# Байесовский классификатор

# Текстовые задачи

- Классификация текста

- Определение авторства

- Определение эмоциональной окраски текста

- Парсинг научных статей

# Байесовская постановка задачи

$y$ – класс, x - документ $\qquad P(y|\mathrm{x}) = \dfrac{P(y)P(\mathrm{x}|y)}{P(\mathrm{x})}$

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

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

Наивное предположение (о независимости):

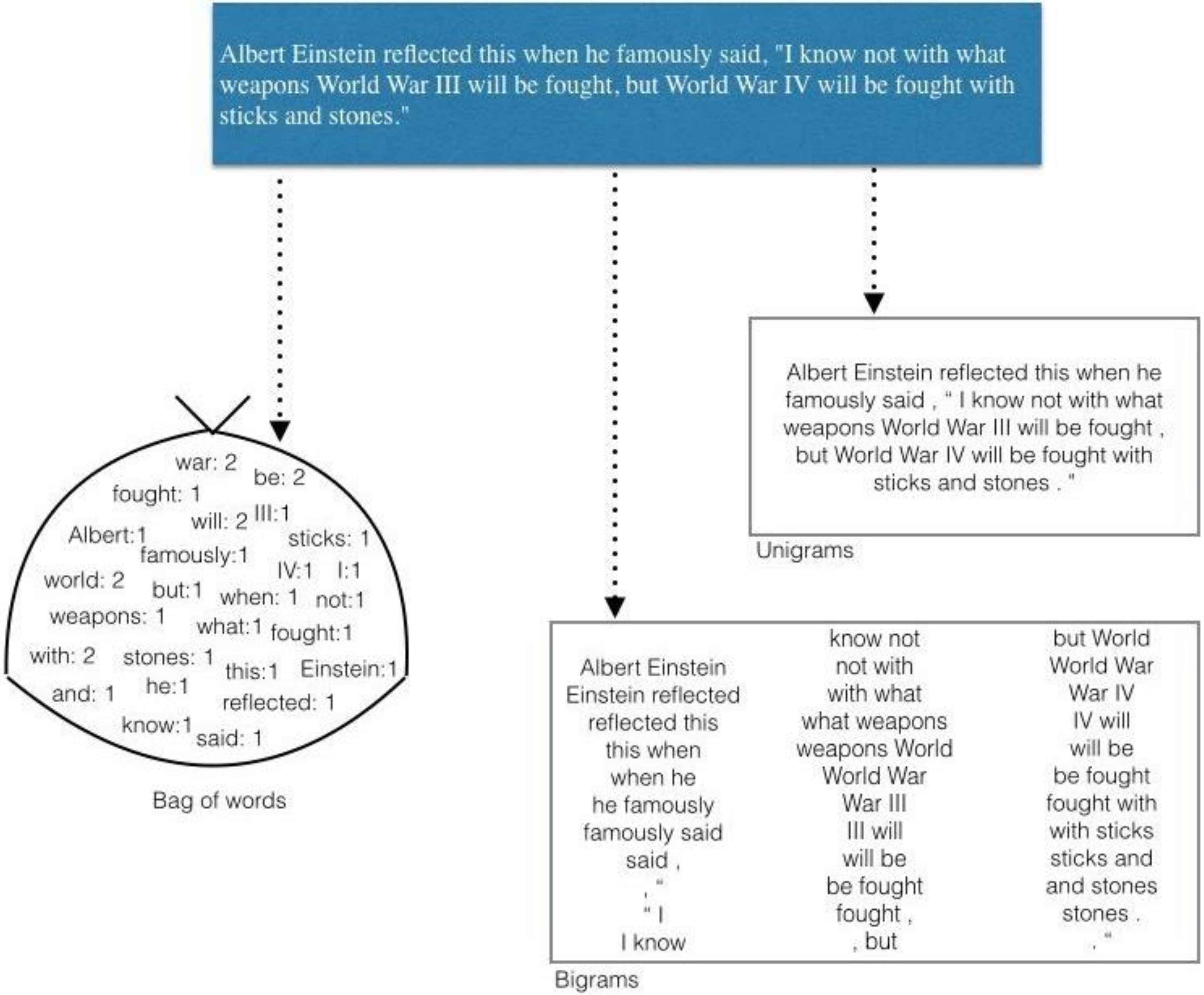
$$P(x_1, x_2, \ldots, x_n|y) = P(x_1|y)\, P(x_2|y)\, P(x_3|y) \ldots P(x_n|y)$$

# Bag-of-Words (BoW)

Важен только набор слов, не важен порядок.

$x_1, x_2, \ldots, x_n$- 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."
war: 2
be: 2
fought: 1
will: 2
III:1
Albert:1
sticks: 1
famously:1
IV:1
I:1
world: 2
but:1
when: 1
not:1
weapons: 1
what:1
fought:1
with: 2
stones: 1
this:1
Einstein:1
and: 1
he:1
reflected: 1
know:1
said: 1
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
Einstein reflected
reflected this
this when
when he
he famously
famously said
said ,
, "
" I
I know
know not
not with
with what
what weapons
weapons World
World War
War III
III will
will be
be fought
fought ,
, but
but World
World War
War IV
IV will
will be
be fought
fought with
with sticks
sticks and
and stones
stones .
. "
Bigrams

# Наивный байесовский классификатор

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

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

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

# Слова в документах

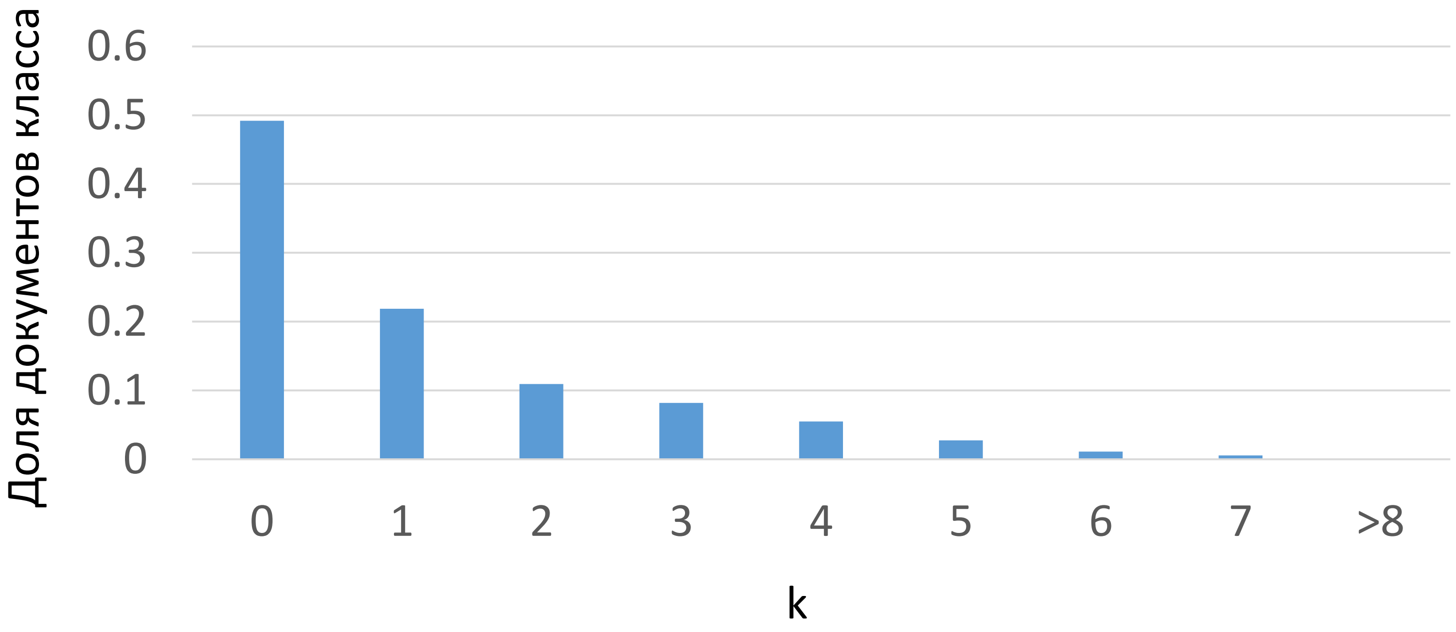
$P(y)$ – частота класса $y$.

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

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

Проблема – $count(x_i, y) = 0$

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



# Обработка текста

- Уменьшение словаря:
  - 35, 535, 17, 200000 → $number
  - (5+3), $\frac{1}{2}\mathrm{w}^{\mathrm{T}}\mathrm{w} + C$ → $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:

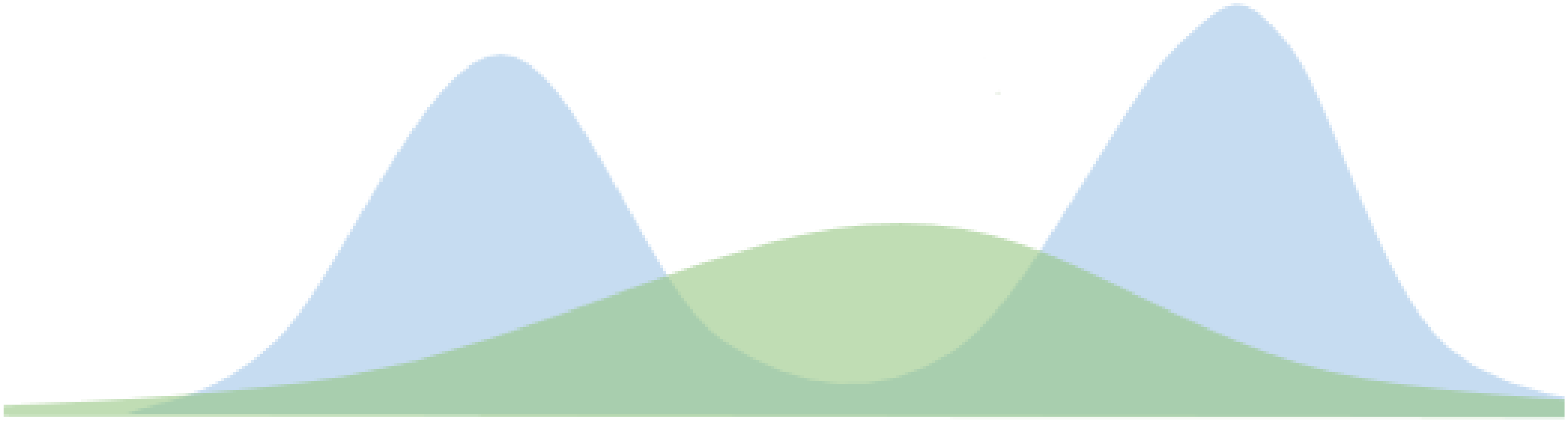
$$P(x_i|y) = P(x_i = 1|y)x_i + \big(1 - P(x_i = 1|y)\big)(1 - x_i), \qquad x_i \in \{0,1\}$$

Распределение Гаусса:

$$p(x_i|y) = \frac{1}{\sqrt{2\pi\left(\sigma_i^y\right)^2}} e^{-\frac{\left(x_i - \mu_i^y\right)^2}{2\left(\sigma_i^y\right)^2}}$$

# 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 ковариций
$\alpha_k$ – "вес" гауссианы, вероятность того, что случайная точка принадлежит к Гауссиане $k$
$\sum \alpha_k = 1$

Принадлежность объекта $x_i$ к k-му распределению :

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

**E-Step:** считаем $w_{ik}$

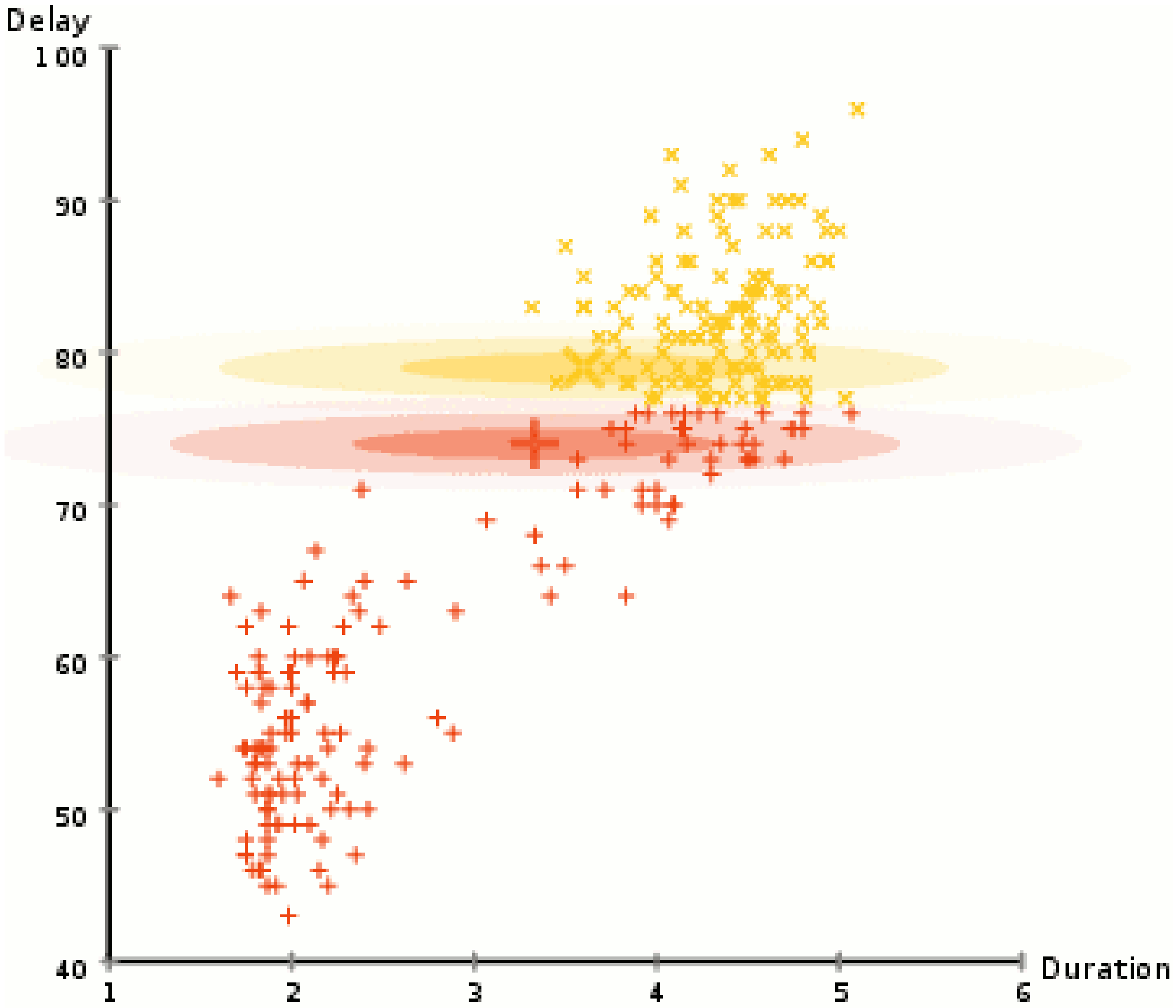
**M-Step:**

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

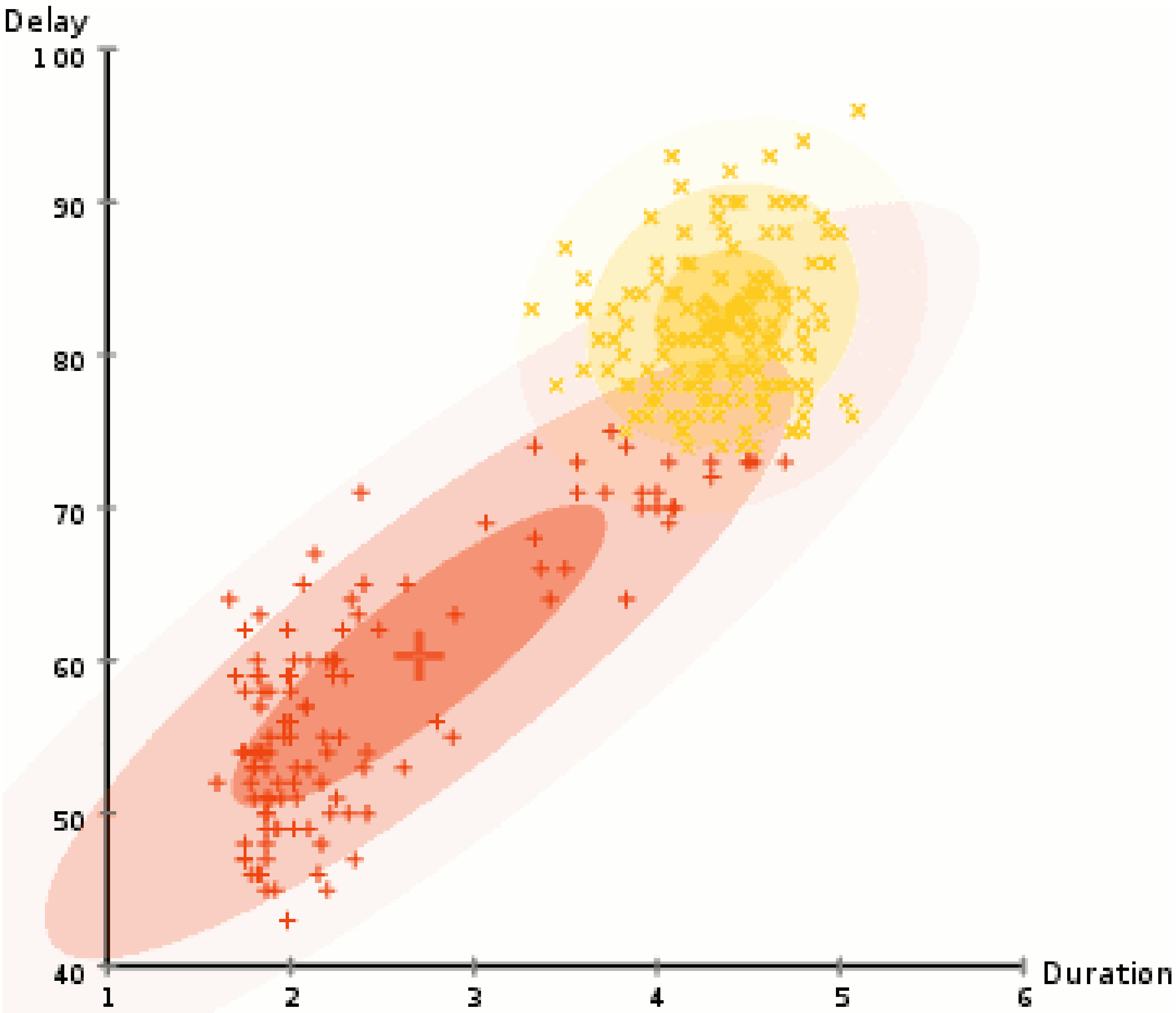
$$\mu_k^{new} = \left(\frac{1}{N_k}\right) \sum_{i=1}^{N} w_{ik} \cdot x_i$$

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

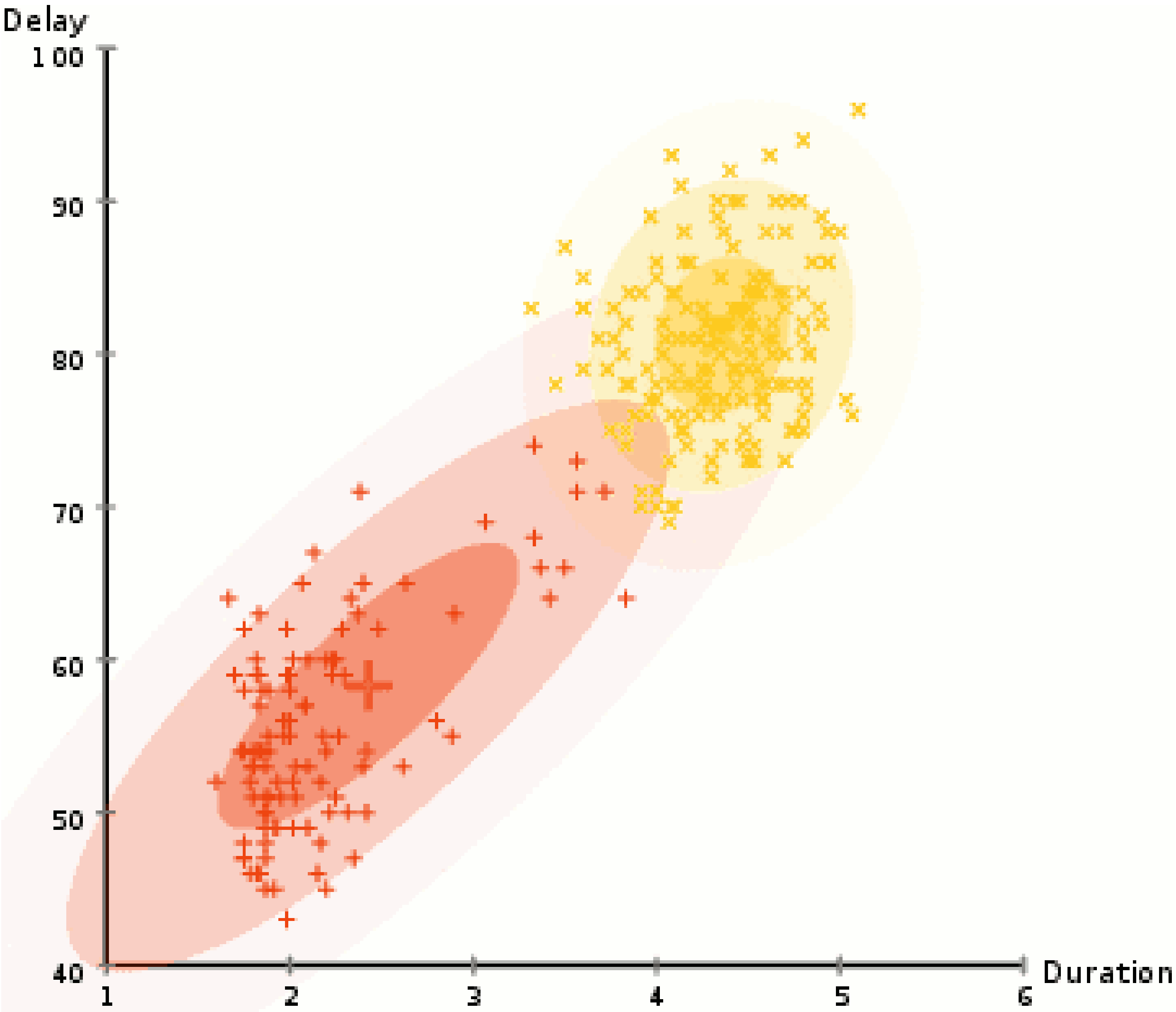
# EM Clustering



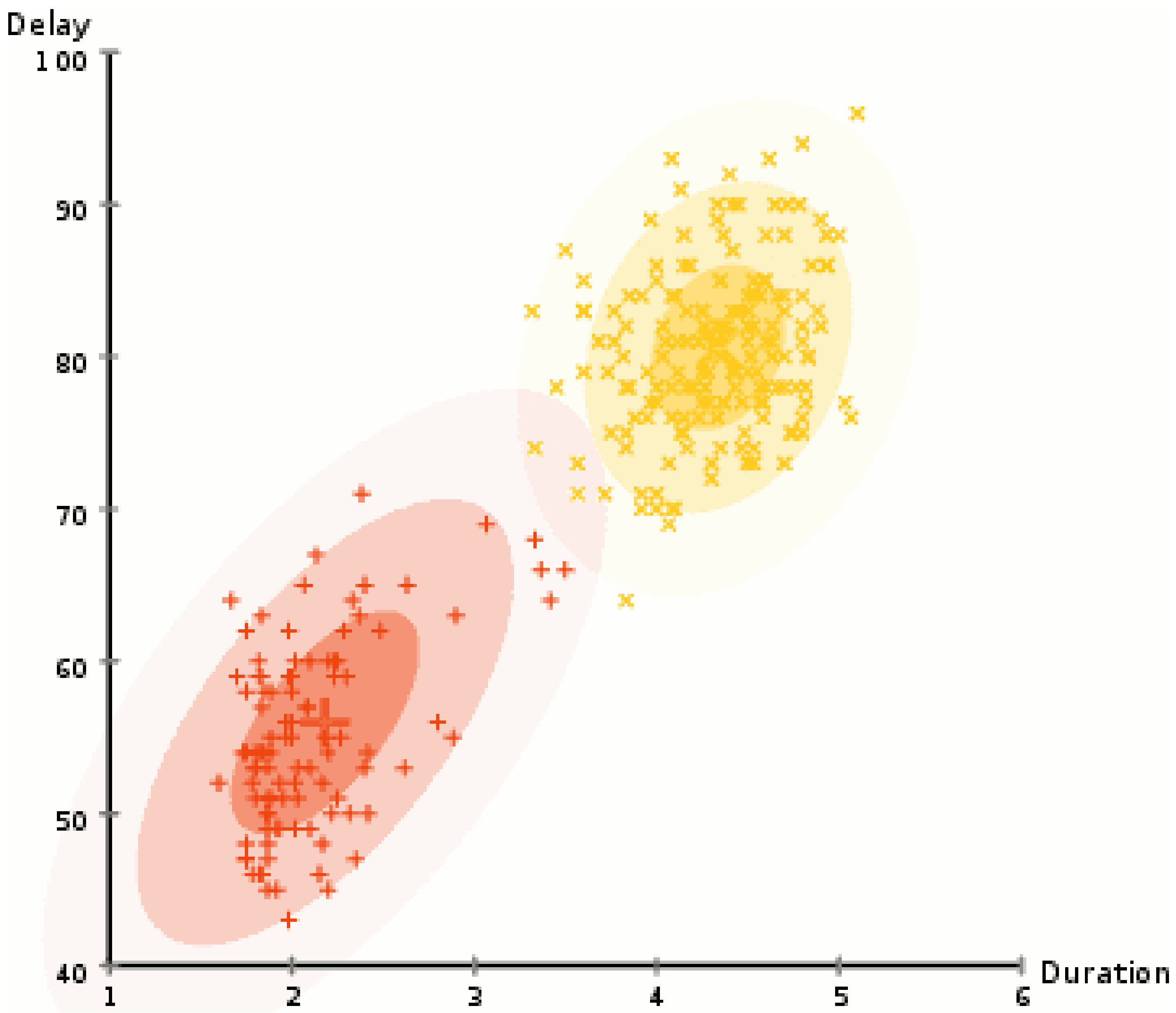
# EM Clustering



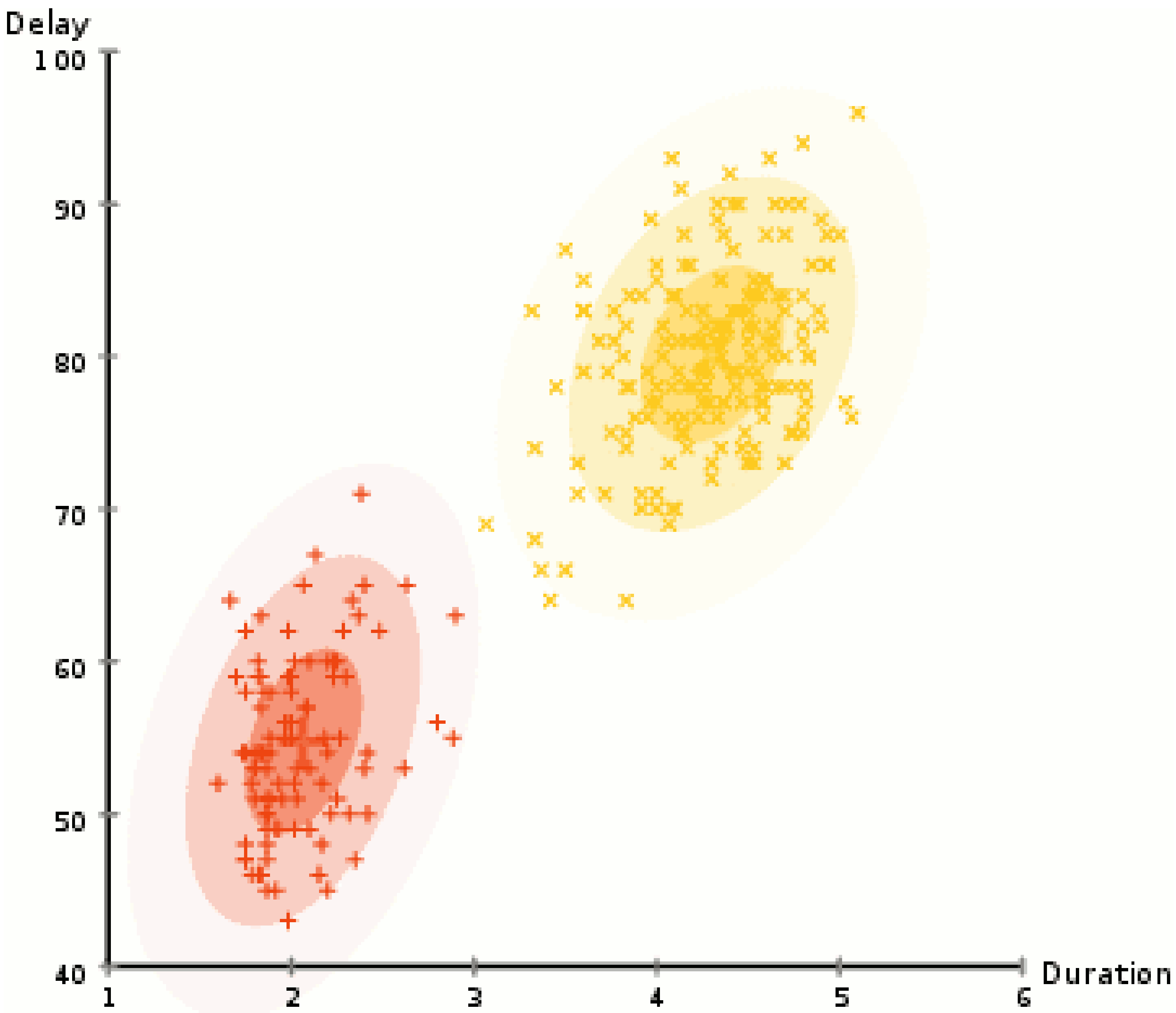
# EM Clustering



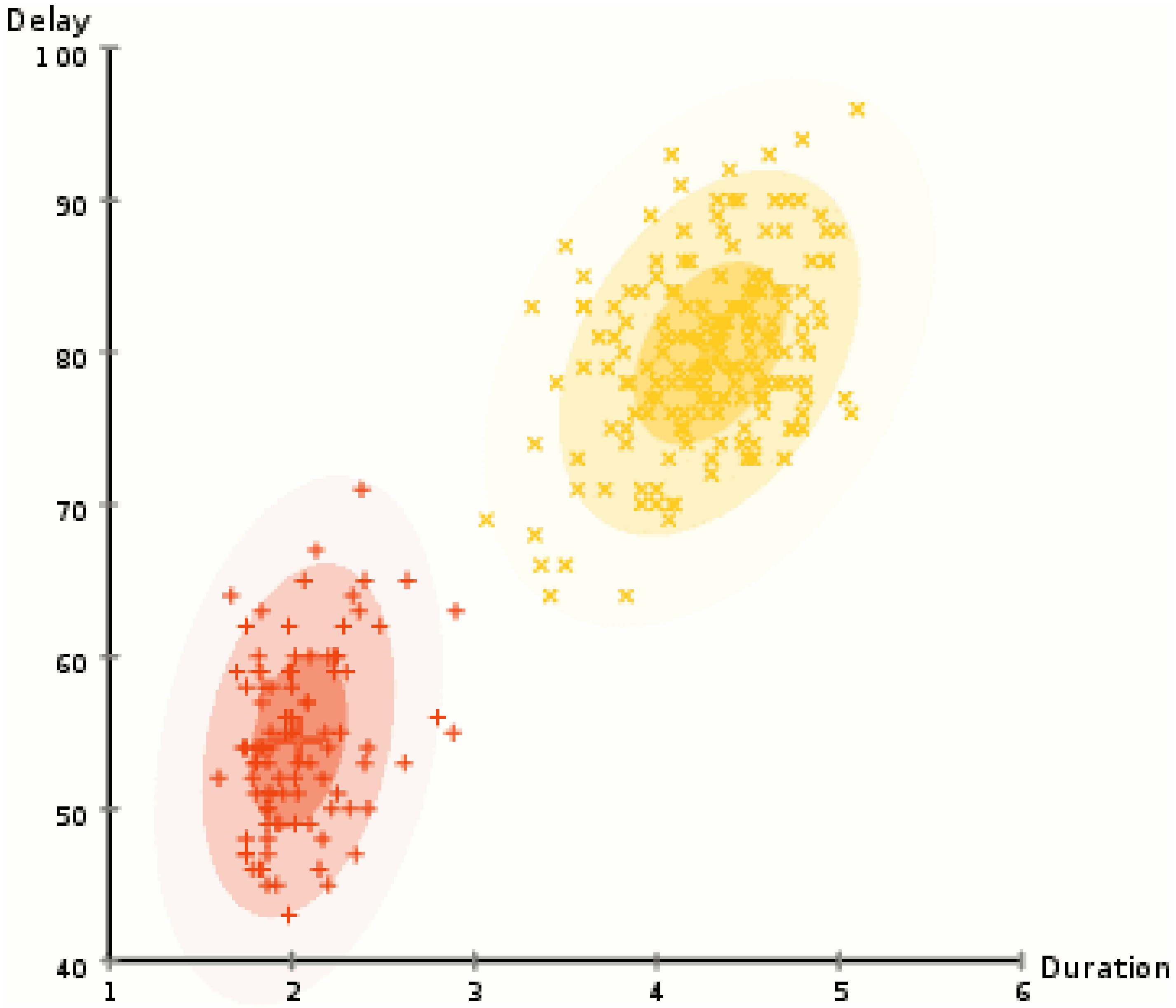
# EM Clustering



# EM Clustering



# EM Clustering



# Naïve Bayes – отличный Baseline!