Решающие деревья и случайные леса (Binary decision trees & Random forests)

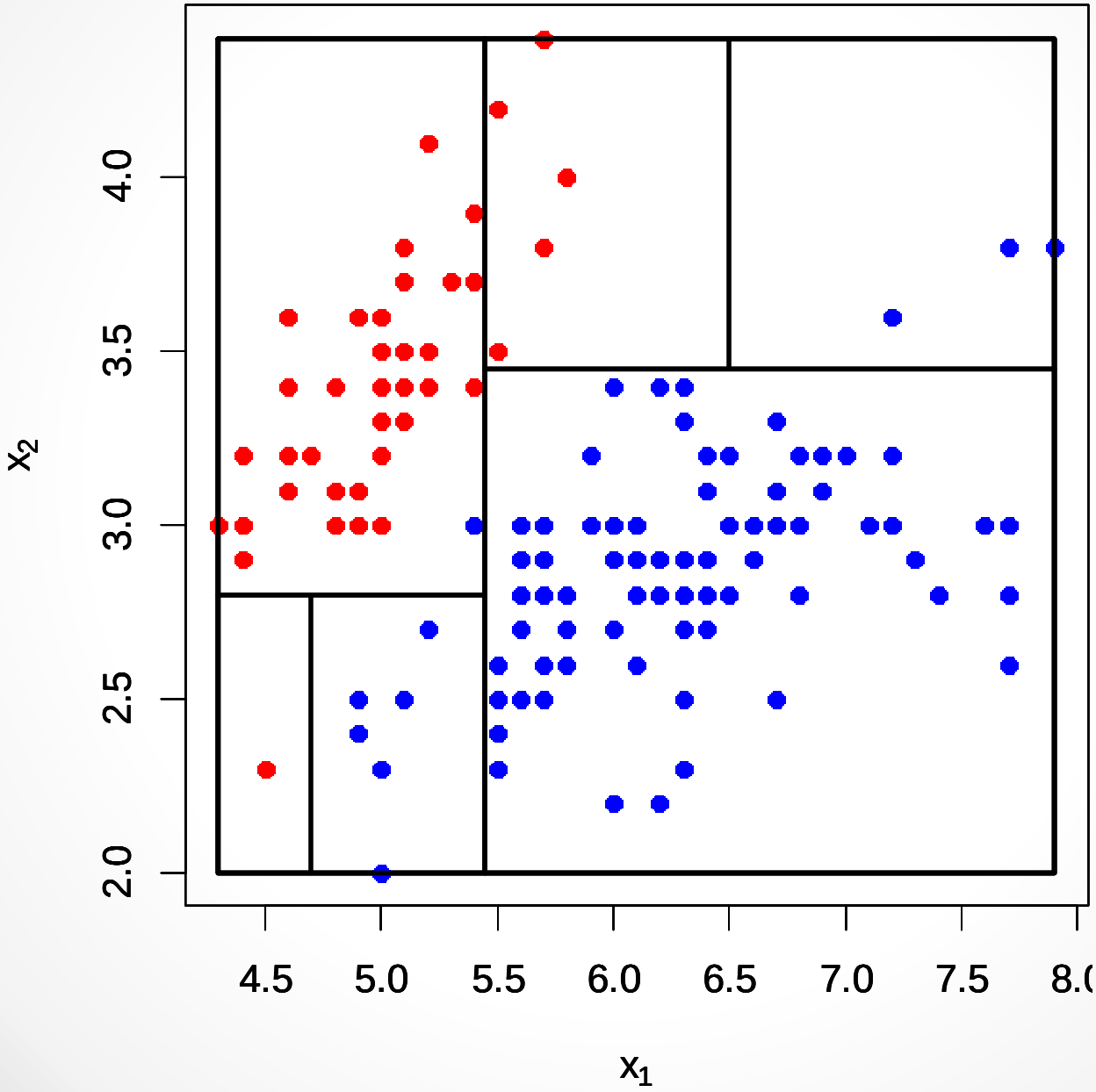
ЦМФ

Основная идея

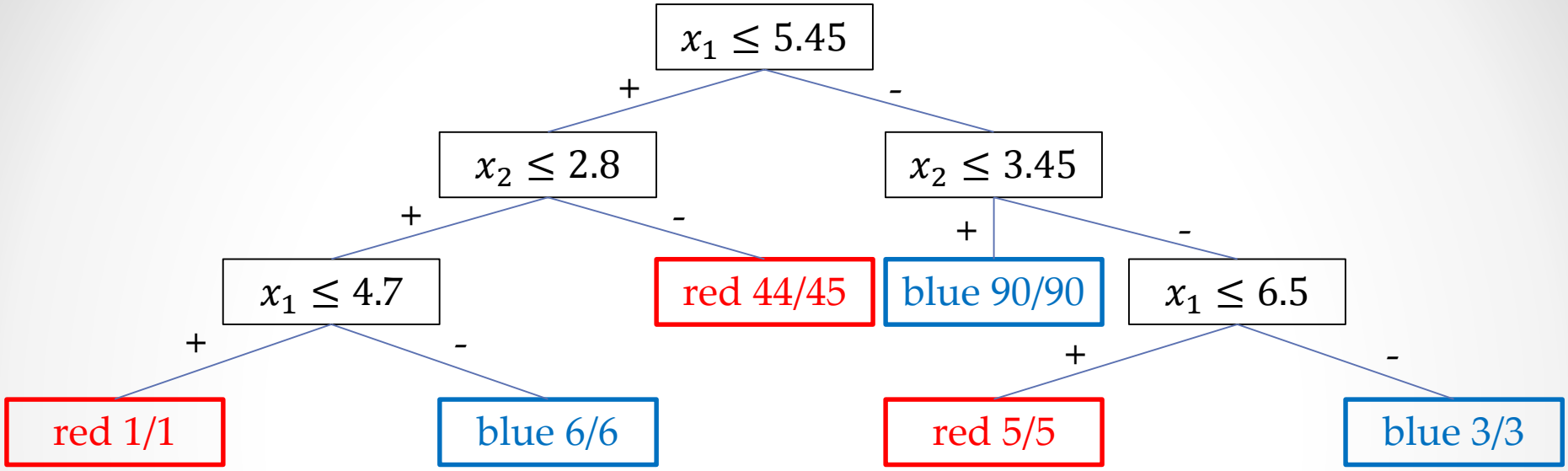
Классификация наблюдений на основе последовательного применения критериев (>, <, ∈) к тому или иному признаку

При этом пространство $R \supset X$ рекурсивно разделяется гиперплоскостями, параллельными оси одного из признаков, до тех пор, пока в каждой из получившихся областей не образуется значительное большинство наблюдений одного класса

Классификация бинарными деревьями



Дерево решений



Основные понятия

Пусть условие $x_j \leq v$ разделяет пространство R на 2 части: R_Y и R_N, тогда множество наблюдений $D = \left\{\vec{x}^{(i)}, i \in \{1; \dots; m\}\right\}$ также разделяется на $D_Y = \left\{\vec{x}^{(i)} \in D: x_{i,j} \leq v\right\}$ и $D_N = \left\{\vec{x}^{(i)} \in D: x_{i,j} > v\right\}$

Однородность j-й области: $purity(D_j) = \max\limits_k \frac{m_{j,k}}{m_j}$, где

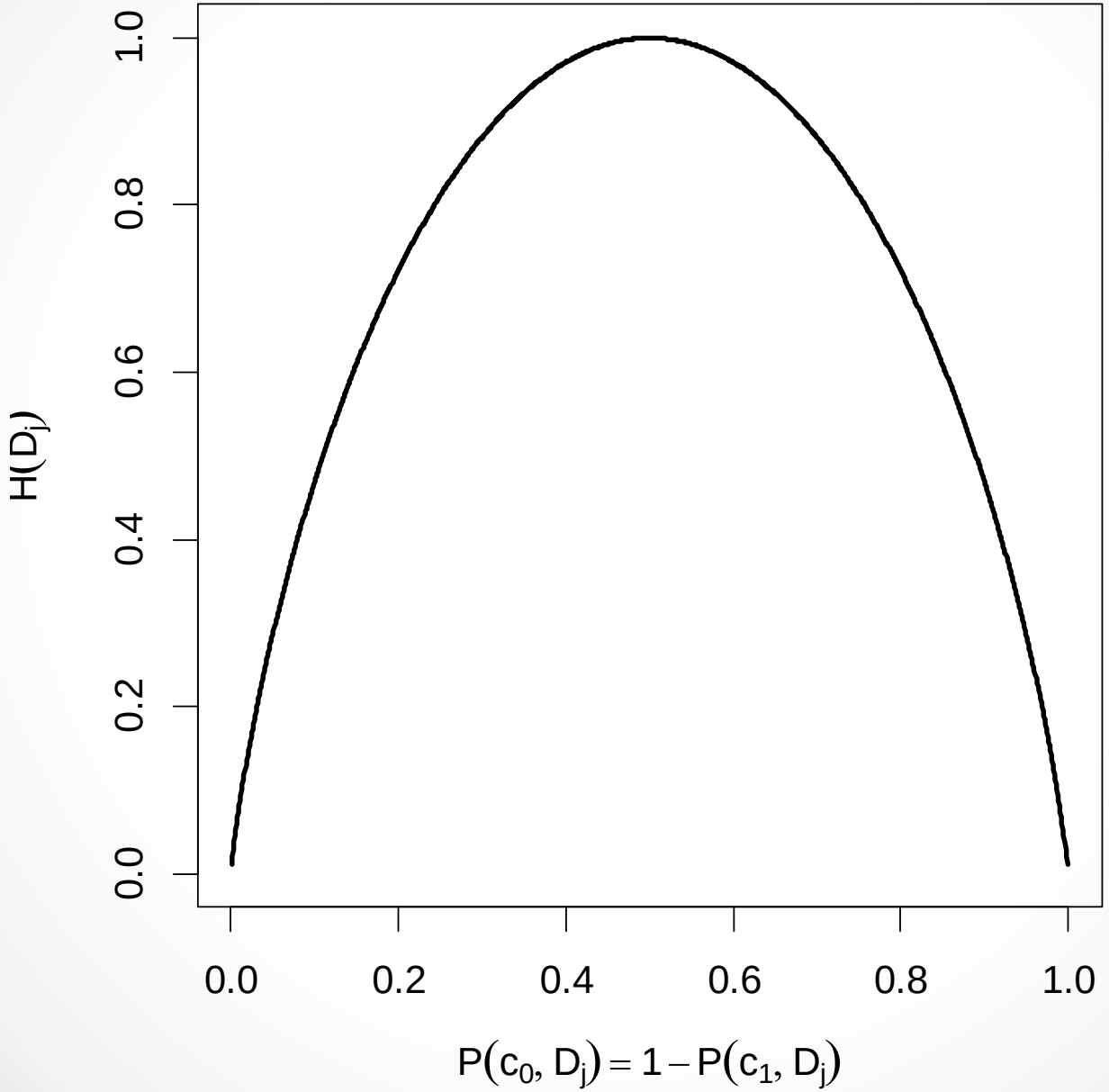
$m_{j,k} = \sum_{i=1}^{m} \boldsymbol{I}\left(\vec{x}^{(i)} \in D_j, y_i = c_k\right)$, $m_j = \sum_{i=1}^{m} \boldsymbol{I}\left(\vec{x}^{(i)} \in D_j\right) = \sum_{k=1}^{K} m_{j,k}$

Энтропия области: $H(D_j) = -\sum_{k=1}^{K} P(c_k|D_j) \log_2 P(c_k|D_j)$, где

$P(c_k|D_j) = \frac{m_{j,k}}{m_j}$ — вероятность нахождения наблюдения k-го класса в области D_j

Энтропия разделения: $H(D_Y, D_N) = \frac{m_Y}{m} H(D_Y) + \frac{m_N}{m} H(D_N)$

Энтропия



Оценка эффективности разделения

Информативность (сокращение энтропии):

$$gain(D, D_Y, D_N) = H(D) - H(D_Y, D_N)$$

Вместо энтропии можно использовать коэффициент Джини:

$$G(D) = 1 - \sum_{k=1}^{K} P^2(c_k|D), \;\; G(D_Y, D_N) = \frac{m_Y}{m} G(D_Y) + \frac{m_N}{m} G(D_N)$$

Если величины $x_{i,j}$ непрерывны, то в качестве кандидатов на точки разделения рассматриваются середины интервалов между последовательными уникальными значениями

Если $x_{i,j}$ дискретны, то каждое их уникальное значение рассматривается как возможная точка разделения

Случайный лес

Случайный лес представляет собой совокупность моделей — бинарных деревьев решений, — отличающихся случайным выбором экзогенных параметров

Такими параметрами могут быть: выбор точек разделения областей, выбор набора обучающих наблюдений из тренировочной совокупности и др.

Прогнозным значением в задачах классификации может являться наиболее часто встречающийся номер класса среди прогнозов по деревьям, составляющим лес

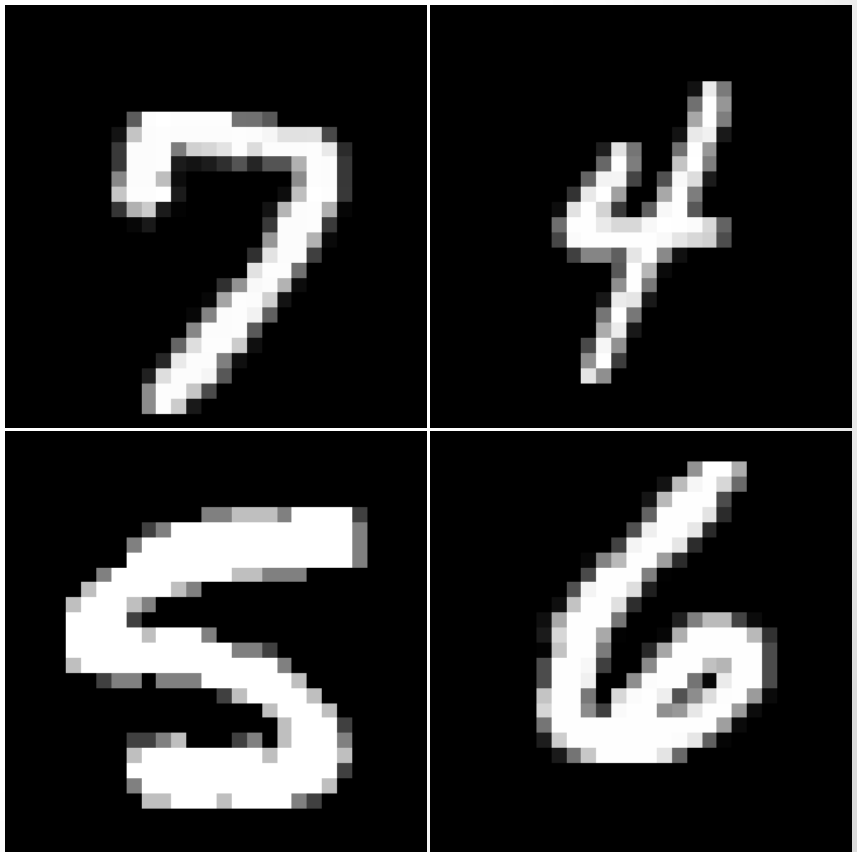
Случайный лес в Python

```
import numpy as np
import pandas as pd
from sklearn.ensemble import RandomForestClassifier as RFC

# загрузка исходных данных из файла
X = pd.read_csv('train_data_file.csv', header=0).as_matrix()

# обучение модели
rf = RFC(n_estimators=15) # количество деревьев = 15
rf.fit(X=X[:, 1:], y=X[:, 0]) # целевая переменная - в первом столбце

# предсказания
X_test = pd.read_csv('test_data_file.csv', header=0).as_matrix()
test_classes = rf.predict(X_test) # номера классов
test_probs = rf.predict_proba(X_test) # вероятности принадлежности
```

Домашнее задание

Классифицировать рукописные цифры из папки «digits_test»

Ответы принимаются на
https://kaggle.com/join/cmf_digits



Полезные функции

```
# список файлов в каталоге mypath
from os import listdir
from os.path import isfile, join
onlyfiles = [f for f in listdir(mypath) if isfile(join(mypath, f))]

# преобразование изображения в numpy-матрицу
import numpy as np
import cv2
img = cv2.imread('file_full_name.png', 0)
```