Bayes classifier

Bayes classifier

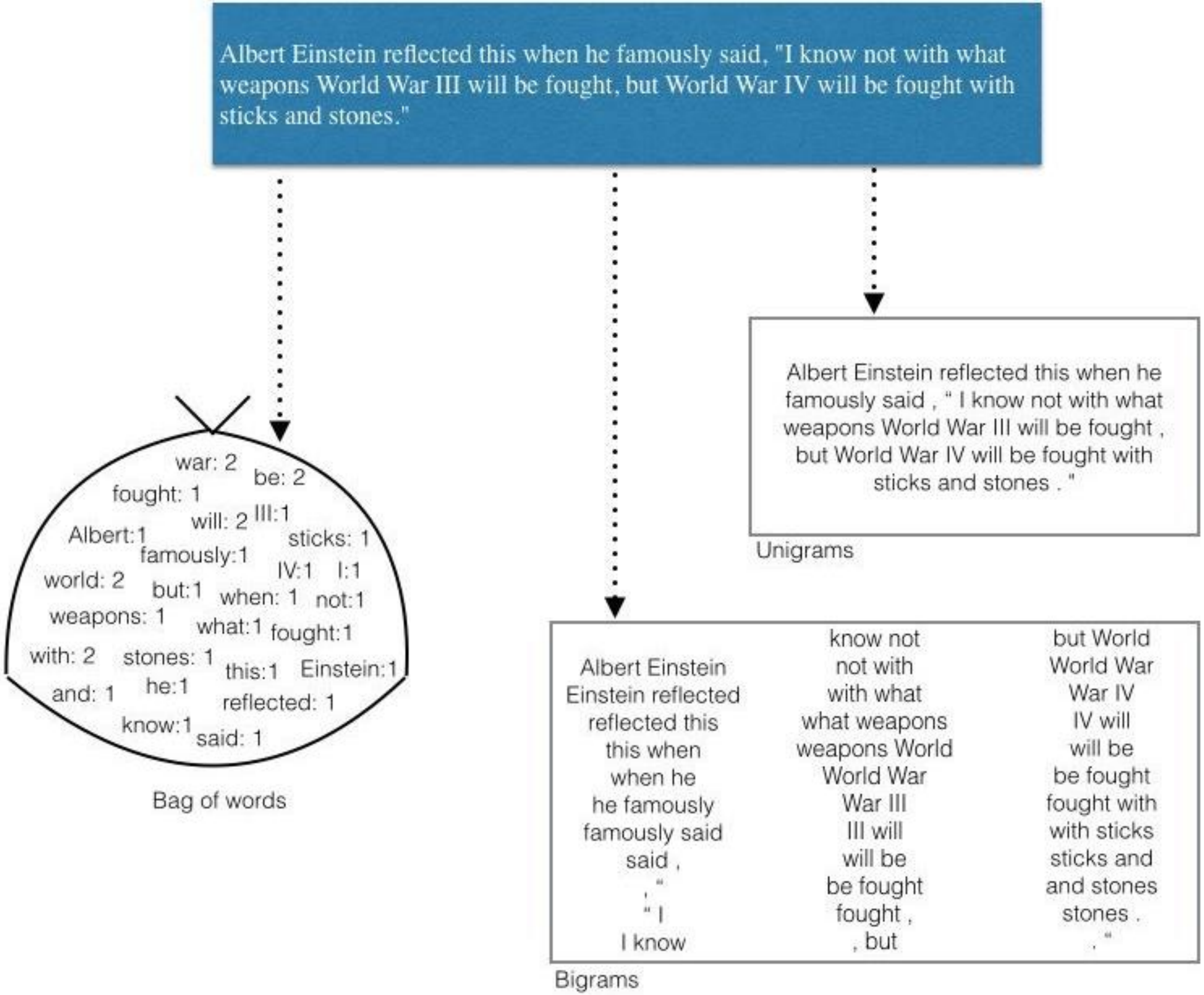
$$P(y|\mathrm{x}) = \frac{P(y)P(\mathrm{x}|y)}{P(\mathrm{x})}$$

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

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

Naïve assumption (features are independent):

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


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

Naïve Bayes classifier

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

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

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

Words in documents

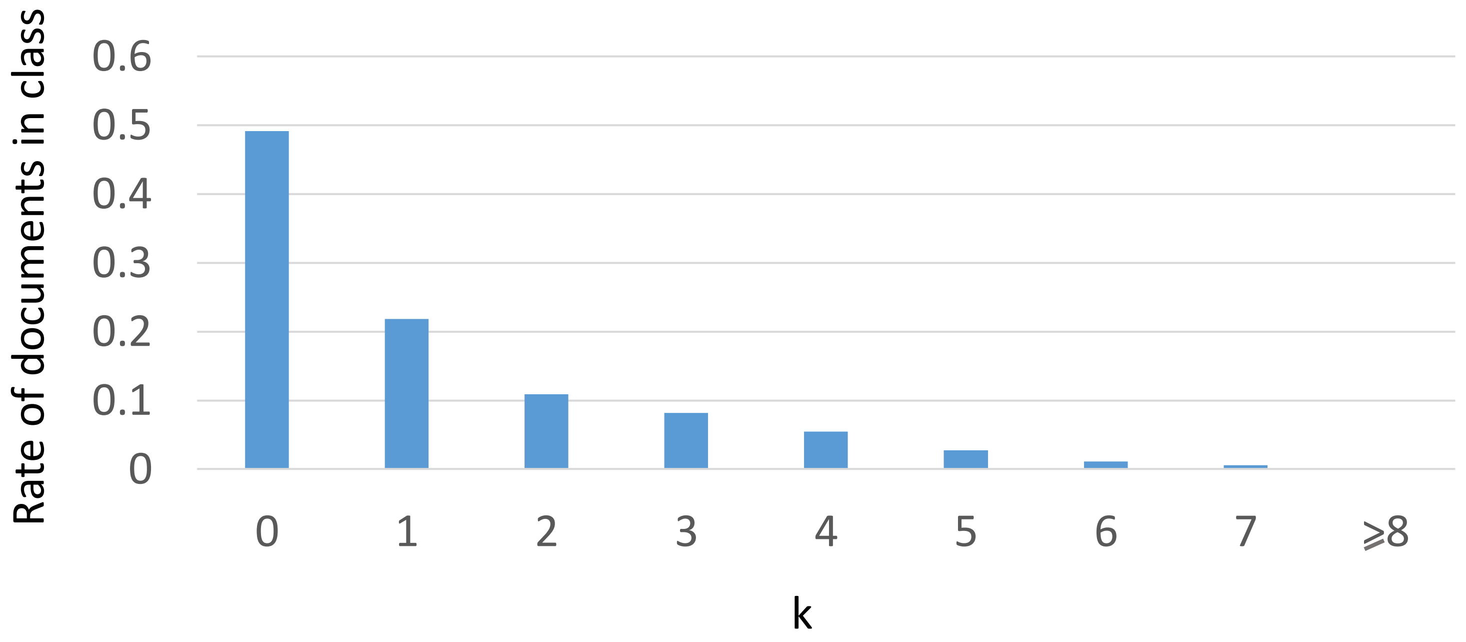
$P(y)$ is the frequency of class y.

$P(x_i = k|y)$ is the probability of value k of feature x_i in class y. Such as a proportion of documents in class with $x_i = k$.

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

Problem – $count(x_i, y_j) = 0$

$$\hat{P}(x_i|y_j) = \frac{count(x_i, y_j) + 1}{count(y_j) + K}$$



Bayes classifier with various types of features

Binary features:

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

Distribution for continuous features:

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

Distribution evaluation

A simple option is to use sample mean and variance for μ and σ.

A more complex option is expectation-maximization algorithm (EM) with Gaussian mixtures.

Expectation-maximization (EM)

Define Gaussian mixture of K distributions by:

μ_k is the mean vector, Σ_k is the covariance matrix
α_k is the weight of distribution k, a probability that random point belongs to it.
$\sum \alpha_k = 1$

Affinity of object $\mathrm{x_i}$ to distribution k:

$$w_{ik} = P(\mu_k, \Sigma_k | \mathrm{x_i}) = \frac{p(\mathrm{x}_i | \mu_k, \Sigma_\mathrm{k}) \cdot \alpha_k}{\sum_j p_j \left(\mathrm{x}_i \middle| \mu_j, \Sigma_j\right) \cdot \alpha_j}$$

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

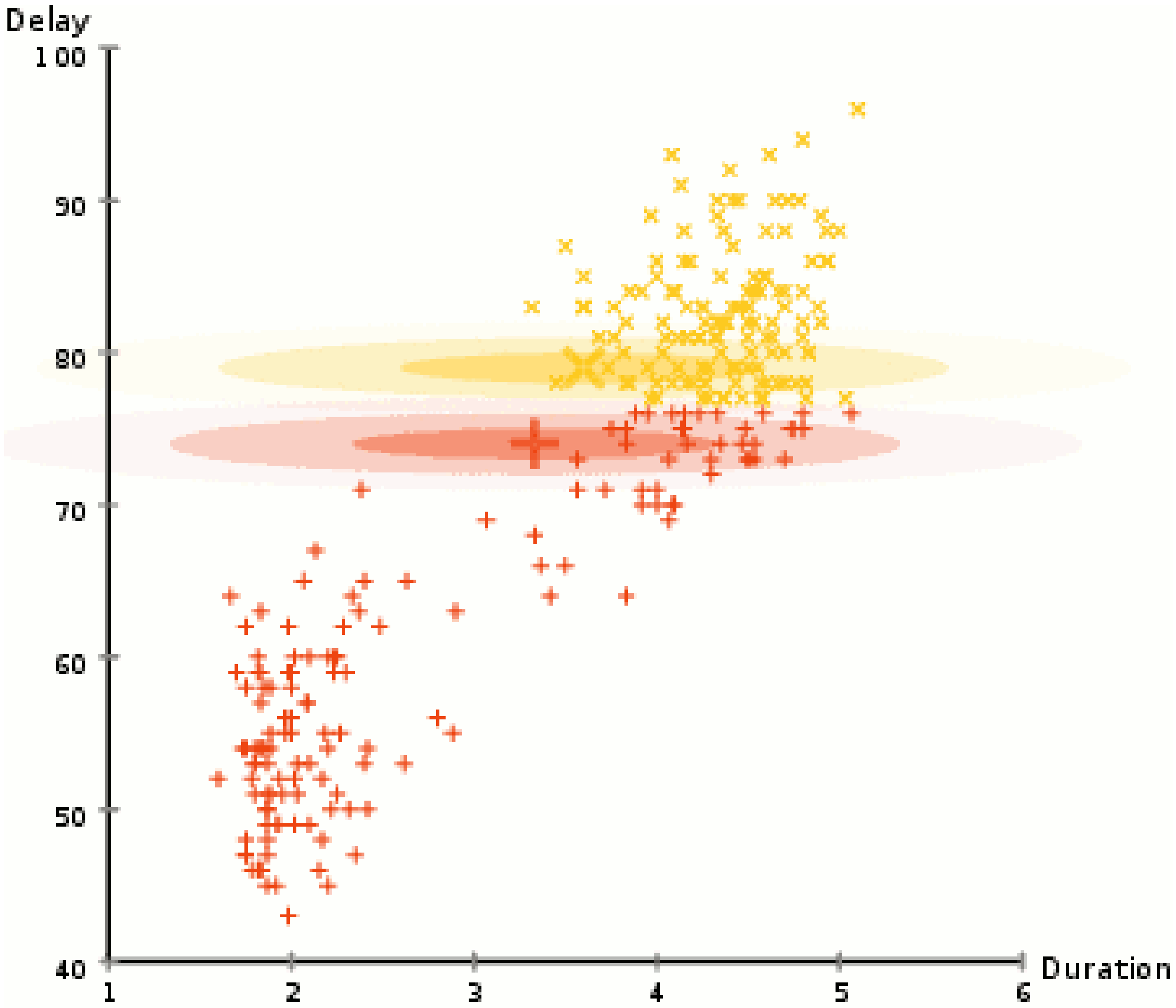
E-Step: calculate w_{ik}

M-Step:

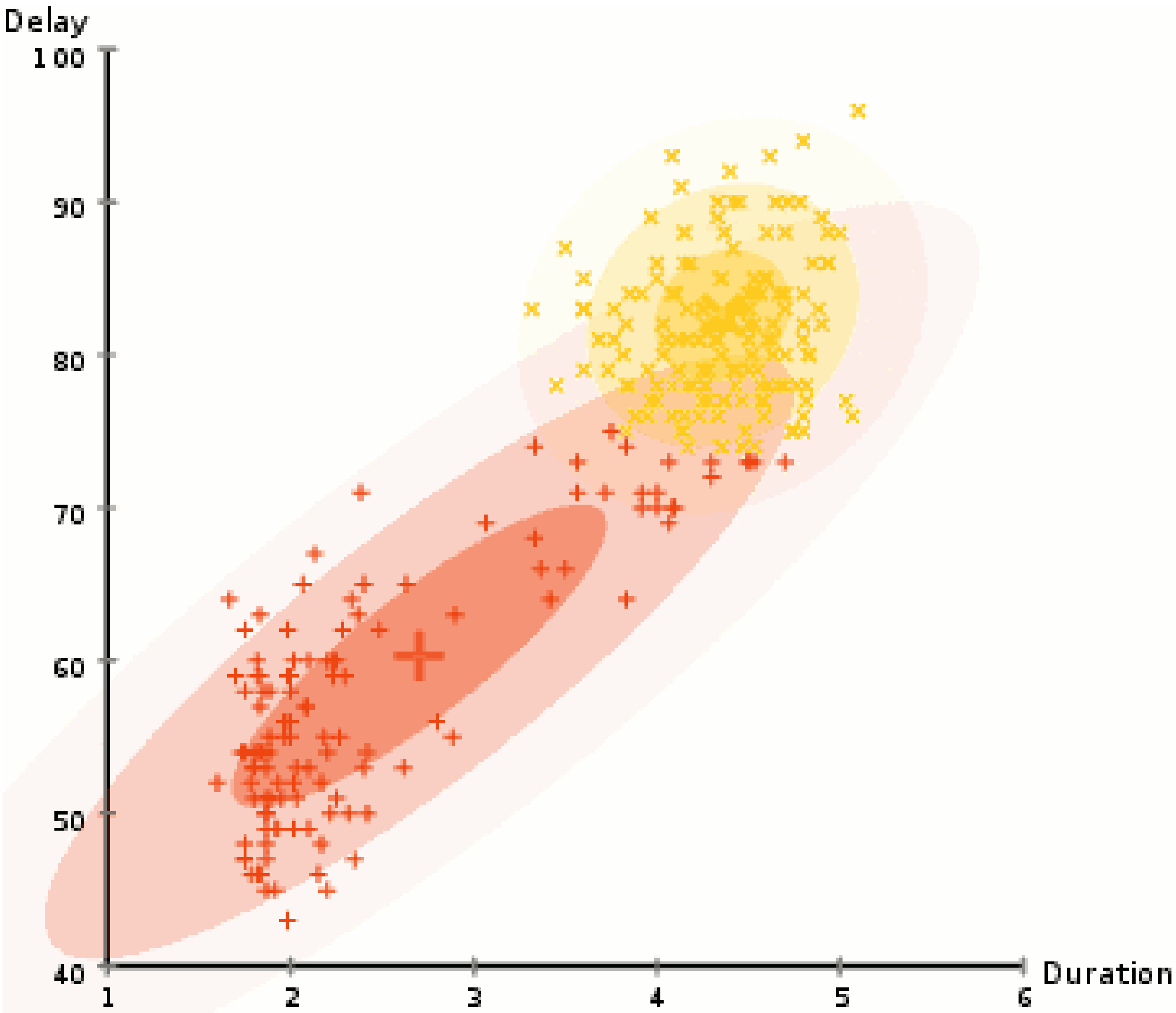
$$\mu_k^{new} = \left(\frac{1}{N_k}\right) \sum_{i=1}^{N} w_{ik} \cdot \mathrm{x}_i$$

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

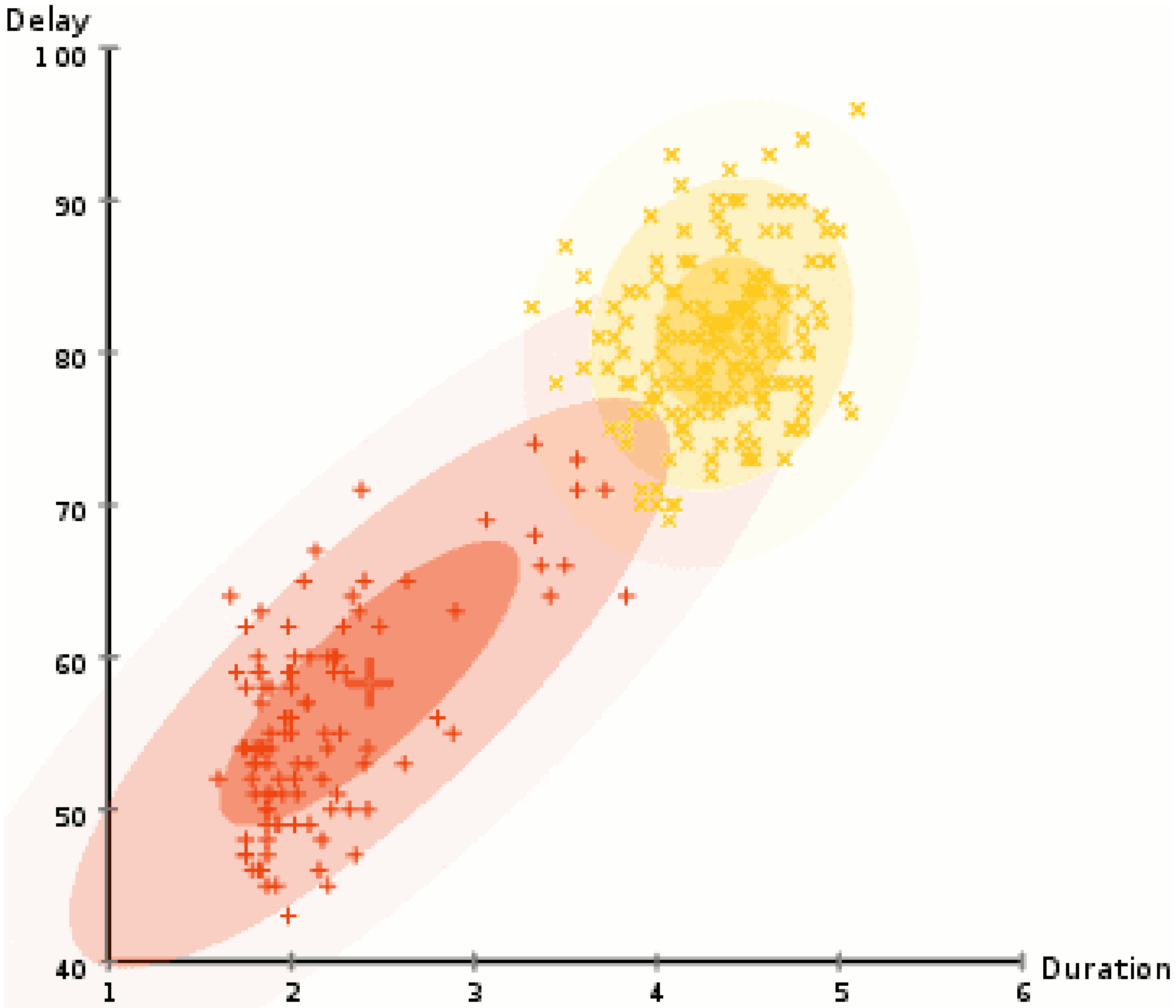
EM Clustering



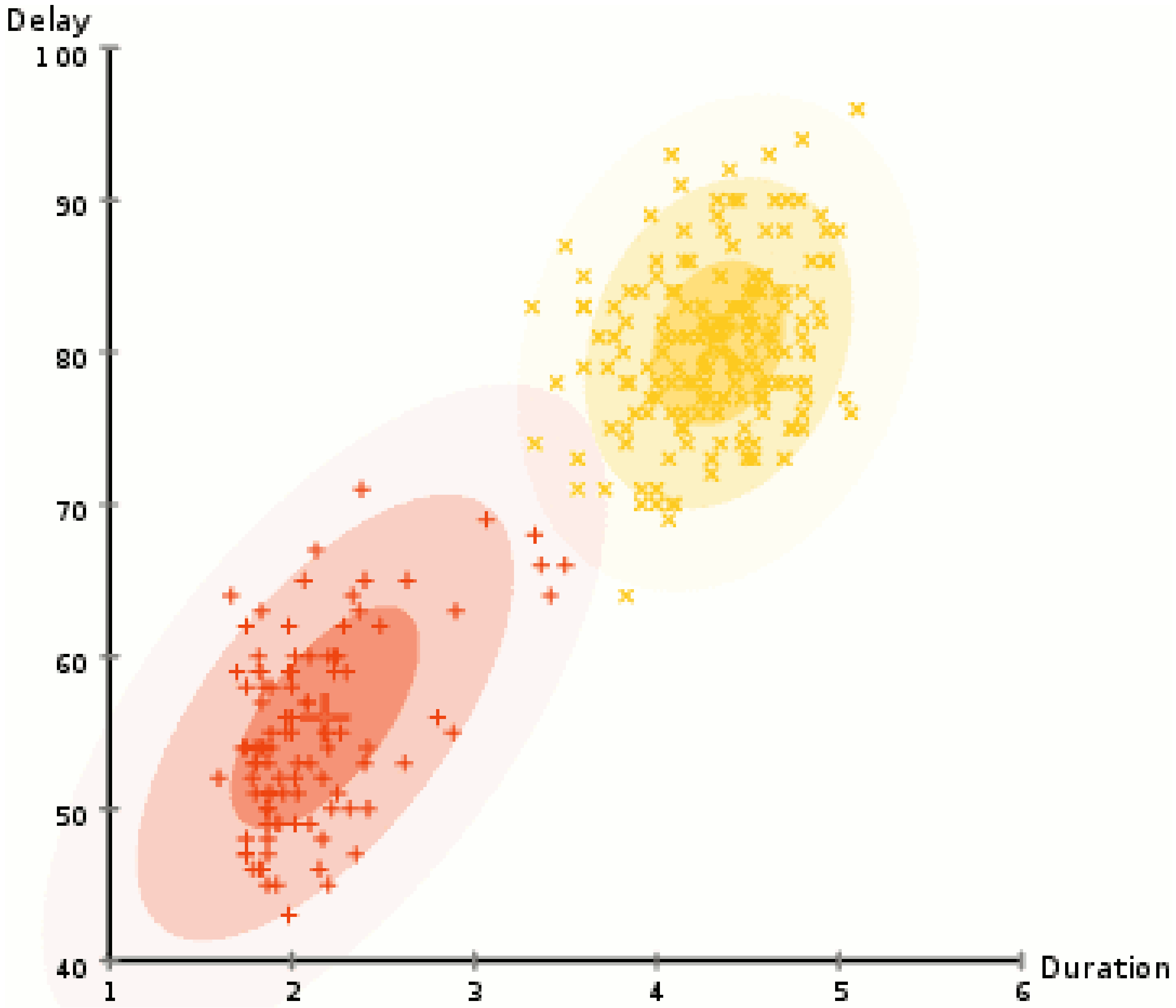
EM Clustering



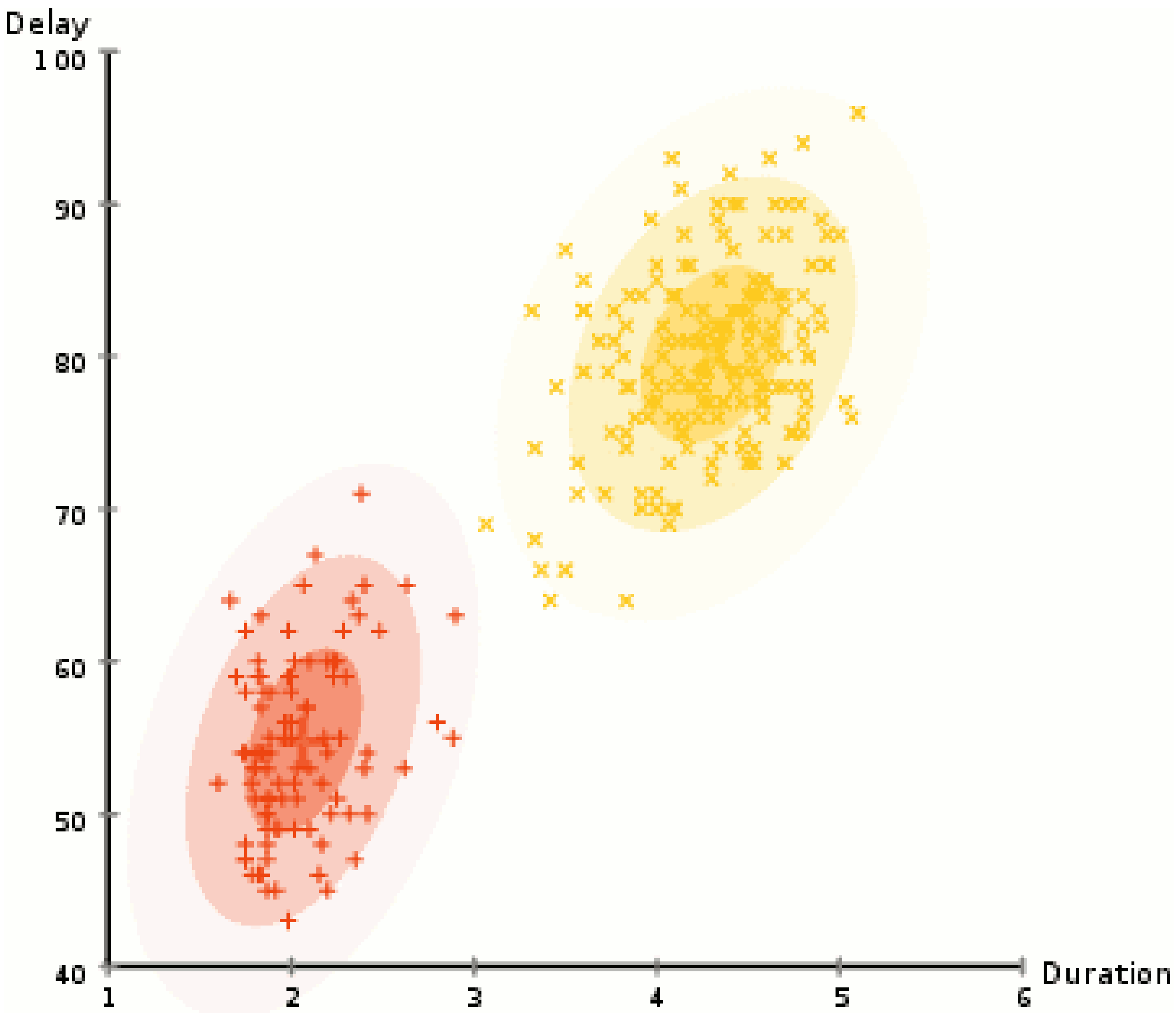
EM Clustering



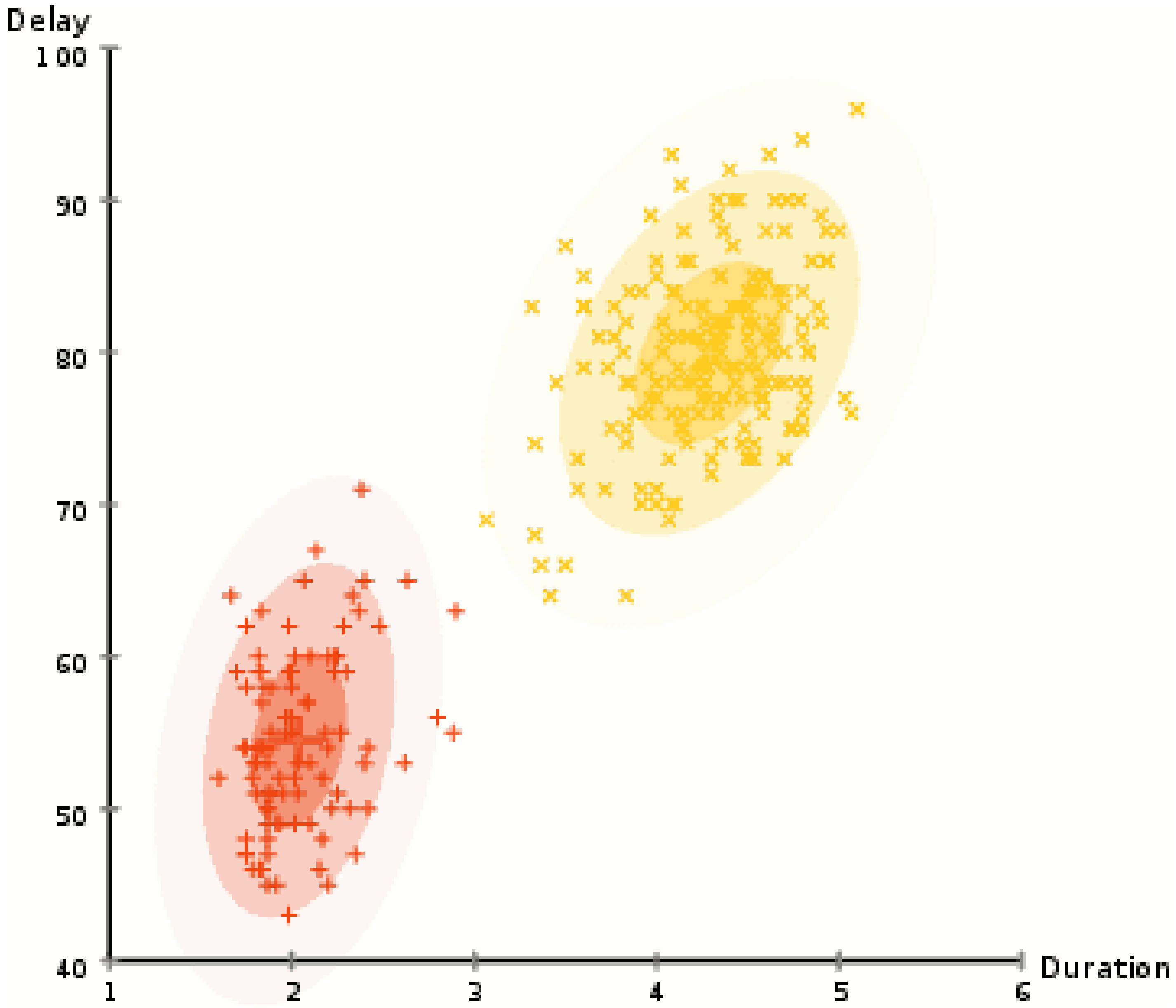
EM Clustering



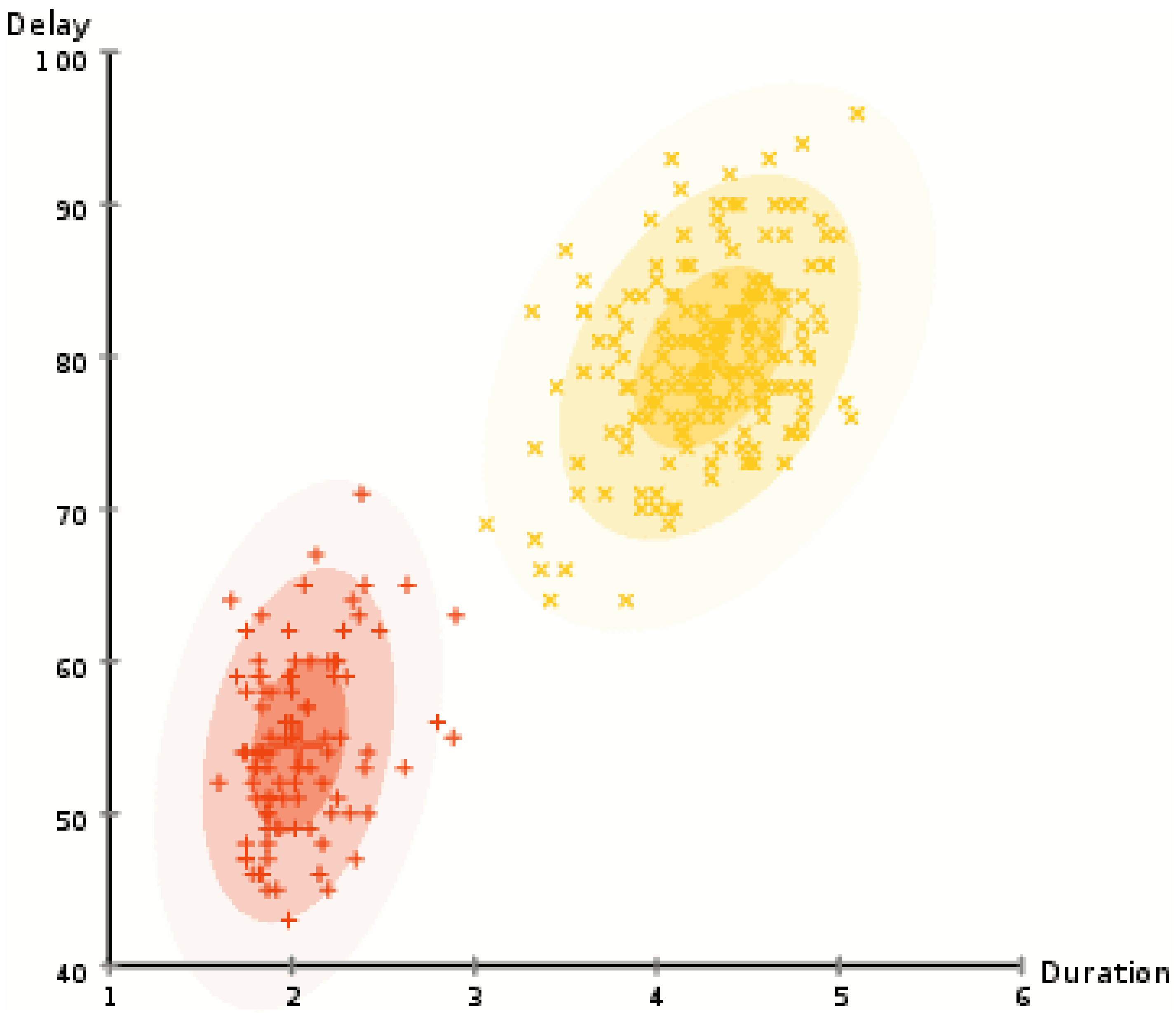
EM Clustering

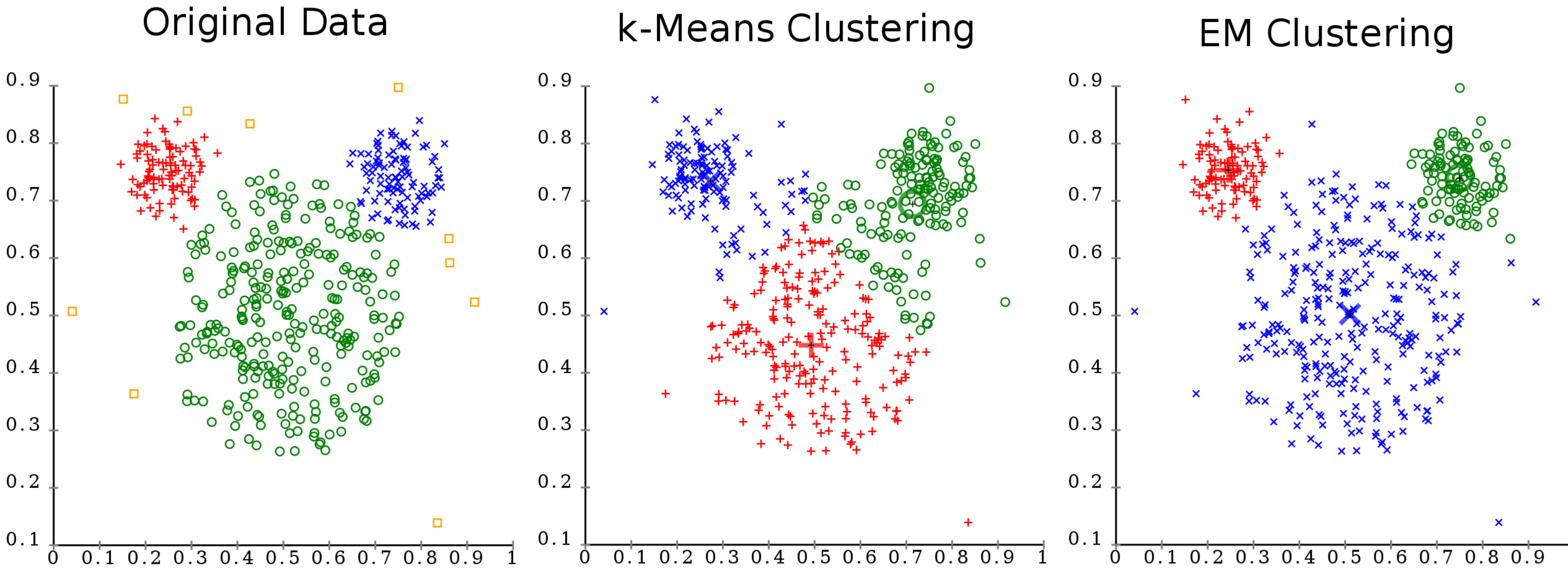


EM Clustering



EM Clustering




Original Data
k-Means Clustering
EM Clustering
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

Naïve Bayes is a very nice baseline!