3F3: Signal and Pattern Processing

Lecture 4: Clustering

Zoubin Ghahramani
zoubin@eng.cam.ac.uk

Department of Engineering
University of Cambridge

Lent Term

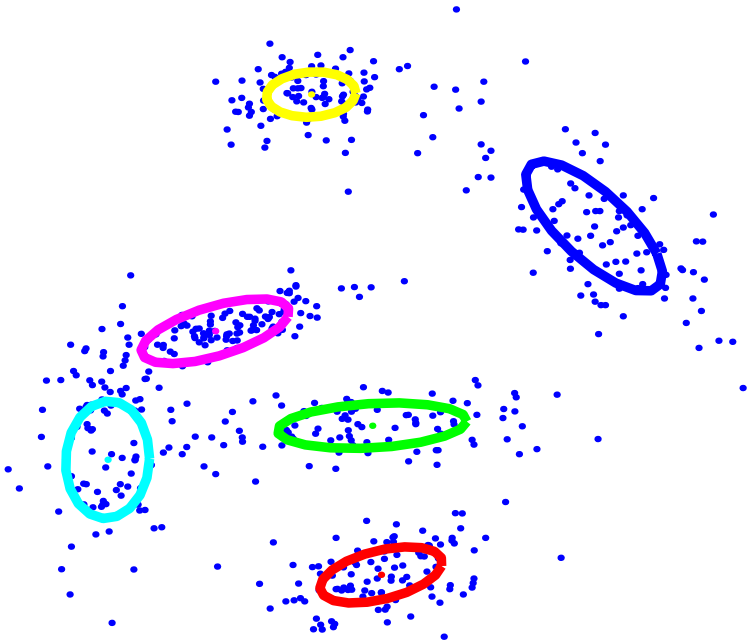
Clustering

Given some data, the goal is to discover "clusters" of points.

Roughly speaking, two points belonging to the same cluster are generally more similar to each other or closer to each other than two points belonging to different clusters.

Examples:

- cluster news stories into topics
- cluster genes by similar function
- cluster movies into categories
- cluster astronomical objects



The K-Means Algorithm

Input: Data Set $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ where $\mathbf{x}_n \in \Re^D$

Initialize Centers: $\mathbf{m}_k \in \Re^D$ for $k = 1 \ldots K$.

repeat:

  for $n = 1 \ldots N$:

    let $s_n = \arg\min_k \|\mathbf{x}_n - \mathbf{m}_k\|$ % assign data points to nearest center

  end for

  for $k = 1 \ldots K$:

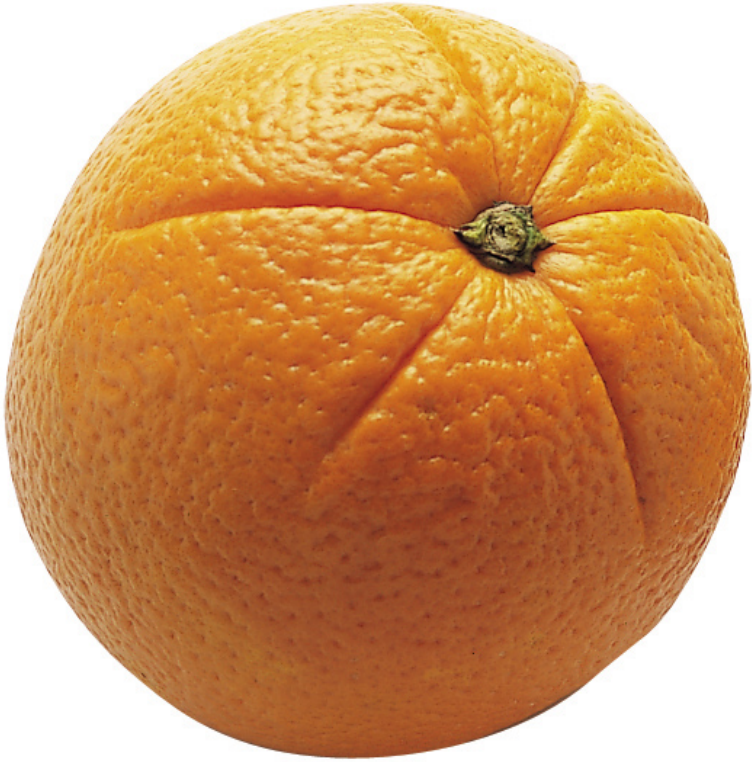
    let $\mathbf{m}_k = \text{mean}\{\mathbf{x}_n : s_n = k\}$ % re-compute means

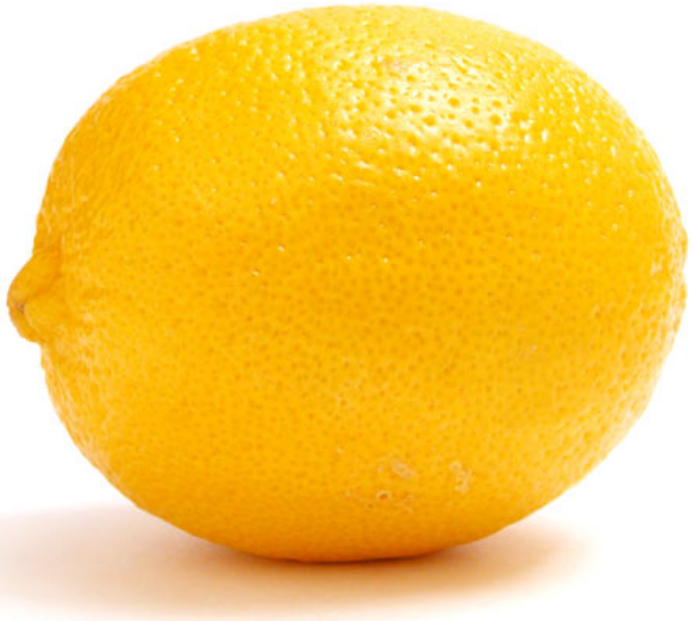
  end for

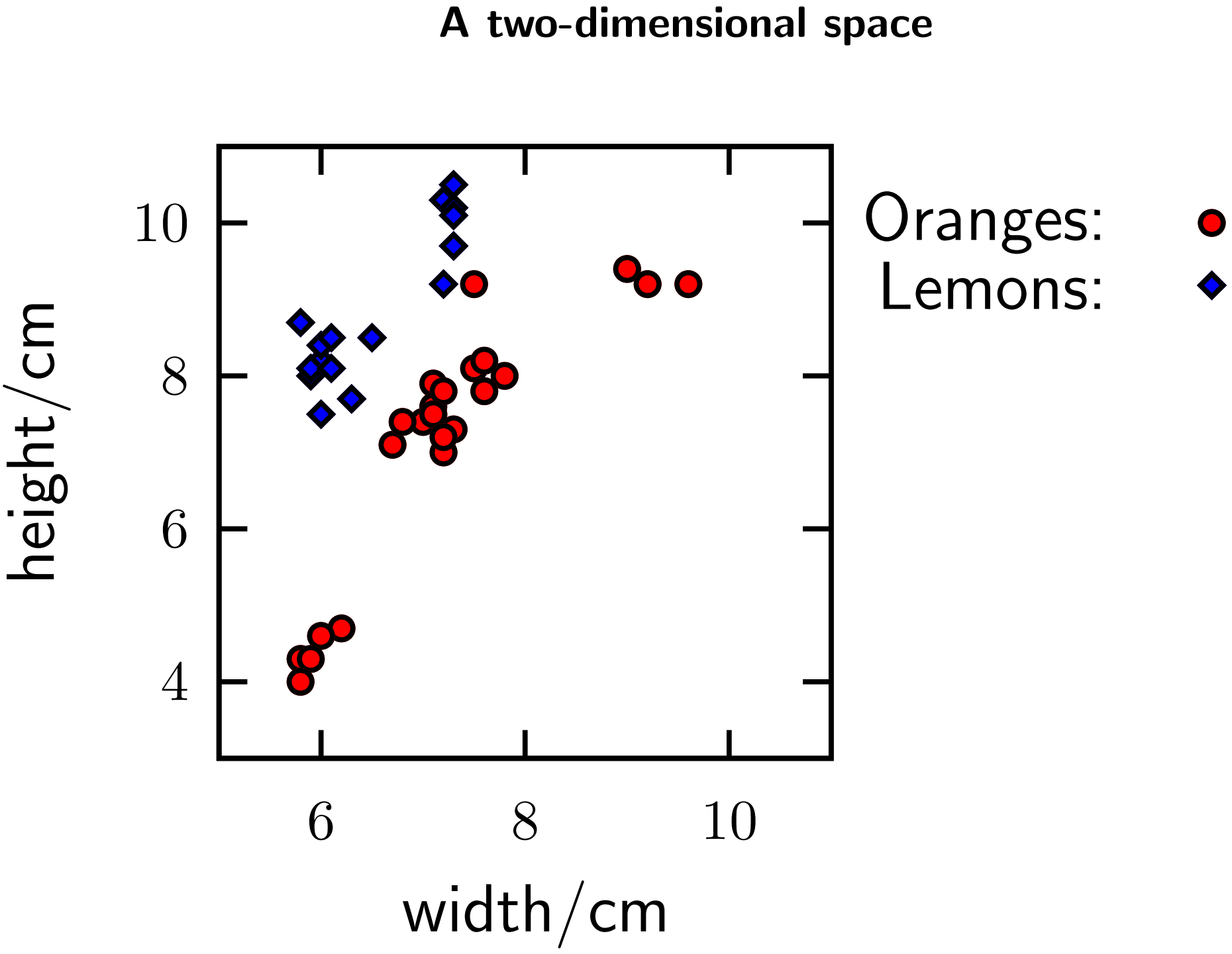
until convergence (s has not changed)

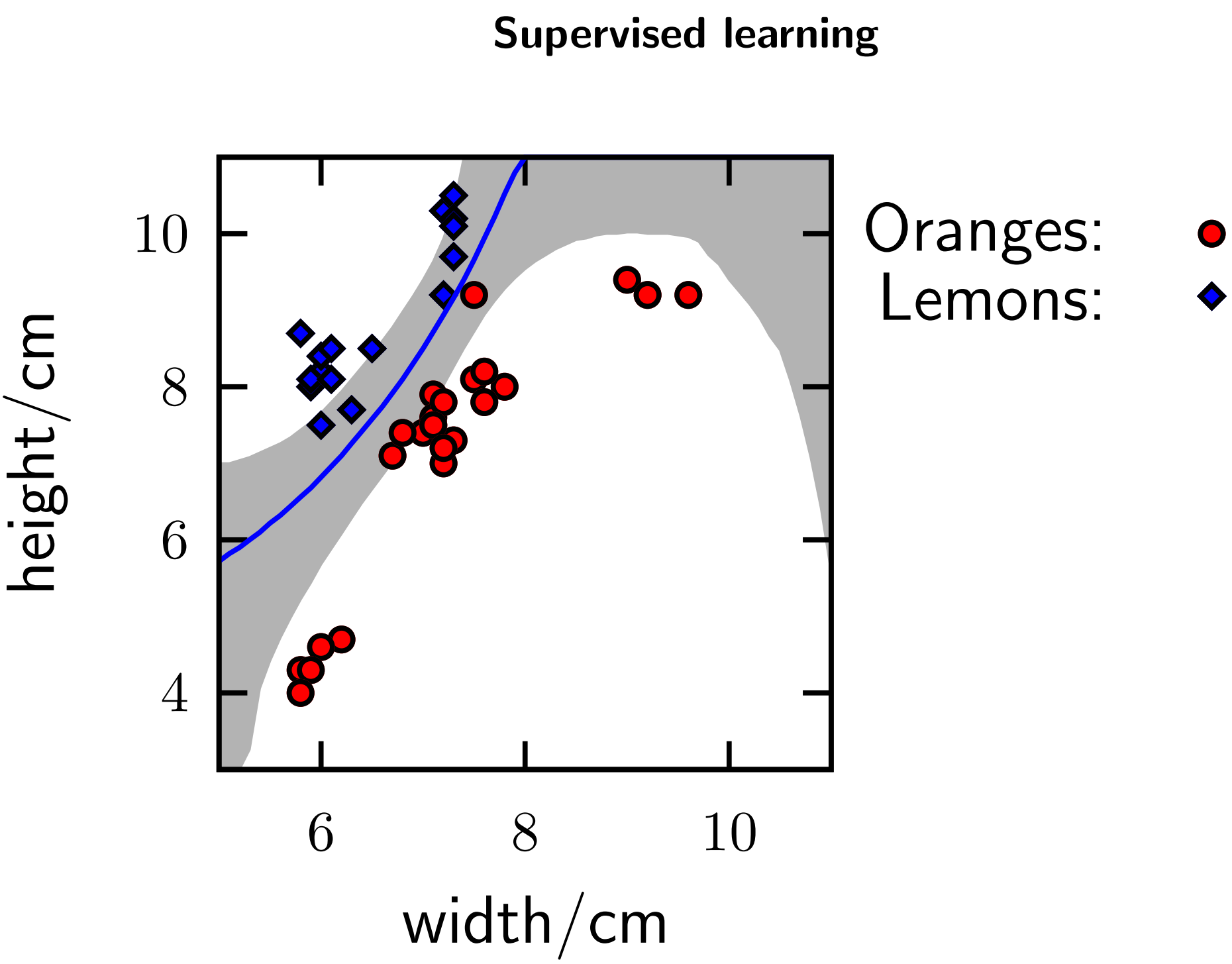
kmeansdemo

Oranges and Lemons
Thanks to Iain Murray

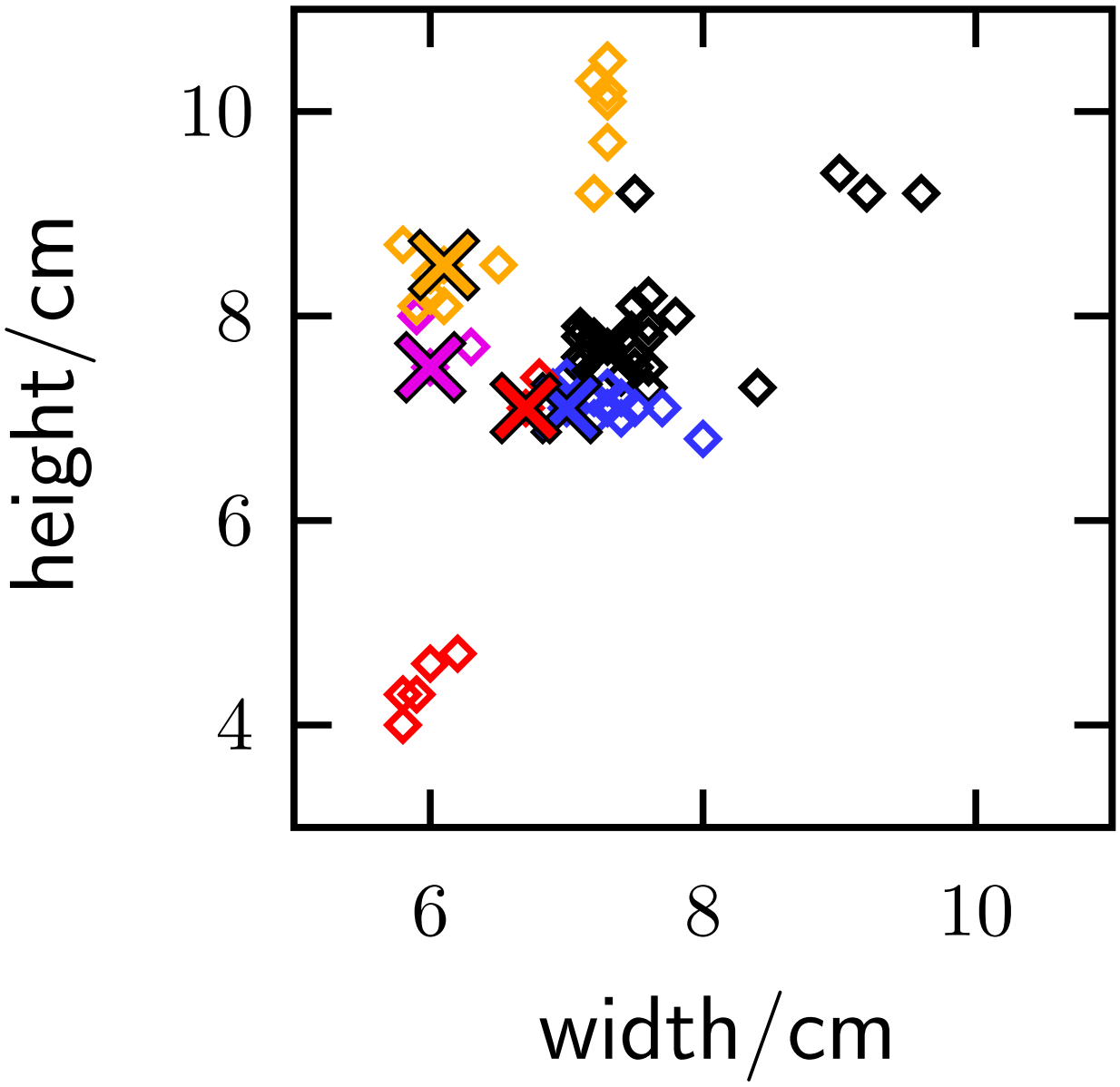




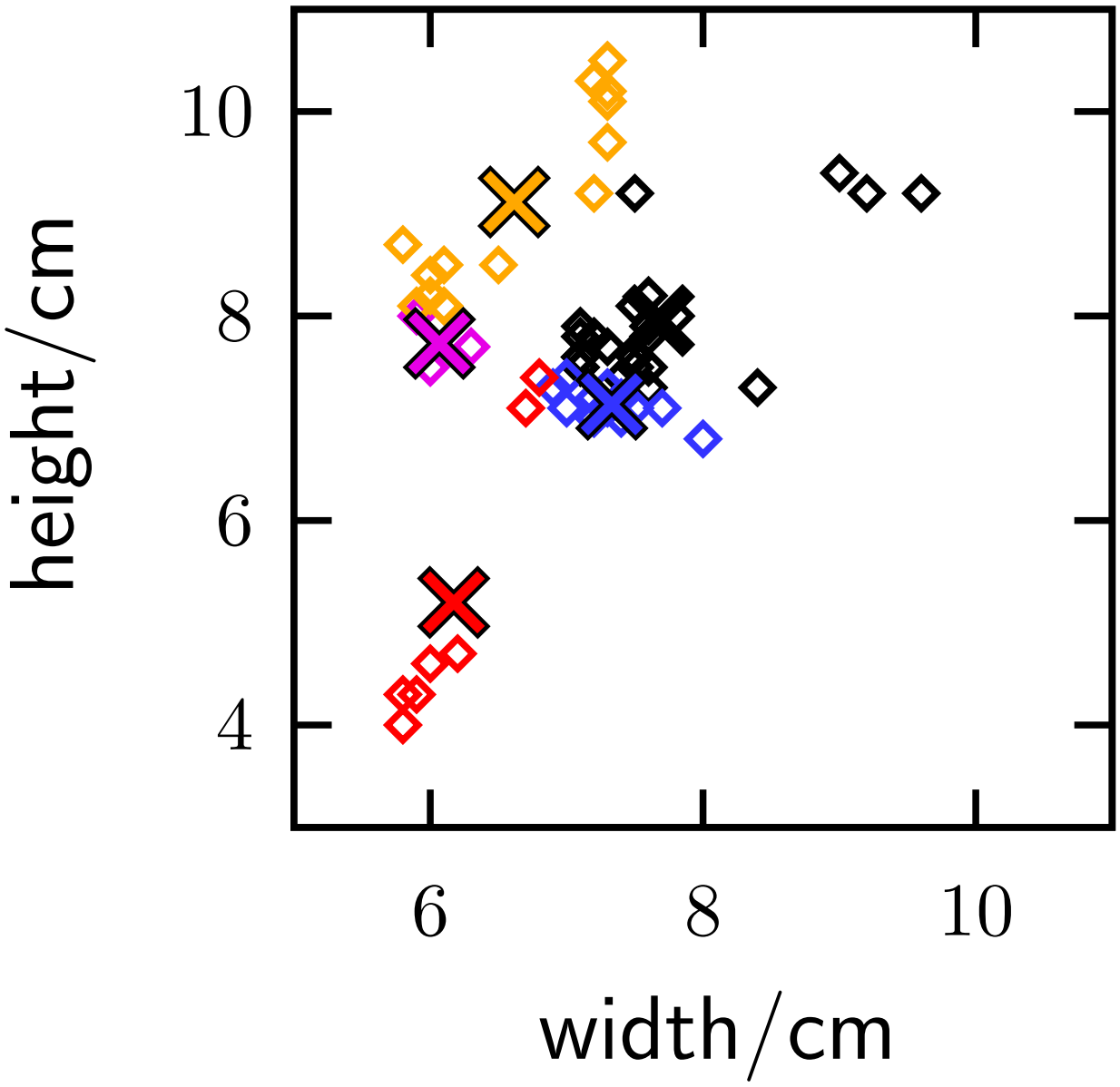

A two-dimensional space
height/cm
10
8
6
4
6
8
10
width/cm
Oranges:
Lemons:


Supervised learning
10
8
6
4
height/cm
6
8
10
width/cm
Oranges:
Lemons:

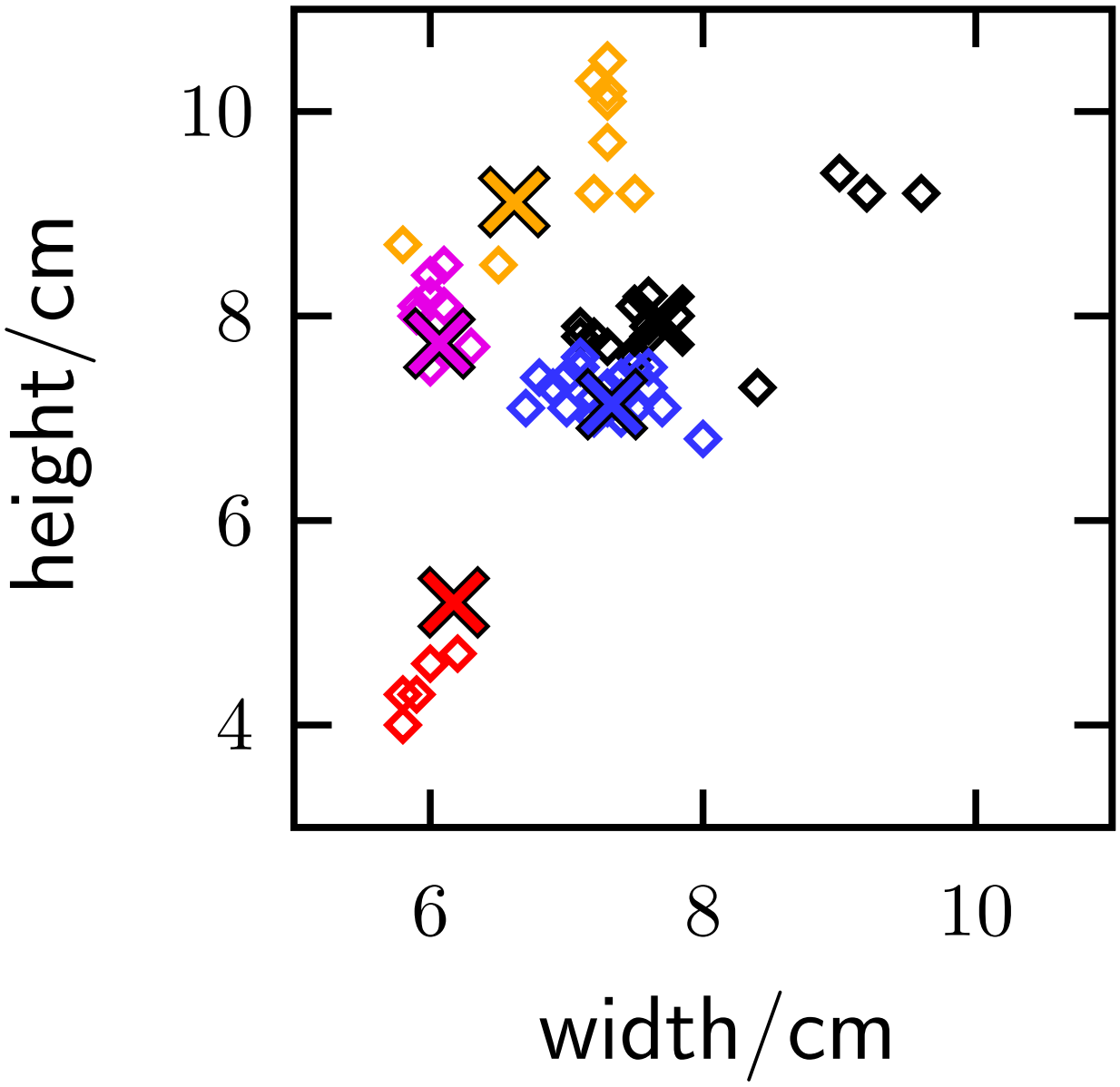
Clustering



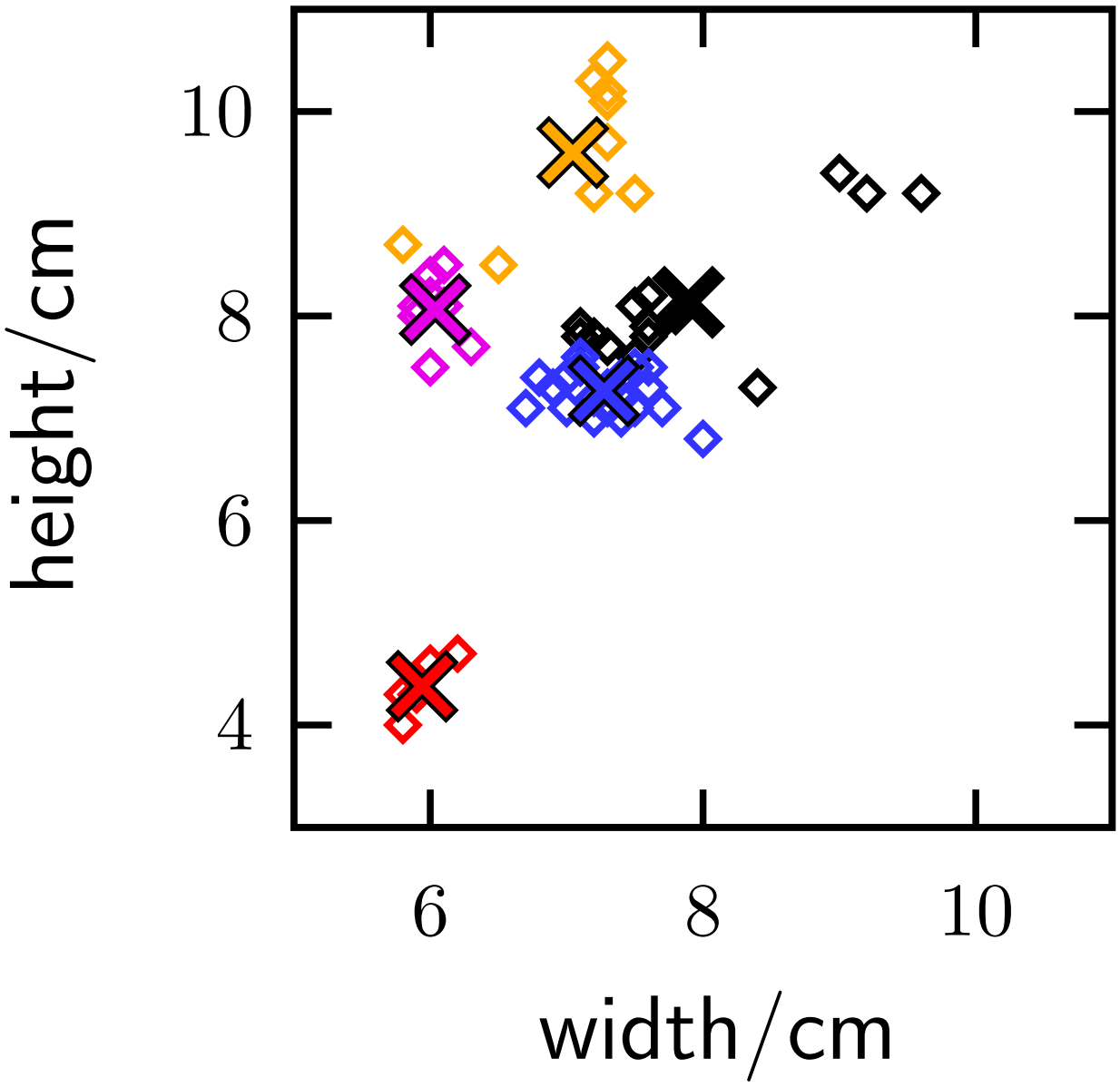
Clustering



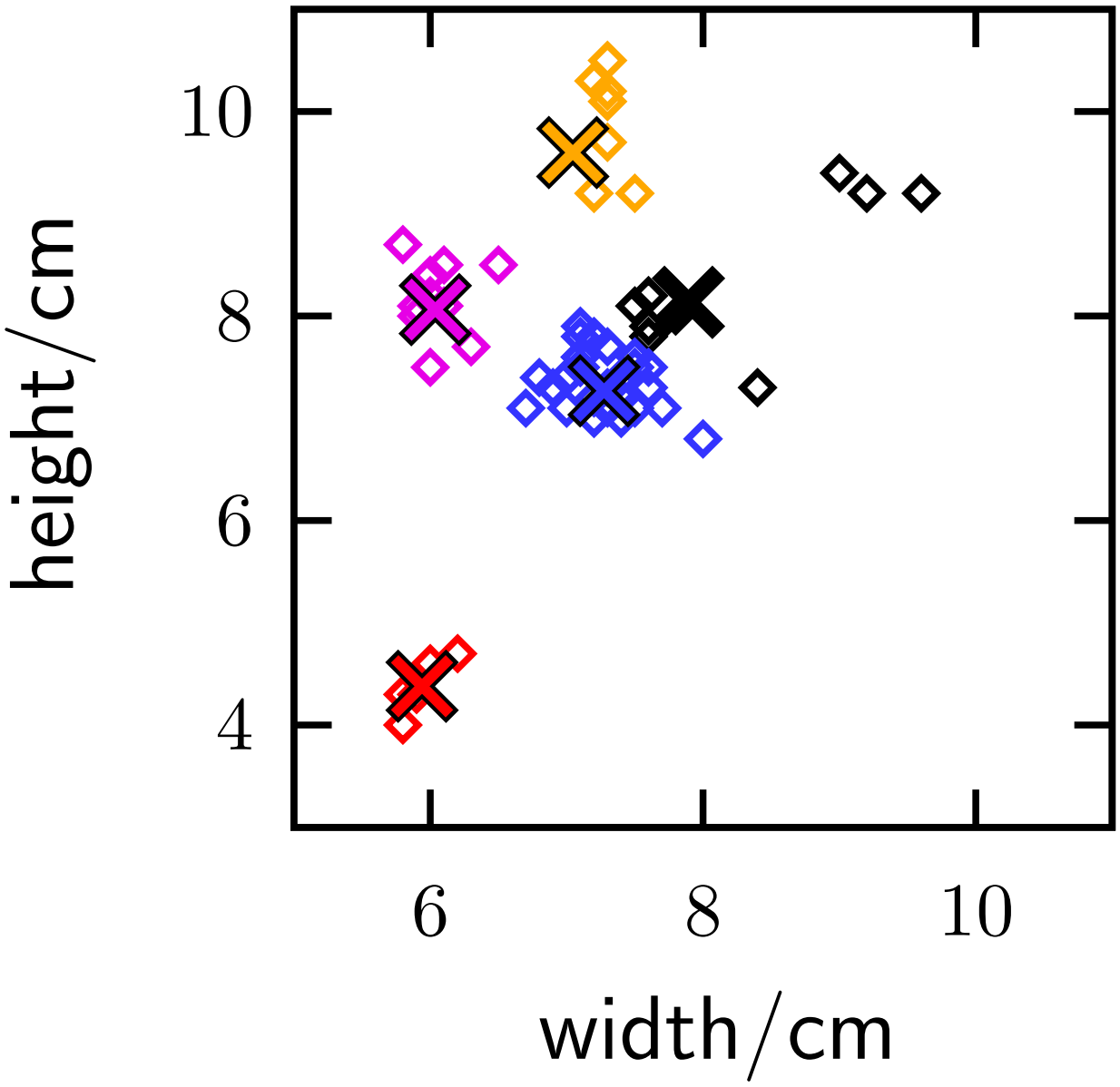
Clustering



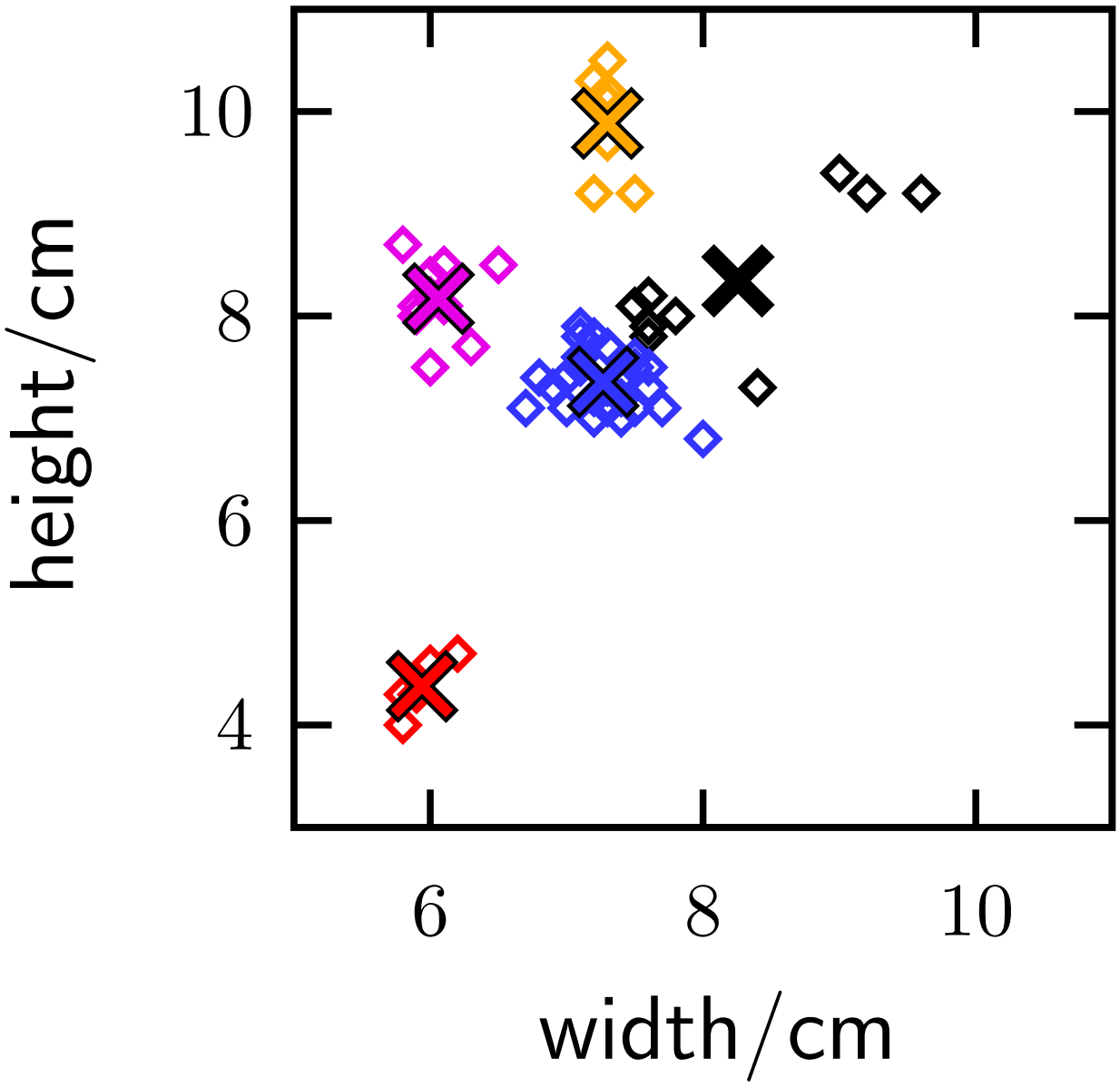
Clustering



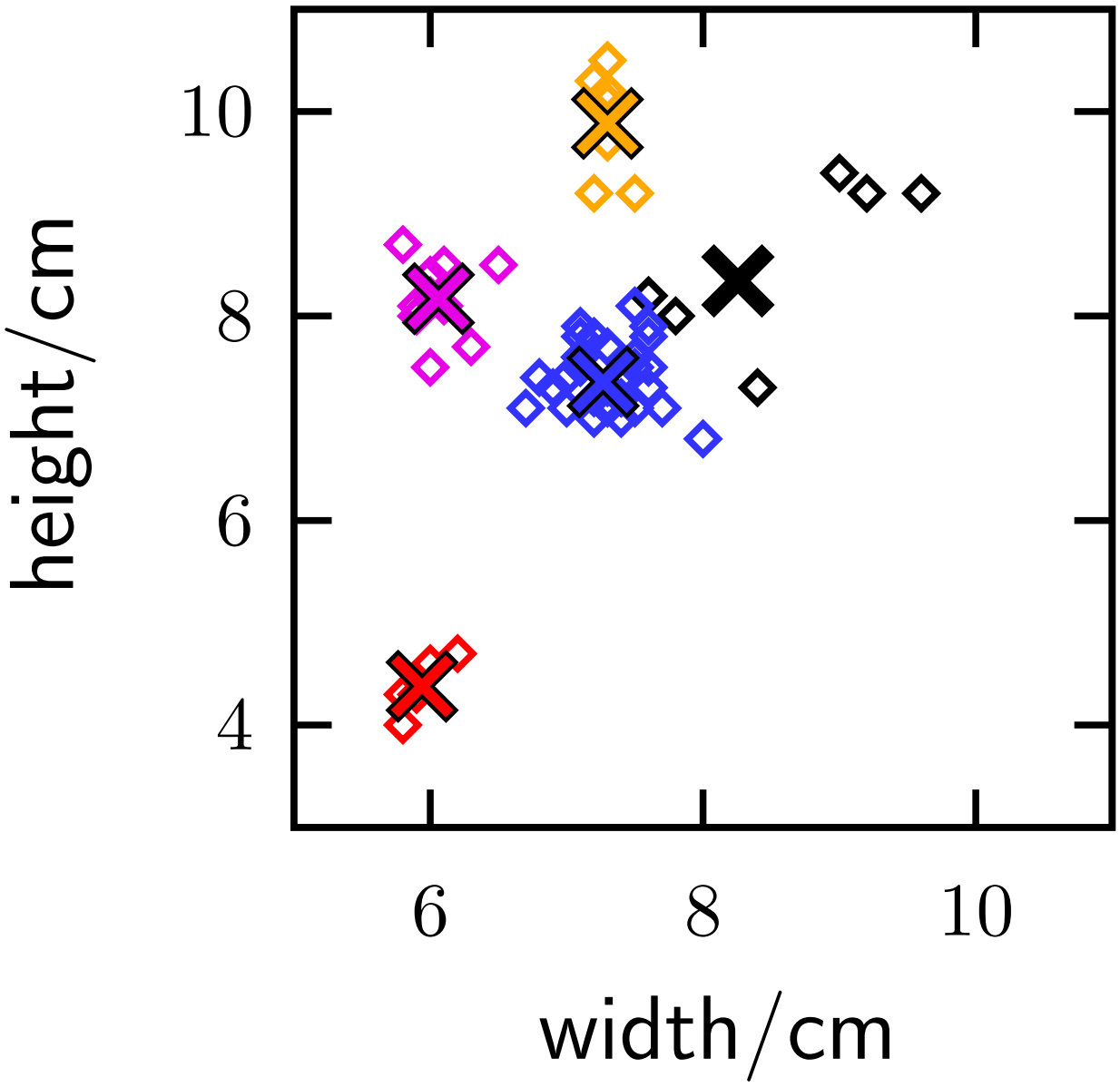
Clustering



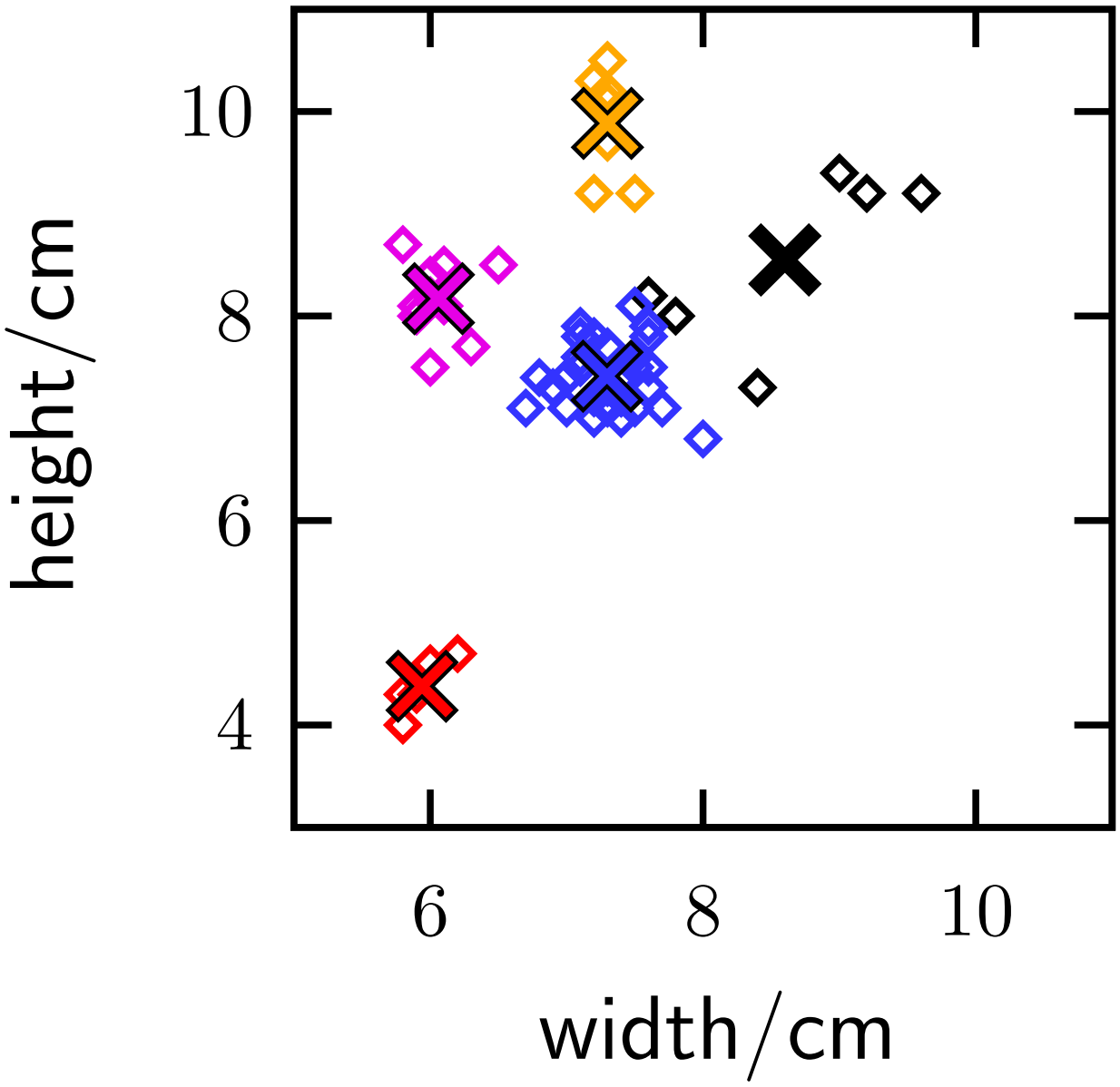
Clustering



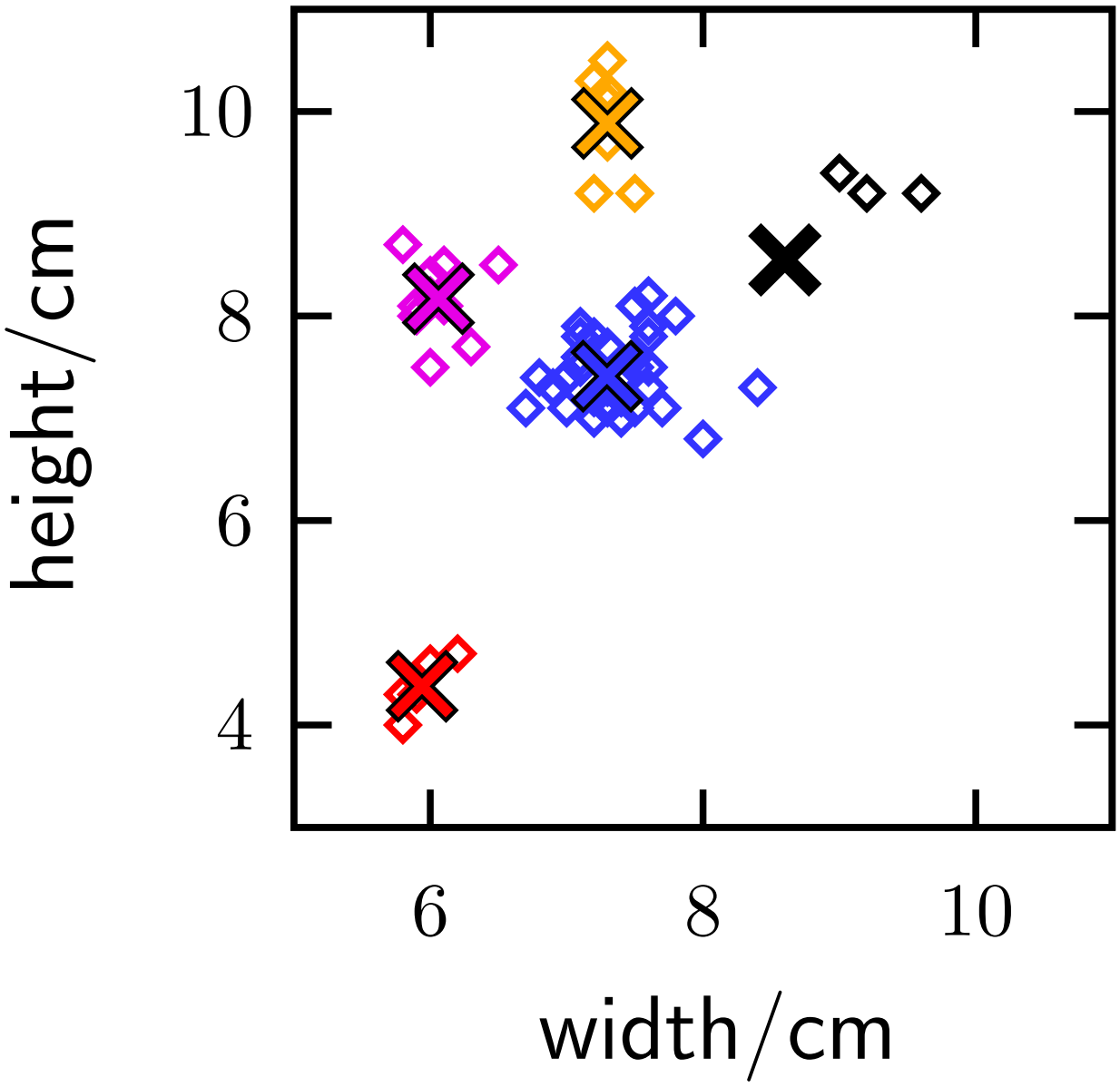
Clustering



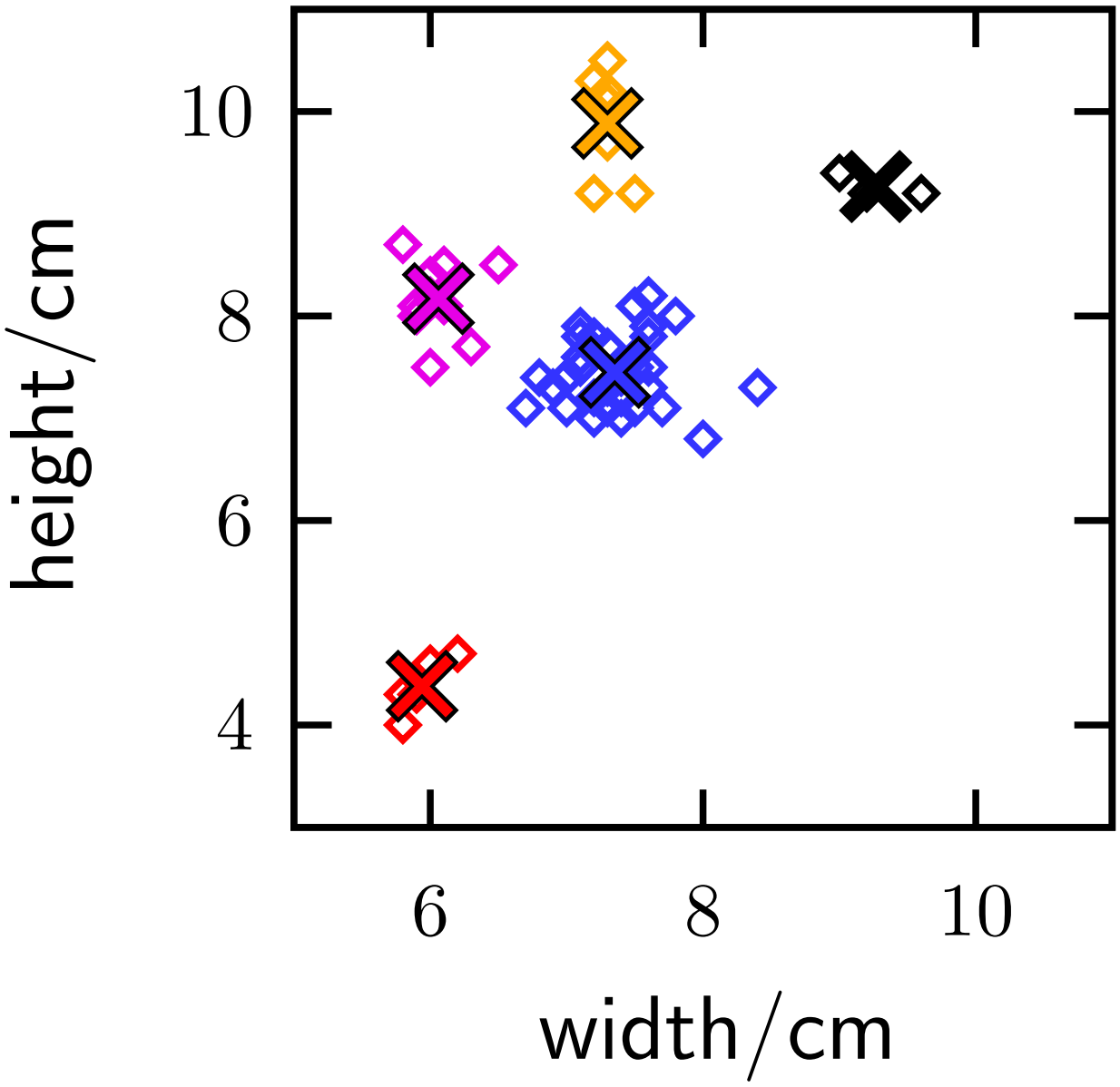
Clustering



Clustering



Clustering



A Cost Function for K-means

Let $s_{nk} = 1$ if data point n is assigned to cluster k and zero otherwise.

Note: $\sum_{k=1}^{K} s_{nk} = 1$.

Cost

$$C = \sum_{n=1}^{N} \sum_{k=1}^{K} s_{nk} \|\mathbf{x}_n - \mathbf{m}_k\|^2$$

The K-means algorithm tries to minimize the cost function C with respect to $\{s_{nk}\}$ and $\{\mathbf{m}_k\}$, subject to $\sum_k s_{nk} = 1$ and $s_{nk} \in \{0, 1\}$.

K-means:

- minimize C with respect to $\{s_{nk}\}$, holding $\{\mathbf{m}_k\}$ fixed.
- minimize C with respect to $\{\mathbf{m}_k\}$, holding $\{s_{nk}\}$ fixed.

Finding the global optimum of C is a *hard* problem.

A probabilistic interpretation of K-means

Multivariate Gaussian density ($\mathbf{x} \in \Re^D$):

$$p(\mathbf{x}|\mu, \Sigma) = |2\pi\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu)\right\}$$

Multivariate Gaussian density with mean $\mathbf{m}_k$ and identity covariance matrix I.

$$\begin{aligned} p(\mathbf{x}|\mathbf{m}_k) &= |2\pi I|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mathbf{m}_k)^\top (\mathbf{x}-\mathbf{m}_k)\right\} \\ p(\mathbf{x}|\mathbf{m}_k) &= \frac{1}{(2\pi)^{D/2}} \exp\left\{-\frac{1}{2}\|\mathbf{x}-\mathbf{m}_k\|^2\right\} \end{aligned}$$

A mixture model:

$$p(\mathbf{x}_n|\{\mathbf{m}_k\}) = \sum_k w_k \; p(\mathbf{x}_n|\mathbf{m}_k)$$

where w_k is the mixing proportion (e.g. set $w_k = 1/K$).

A probabilistic interpretation of K-means

Multivariate Gaussian density with mean $\mathbf{m}_k$ and identity covariance matrix I.

$$p(\mathbf{x}|\mathbf{m}_k) = \frac{1}{(2\pi)^{D/2}} \exp\left\{-\frac{1}{2}\|\mathbf{x} - \mathbf{m}_k\|^2\right\}$$

A mixture model:

$$p(\mathbf{x}_n|\{\mathbf{m}_k\}) = \sum_k w_k \, p(\mathbf{x}_n|\mathbf{m}_k)$$

where w_k is the mixing proportion (e.g. set $w_k = 1/K$).

Imagine we observed which data points came from which Gaussians (i.e. we knew $\{s_{nk}\}$), then:

$$p(\mathbf{x}_n, \mathbf{s}_n|\{\mathbf{m}_k\}) = \prod_k \left[w_k \, p(\mathbf{x}_n|\mathbf{m}_k)\right]^{s_{nk}}$$

Likelihood:

$$p(\mathbf{X}, \mathbf{S}|\{\mathbf{m}_k\}) = \prod_n p(\mathbf{x}_n, \mathbf{s}_n|\{\mathbf{m}_k\}) = \prod_{nk} \left[w_k \, p(\mathbf{x}_n|\mathbf{m}_k)\right]^{s_{nk}}$$

A probabilistic interpretation of K-means

Multivariate Gaussian density with mean $\mathbf{m}_k$ and identity covariance matrix I.

$$p(\mathbf{x}|\mathbf{m}_k) = \frac{1}{(2\pi)^{D/2}} \exp\left\{-\frac{1}{2}\|\mathbf{x} - \mathbf{m}_k\|^2\right\}$$

Likelihood: $$p(\mathbf{X}, \mathbf{S}|\{\mathbf{m}_k\}) = \prod_n p(\mathbf{x}_n, \mathbf{s}_n|\{\mathbf{m}_k\}) = \prod_{n,k} \left[w_k\, p(\mathbf{x}_n|\mathbf{m}_k)\right]^{s_{nk}}$$

Log Likelihood if we set $w_k = 1/K$:

$$\begin{aligned}
\ln p(\mathbf{X}, \mathbf{S}|\{\mathbf{m}_k\}) &= \sum_{n,k} s_{nk} \left[\log w_k + \log p(\mathbf{x}_n|\mathbf{m}_k)\right] \\
&= \sum_{n,k} s_{nk} \log p(\mathbf{x}_n|\mathbf{m}_k) - N \log K \\
&= -\frac{1}{2}\left[\sum_{n,k} s_{nk}\|\mathbf{x}_n - \mathbf{m}_k\|^2\right] - \frac{ND}{2}\log(2\pi) - N \log K
\end{aligned}$$

Maximizing $\ln p(\mathbf{X}, \mathbf{S}|\{\mathbf{m}_k\})$ with respect to $\{s_{nk}\}$ and $\{\mathbf{m}_k\}$ is equivalent to minimizing the K-means cost function. (Note: better to treat the $\{s_{nk}\}$ as unknown variables).