

# Байесовский классификатор

# Текстовые задачи

- Классификация текста
- Определение авторства
- Определение эмоциональной окраски текста
- Парсинг научных статей

# Байесовская постановка задачи

$y$  – класс,  $x$  – документ

$$P(y|x) = \frac{P(y)P(x|y)}{P(x)}$$

$$y_{MAP} = \arg \max_{y \in Y} P(y|x) = \arg \max_{y \in Y} \frac{P(y)P(x|y)}{P(x)} = \arg \max_{y \in Y} P(y)P(x|y)$$

$$\arg \max_{y \in Y} P(y)P(x|y) = \arg \max_{y \in Y} P(x_1, x_2, \dots, x_n|y)P(y)$$

Наивное предположение (о независимости):

$$P(x_1, x_2, \dots, x_n|y) = P(x_1|y) P(x_2|y) P(x_3|y) \dots P(x_n|y)$$

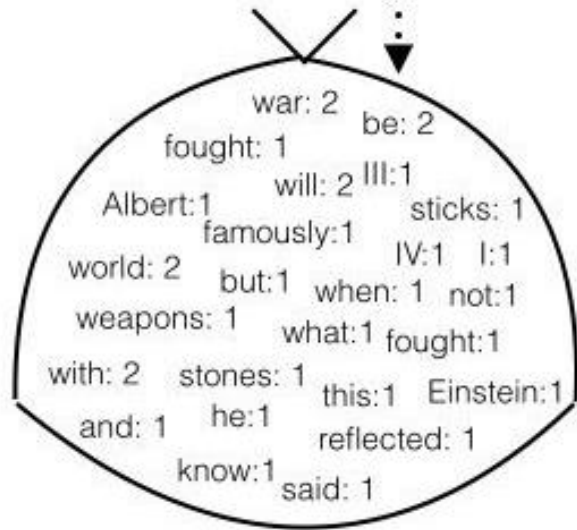
# Bag-of-Words (BoW)

Важно только набор слов, не важен порядок.

$x_1, x_2, \dots, x_n$  - количества слов из словаря длины  $n$  в документе.

Словарь может быть как полный, так и ограниченный.

Albert Einstein reflected this when he famously said, "I know not with what weapons World War III will be fought, but World War IV will be fought with sticks and stones."



Bag of words

Albert Einstein reflected this when he famously said , " I know not with what weapons World War III will be fought , but World War IV will be fought with sticks and stones . "

Unigrams

Albert Einstein	know not	but World
Einstein reflected	not with	World War
reflected this	with what	War IV
this when	what weapons	IV will
when he	weapons World	will be
he famously	World War	be fought
famously said	War III	fought with
said ,	III will	with sticks
"	will be	sticks and
" I	be fought	and stones
I know	fought ,	stones .
	, but	,"

Bigrams

# Наивный байесовский классификатор

$$P(x_1, x_2, \dots, x_n | y) = P(x_1 | y) P(x_2 | y) P(x_3 | y) \dots P(x_n | y)$$

$$y_{MAP} = \arg \max_{y \in Y} P(y) P(x | y)$$

$$y_{NB} = \arg \max_{y \in Y} P(y) \prod_i P(x_i | y)$$

# Слова в документах

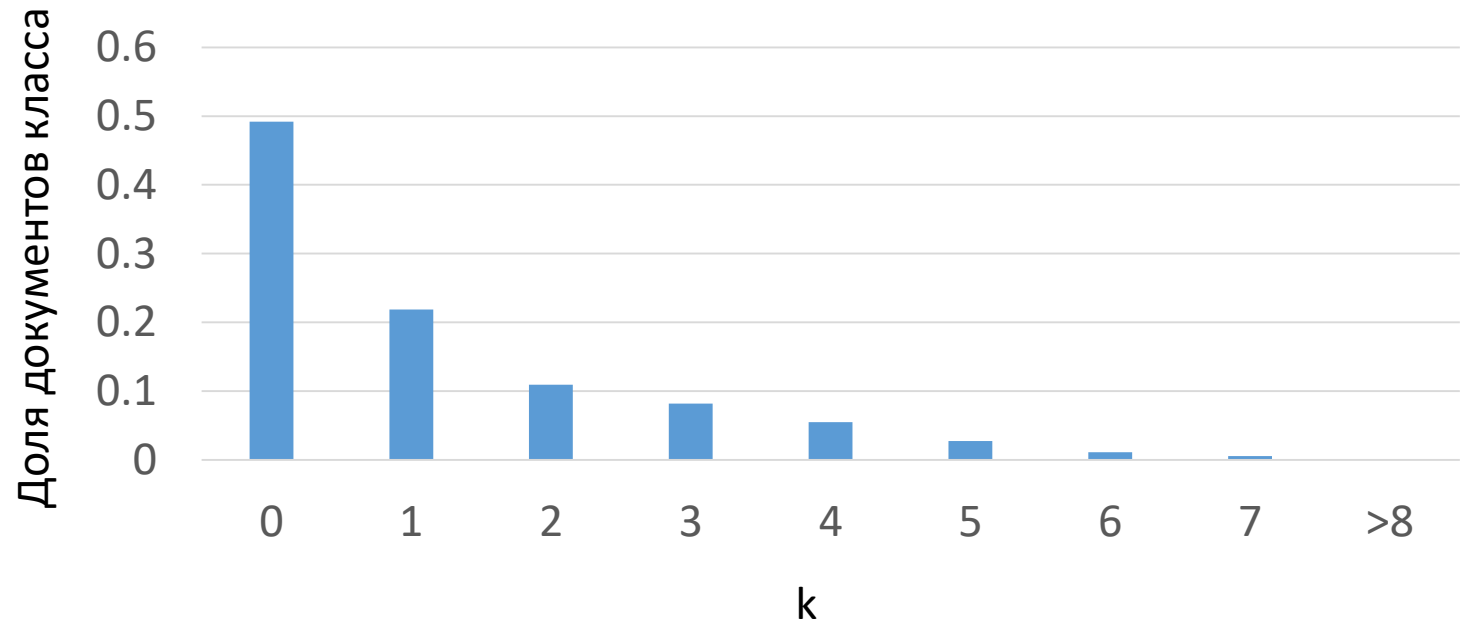
$P(y)$  – частота класса  $y$ .

$P(x_i|y)$  – вероятность значения признака  $x_i$  в классе  $y$ , например доля документов в классе, в которых определенное слово встречается  $k$  раз.

$$P(x_i|y) = \frac{\text{count}(x_i, y_j)}{\text{count}(x_i)}$$

Проблема –  $\text{count}(x_i, y_j) = 0$

$$\hat{P}(x_i|y) = \frac{\text{count}(x_i, y_j) + \alpha}{\text{count}(x_i) + \alpha K}$$



# Обработка текста

- Уменьшение словаря:
  - 35, 535, 17, 200000  $\rightarrow$  \$number
  - $(5+3), \frac{1}{2}w^T w + C \rightarrow$  \$formula
  - Stemming – приведение слова в инфинитивную форму (не всегда работает хорошо)
- Повышение веса:
  - Слова в названии документа (гиперссылка, подписях к картинкам)
  - Слова в предложениях, которые содержат слова из названия
  - Первое предложение в каждом абзаце



# Байесовский классификатор с другими признаками

Бернулли:

$$P(x_i|y) = P(x_i = 1|y)x_i + (1 - P(x_i = 1|y))(1 - x_i), \quad x_i \in \{0,1\}$$

Распределение Гаусса:

$$p(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} e^{-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}}$$

# Оценка распределения

Простой вариант (для домашки, например) – выборочное среднее и дисперсия для  $\mu$  и  $\sigma$ .



Более сложный вариант – EM (Expectation-maximization) со смесью Гауссиан (Gaussian mixture).

# Expectation-maximization (EM)

Смесь K Гауссиан задается параметрами:

$\mu_k$  — вектор среднего,  $\Sigma_k$  — матрица ковариаций

$\alpha_k$  — "вес" гауссианы, вероятность того, что случайная точка принадлежит к Гауссиане  $k$

$$\sum \alpha_k = 1$$

Принадлежность объекта  $x_i$  к  $k$ -му распределению :

$$w_{ik} = p(\mu_k, \Sigma_k | x_i) = \frac{p(x_i | \mu_k, \Sigma_k) \cdot \alpha_k}{\sum_j p(x_i | \mu_j, \Sigma_j) \cdot \alpha_j}$$

**E-Step:** считаем  $w_{ik}$

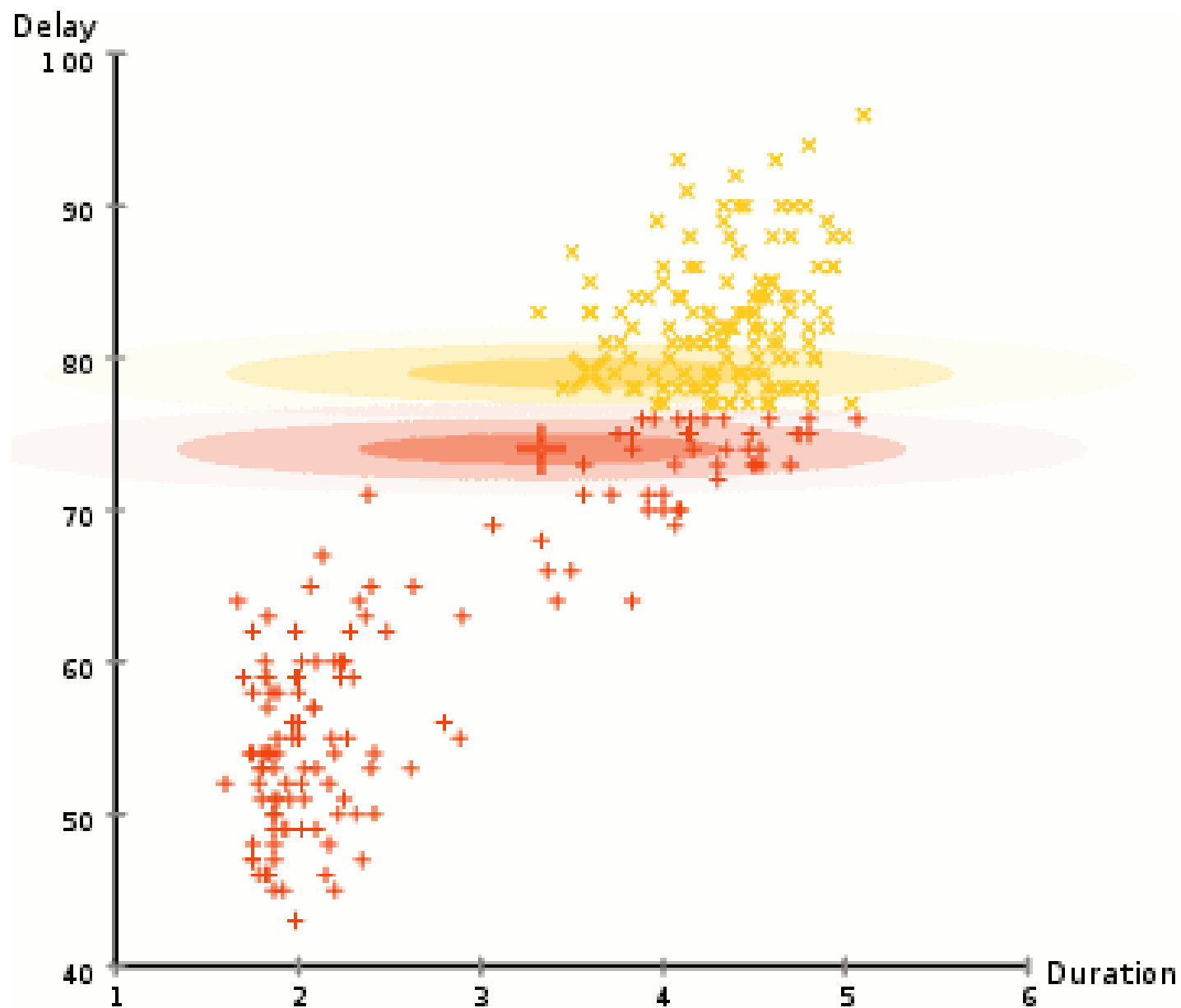
**M-Step:**

$$\alpha_k^{new} = \frac{\sum_{i=1}^N w_{ik}}{N} = \frac{N_k}{N}$$

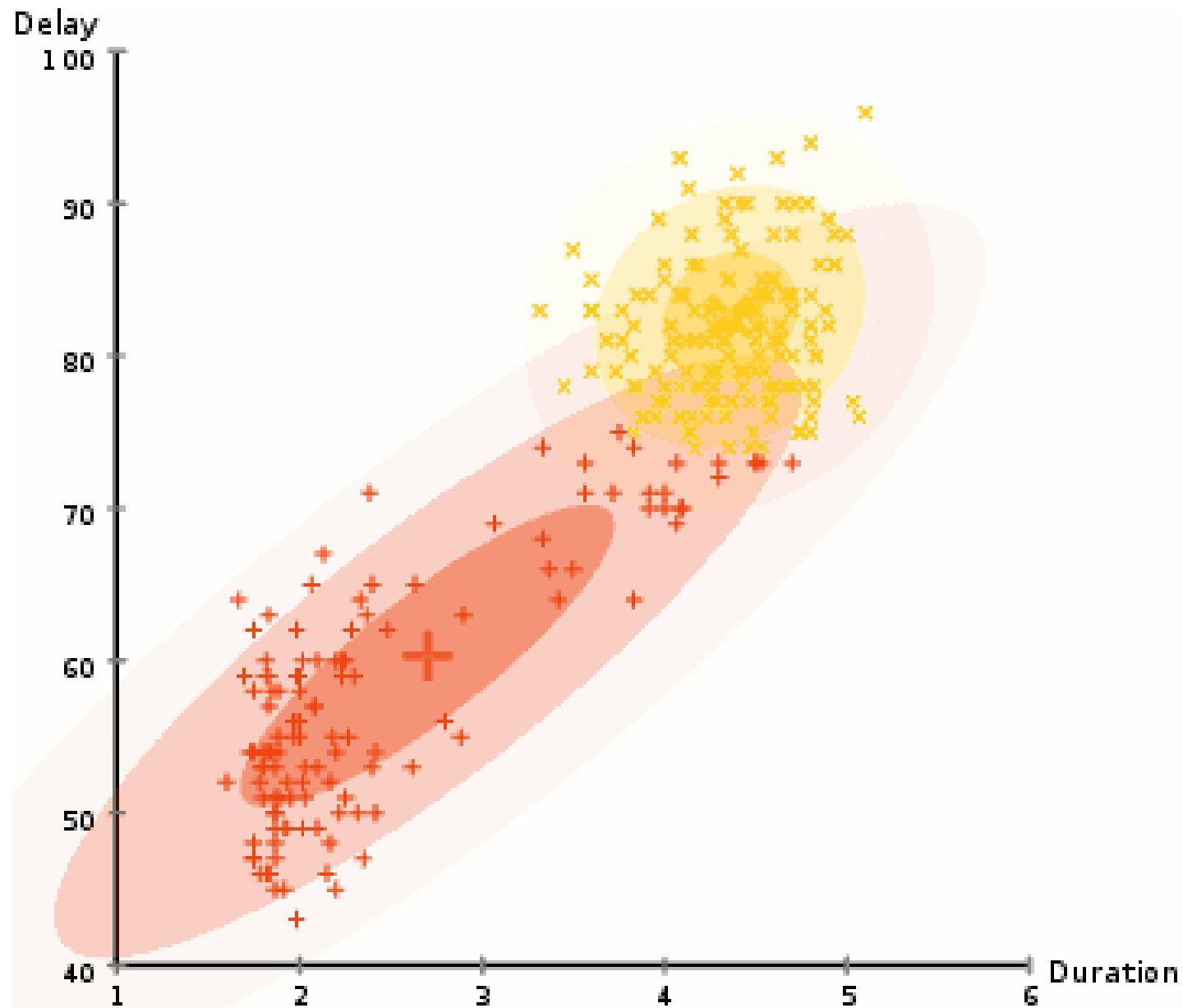
$$\mu_k^{new} = \left( \frac{1}{N_k} \right) \sum_{i=1}^N w_{ik} \cdot x_i$$

$$\Sigma_k^{new} = \left( \frac{1}{N_k} \right) \sum_{i=1}^N w_{ik} \cdot (x_i - \mu_k^{new})(x_i - \mu_k^{new})^T$$

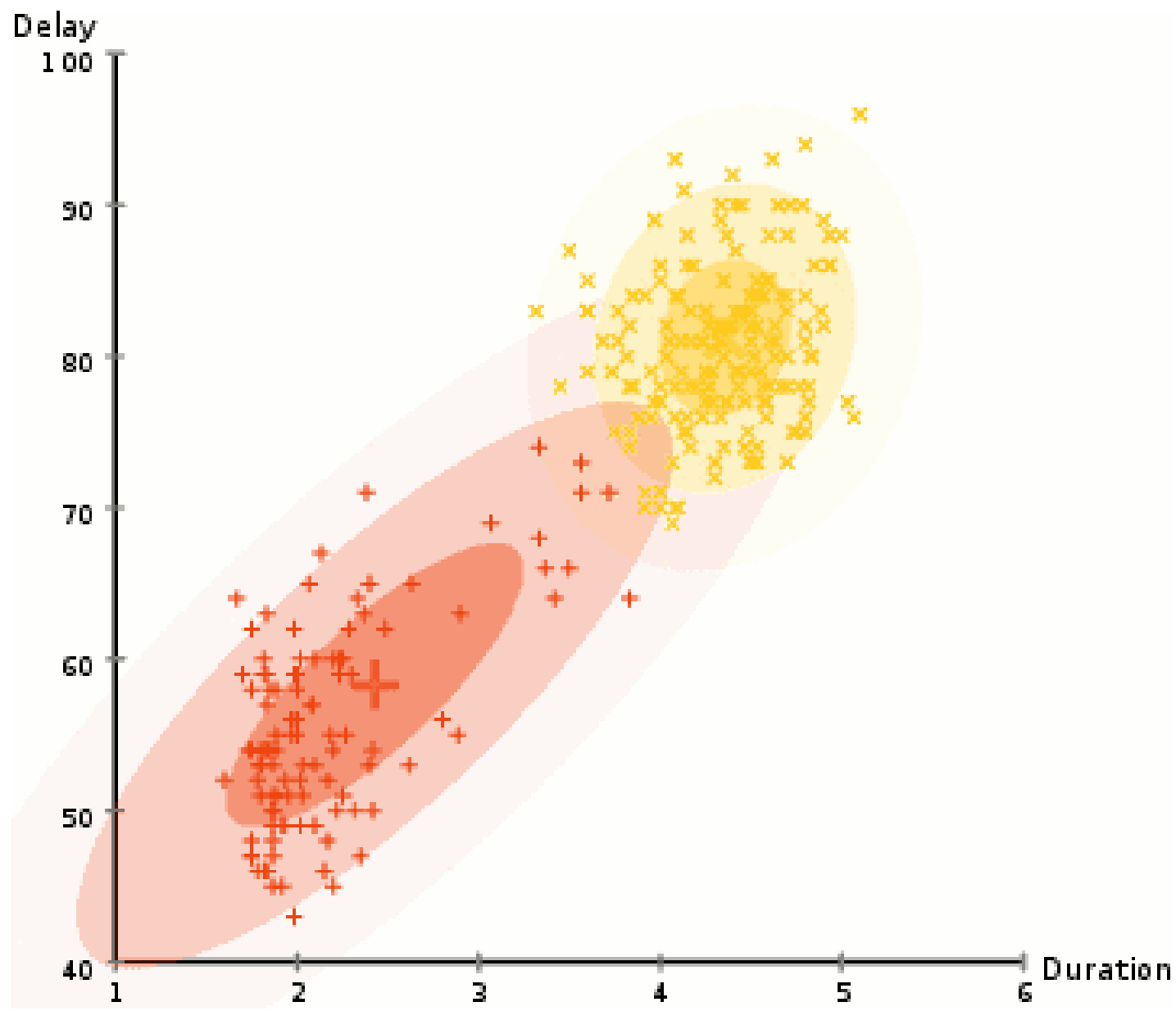
# EM Clustering



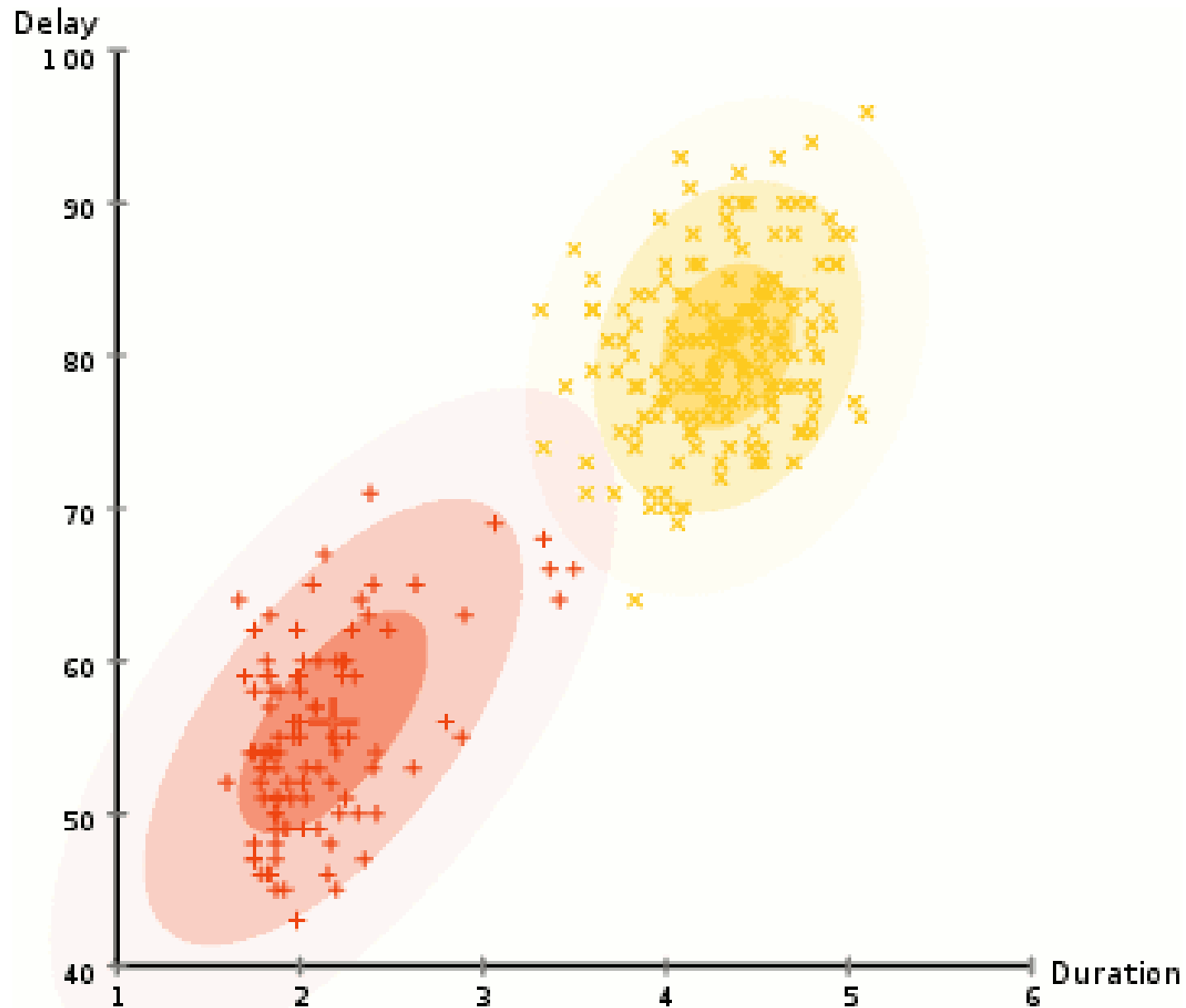
# EM Clustering



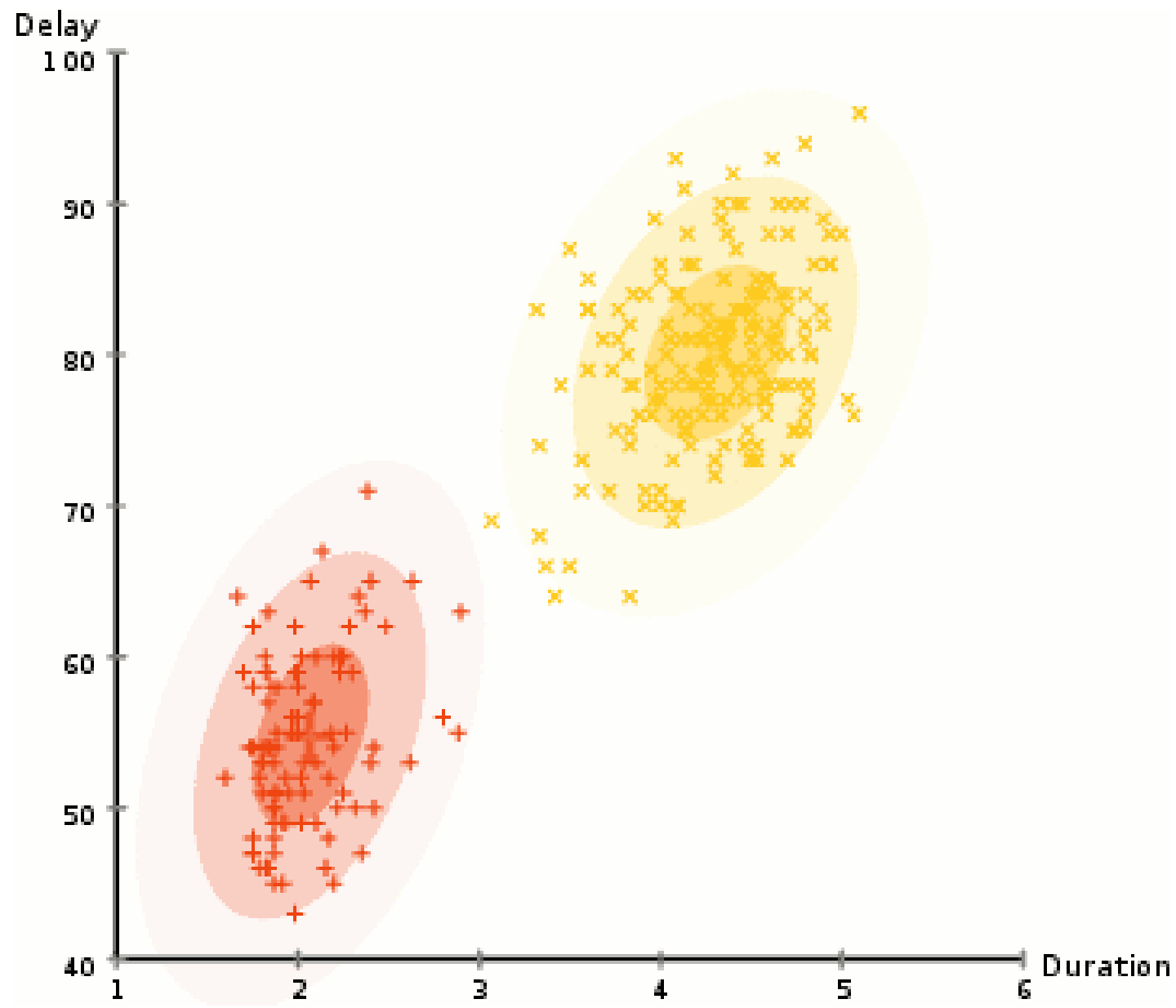
# EM Clustering



# EM Clustering

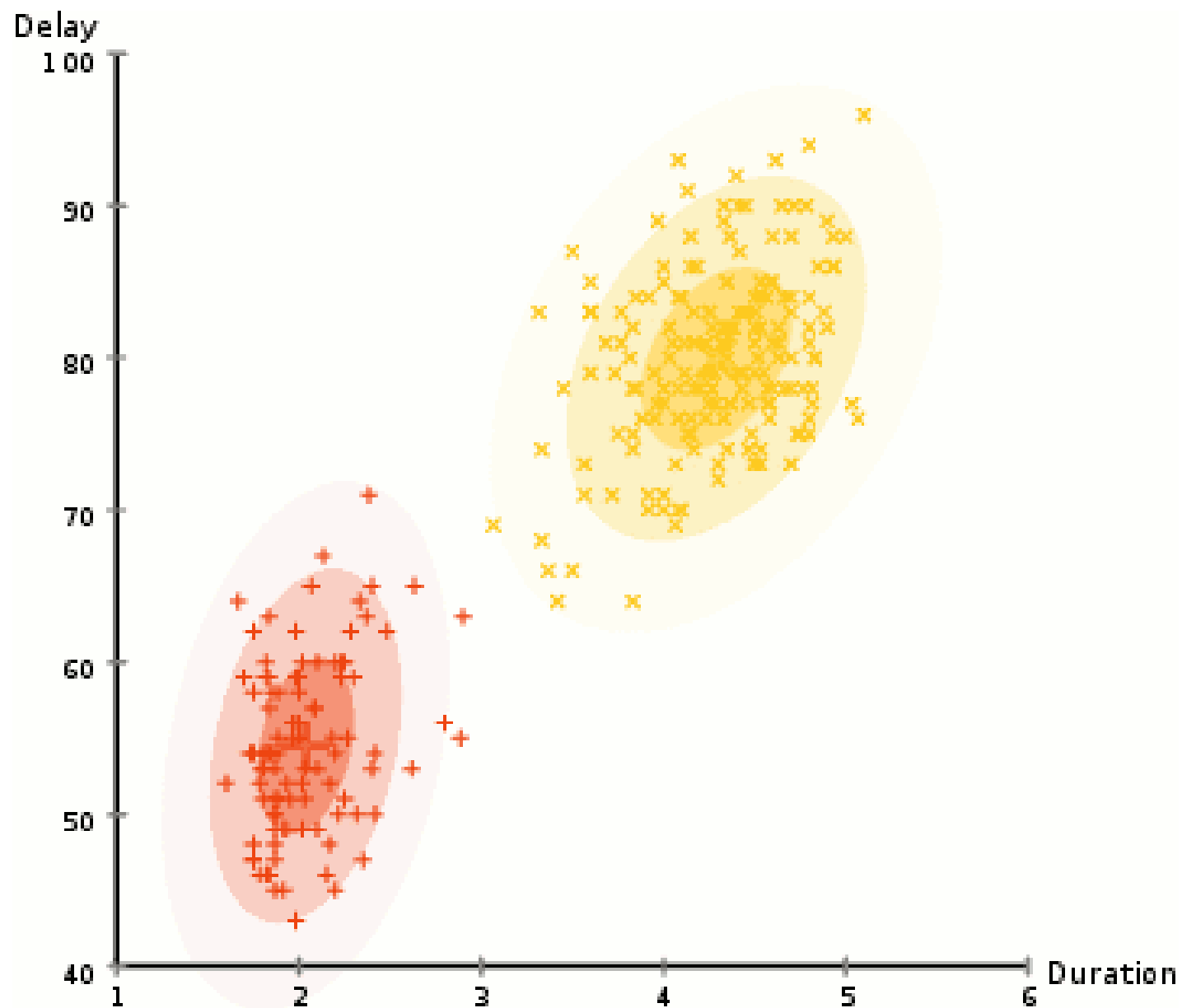


# EM Clustering





# EM Clustering



Naïve Bayes – отличный Baseline!