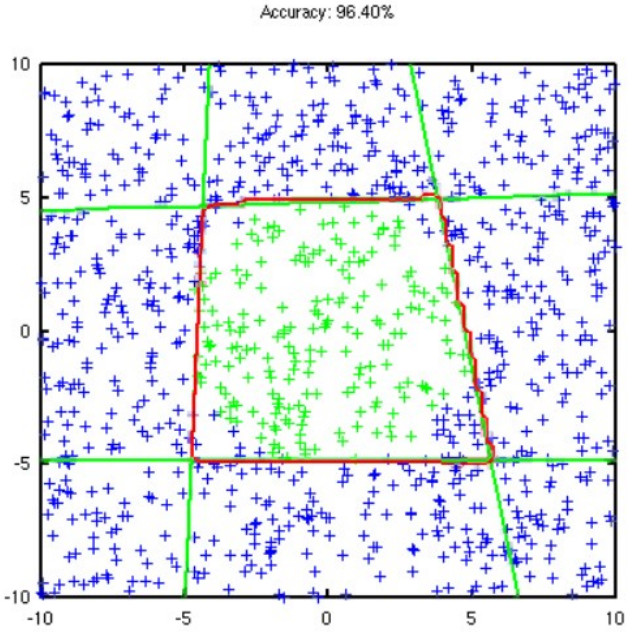




ДР — это набор классификаторов Q.

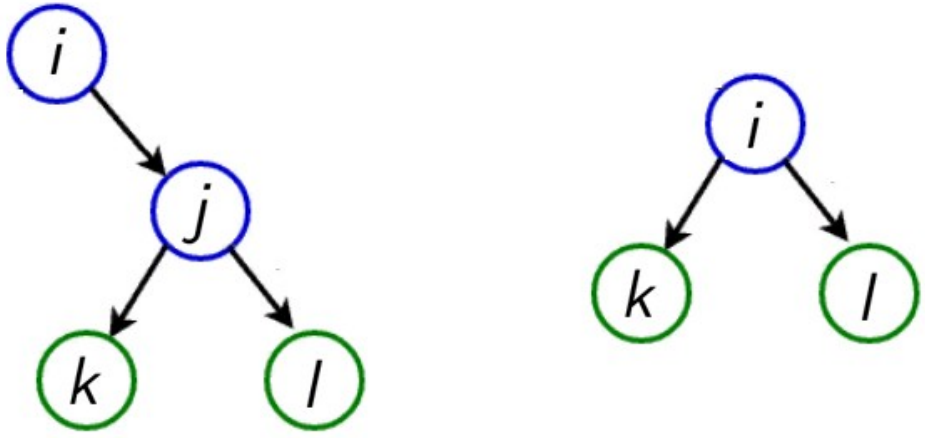
- По мере прохождения дерева множество объектов становится все более однородным.
- Типы классификаторов в узлах дерева
  - координатные: Q: $(x-x_0)<0$ - гиперплоскости, параллельные координатным осям пространства параметров.
  - линейные: Q: $A^t x<0$ - логистическая регрессия
  - нелинейные: Q: $Q(x)<0$





Оптимизация деревьев решений

- Оптимизация выполняется из следующих соображений
  - качество классификации
  - время вычислений
    - классификация
    - обучения



Деревья решений

- До какого уровня ветвим дерево?
  - Ветвление ДР прекращаем как только текущий узел можно объявить листом.
  - Критерий — степень однородности множества объектов.
- S_i — множество объектов в узле i.
- Объявляем вершину i листом если:
  - либо достигнута максимальная глубина дерева d ;
  - либо $|S_i| < N_0$, где N_0 — заданное пороговое значение;
  - либо однородность множества S_i достаточно велика.

ДР. Меры неоднородности

- Мера неоднородности
  - энтропия,
  - дисперсионная неоднородность,
  - неоднородность Джини,
  - неоднородность ошибки классификации.
- Пусть $C^1,..,C^K$ — классы;
  $Si=U_{i=1..K}(S_i^{(k)})$, $S_i^{(k)} \subseteq C^k$;
  S_{ij} — подмножество потомка j узла i , $S_{ij} \subseteq S_i$;
  $N_i = |S_i|$, $N_{ij} = |S_{ij}|$, $N_{ij}^{(k)} = |S_{ij}^{(k)}|$.

  Тогда

  $\Delta Q(S_i) = Q(S_i) - \sum_j (N_{ij}/N_i) * Q(S_{ij}) \rightarrow max$,

  $\Delta Q(S_i) \rightarrow max$, т.е уменьшаем неопределенность в узле i

Энтропия

- Пусть

  $P_k = N_i^{(k)}/N_i$ — частота.

  Тогда

  $H(S_i) = -\sum_k P_k \log P_k$

- Причем можно использовать любое основание больше 1 и

  $P_k = 0 \Rightarrow P_k \log P_k = 0$.

- Максимизируется разность неопределённости исходной подвыборки S_i и средней неопределённости подвыборок S_{ij} .

  $\Delta Q = H(S_i) - \sum_j (N_{ij}/N_i) H(S_i) \rightarrow \max_\theta$

  $H(S_i) = - \sum_k (N_i^{(k)}/N_i) \log(N_i^{(k)}/N_i)$

Выбор параметров классификаторов

- Выбор θ_i классификатора $h(x, \theta_i)$
  - категориальный признак: перебор возможных значений
  - порядковый признак: градиентный спуск
- если оптимальное значение не единственно ($\theta_l \le \theta_i \le \theta_u$), то берём среднее:

  $\theta_i := (\theta_l + \theta_u)/2;$
- взвешенное среднее:

  $\theta_i := (N_i^{(1)}/N_{i)}\ \theta_l + (N_i^{(2)}/N_{i)}\ \theta_u$.

Меры неоднородности

Пусть $P_k = N_i^{(k)}/N_i$ — частота

- Энтропия:

  $Q_1(S_i) = H(S_i) = -\sum_k P_k \log P_k$

- Неоднородность Джини (Gini impurity):

  $Q_2(S_i) = \sum_{i \neq j} P_i P_j = 1 - \sum_i P_i^2$

- Неоднородность ошибочной классификации (misclassification impurity):

  $Q_2(S_i) = 1 - \max_i P_i$.

Меры неоднородности

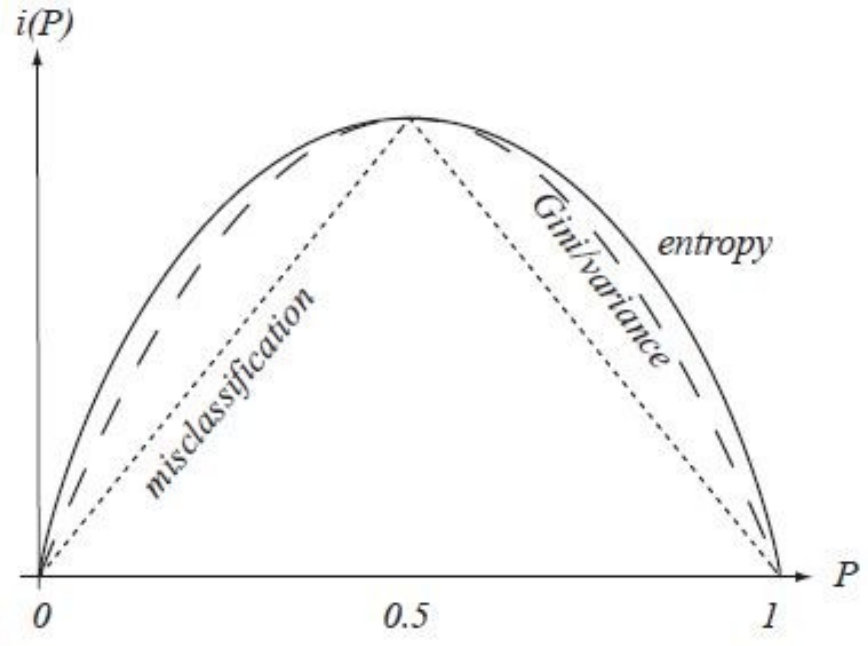


Figure 8.4: For the two-category case, the impurity functions peak at equal class frequencies and the variance and the Gini impurity functions are identical. To facilitate comparisons, the entropy, variance, Gini and misclassification impurities (given by Eqs. 1 – 4, respectively) have been adjusted in scale and offset to facilitate comparison; such scale and offset does not directly affect learning or classification.

Глубина дерева

- Ветвить до тех пор, пока не будет достигнута минимальная неоднородность ⇒ ПЕРЕОБУЧЕНИЕ
  - В наихудшем случае: каждый лист — один элемент выборки
  - просто дерево поиска — обобщения не происходит
- Если ветвление остановить слишком рано ⇒ ошибка классификации слишком велика — НЕДООБУЧЕНИЕ
- Критерии остановки ветвления:
  - минимизация ошибки на тестовой (валидационной) выборке;
  - достижение порогового значения неоднородности;
  - достижение порогового значения на мощность подвыборки;
  - минимизация неоднородности с регуляризацией.

Когда останавливать ветвление

- Минимизация ошибки на тестовой (валидационной) выборке

  Ветвим до тех пор пока не минимизирована ошибка валидирования:

  - производим обучение (построение дерева) до определённого уровня на подвыборке;
  - оставшуюся часть выборки используем для валидирования (тестирования) полученного на текущий момент дерева.

- Достижение порогового значения неоднородности

  Останавливаем ветвление как только

  - $\max_{\theta} i\ \Delta Q(i) \leq \beta$
  - используются все данные для обучения — дерево может быть несбалансированным — листья могут находиться на разных уровнях.

Когда останавливать ветвление

- Достижение порогового значения на мощность подвыборки

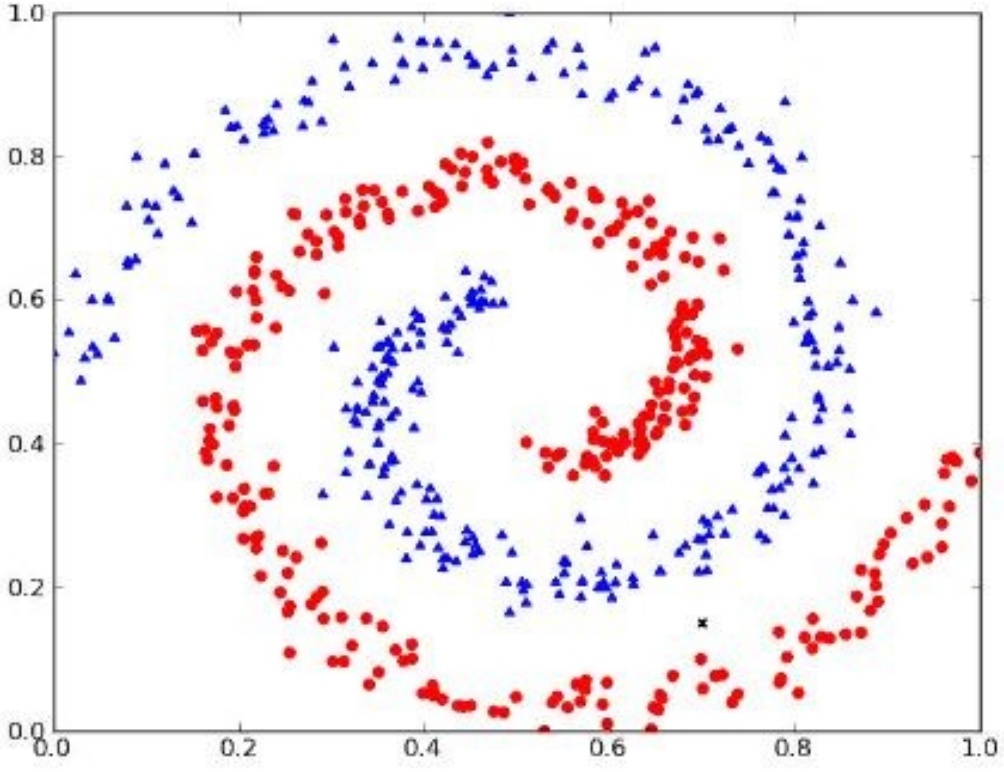
  Останавливаем ветвление, когда $|S_i| \leq M$.

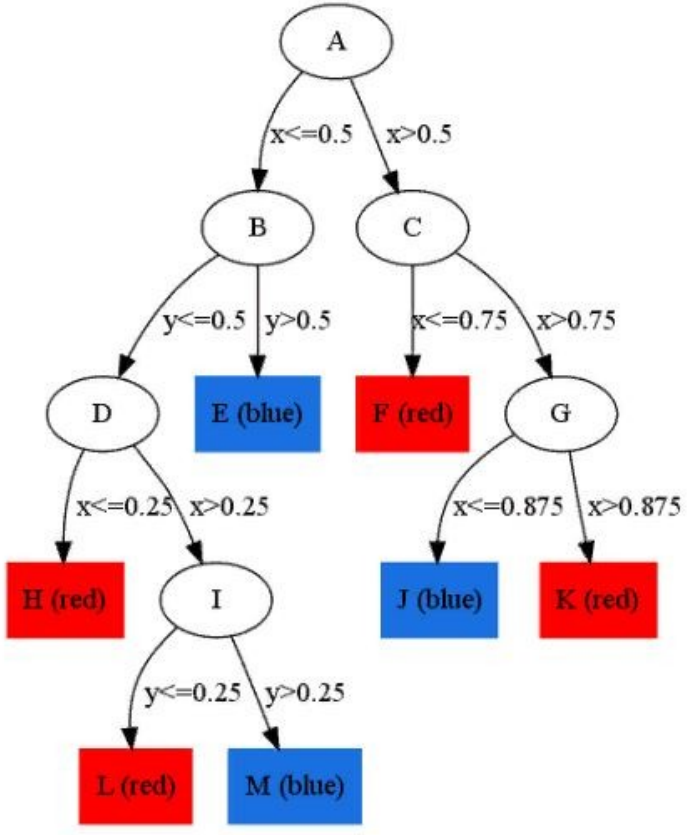
  - Benefit: мелкое разбиение там, где данных много (плотность большая), крупное — где мало (плотность небольшая).
- Минимизация неоднородности с регуляризацией
- Останавливаем ветвление, когда будет достигнут минимум

  $\lambda \cdot size + \sum_{LN} Q(i) \leq \beta,$

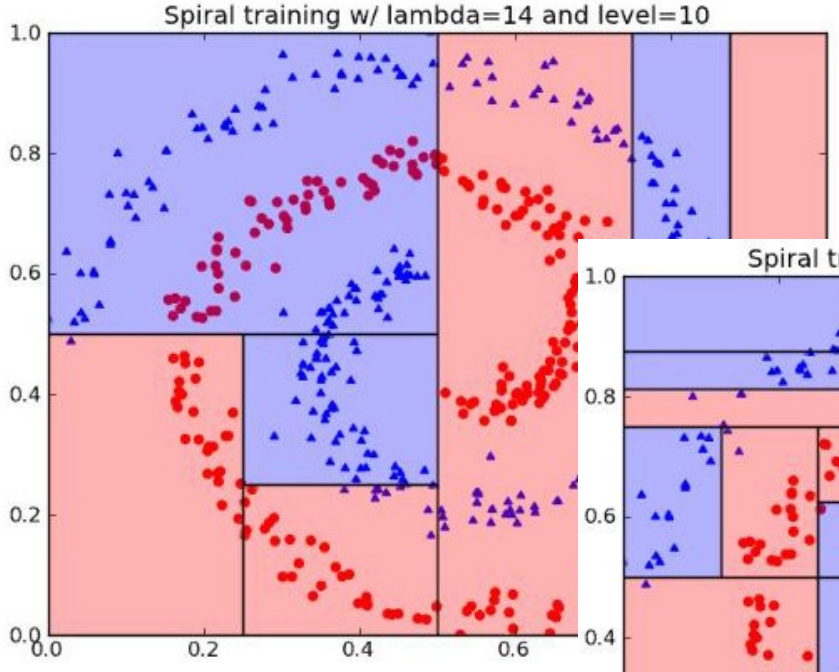
  - $\lambda > 0$ — регуляризационный коэффициент,
  - size — количество рёбер или вершин.

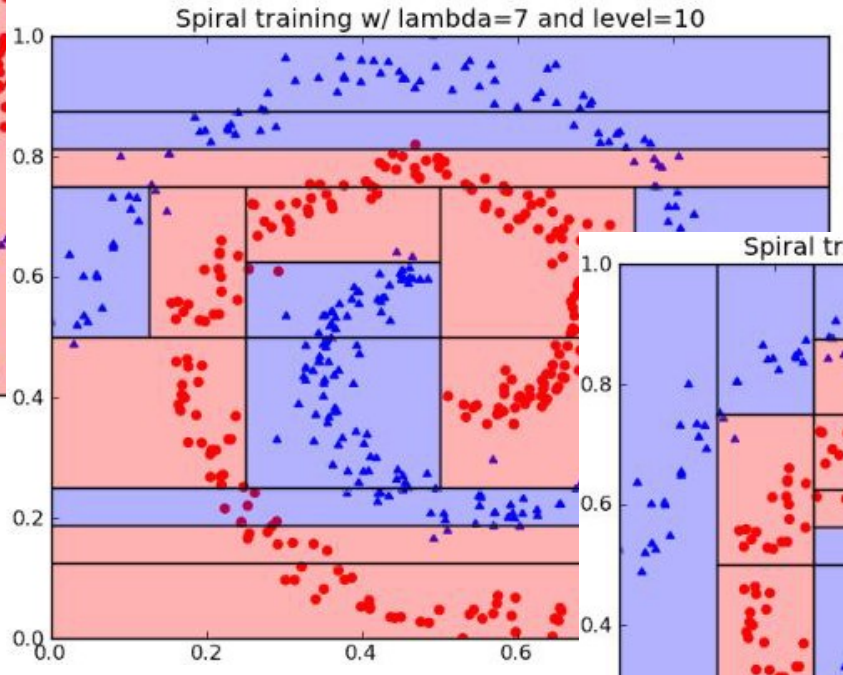
Когда останавливать ветвление

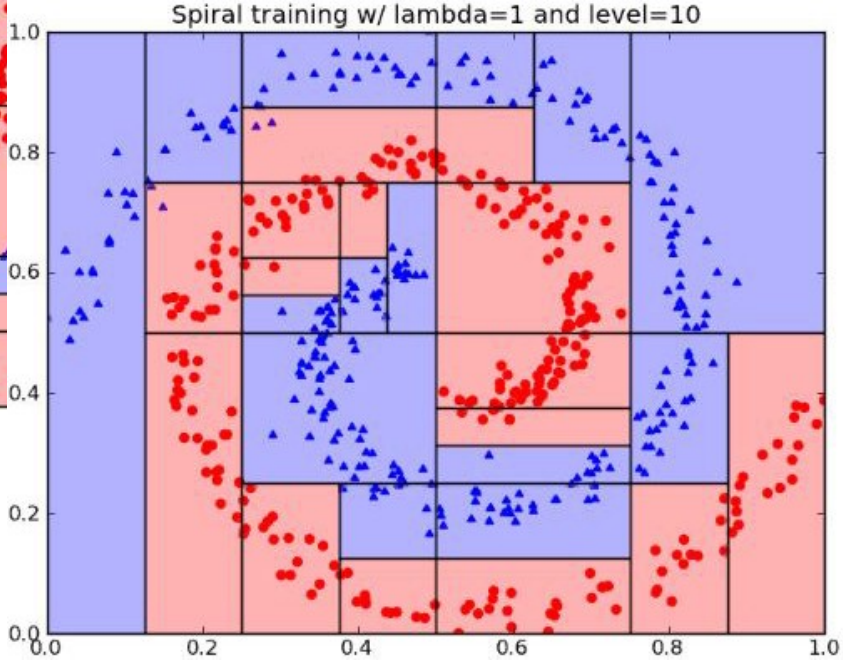




Когда останавливать ветвление







Материалы лекций:

(https://theory.sinp.msu.ru/doku.php/ml_lectures)