10-701 Recitation:
Probabilistic and graphical models

Abulhair Saparov

Probabilistic modeling

- There are a lot of different naming conventions.
  - Seems to be the case a lot in machine learning
  - and outside machine learning as well…
- I will present some definitions using nomenclature that I think is intuitive, but also not far off from what other people use.

What is a **probabilistic model**?

Probabilistic modeling

A **probabilistic model** is a collection of random variables.

The random variables can be divided into two categories:

1. the **observations** (data)
2. the **hidden variables**

Intuitively, a probabilistic model is a *description* of how your observations were generated.

It can be a hypothesis of the mechanism that underlies your data.

Probabilistic modeling

For more intuition, imagine we have a system of equations.

$$x + 2y = 4$$

If x is 0, what is y?

If $x \sim \mathcal{N}(0,1)$, what is y?

You can think of a probabilistic model as a system where the variables can be random.

Probabilistic modeling

A **probabilistic model** is a collection of random variables: **{x,θ}**

The random variables can be divided into two categories:

1. **x:** the **observations** (data)
2. **θ:** the **hidden variables**

the **prior** distribution is p(**θ**)

the **likelihood** is p(**x|θ**)

the **posterior** distribution is p(**θ|x**)

the **joint** distribution is p(**x,θ**)

So then, what's a graphical model?

A **graphical model** is a graphical representation of a probabilistic model.

There are different ways to represent probabilistic models graphically:

- Bayesian networks
- Factor graphs
- Markov random fields

When people say "graphical model", they usually mean

graph + probabilistic model.

Common convention: observations are depicted as shaded nodes, whereas hidden variables are unshaded.

Example

probabilistic model

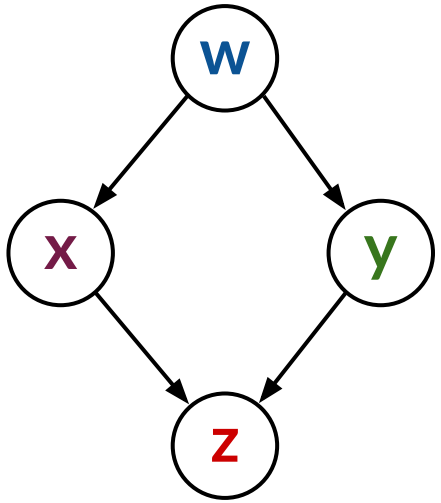
$w \sim \text{Uniform}(0,1)$

$x \sim \text{Bernoulli}(w)$

$y \sim \text{Bernoulli}(w)$

$z = x + y$

graphical model



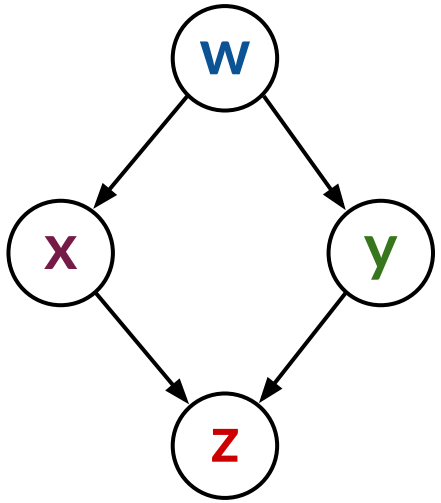
You can think of **x** and **y** as two coin flips, and **w** represents the fairness of the coin.

Example

It is easy to factorize the joint distribution by looking at the graph structure:

$$p(w,x,y,z) = p(w)p(x|w)p(y|w)p(z|x,y)$$

graphical model



$w \sim \text{Uniform}(0,1)$

$x \sim \text{Bernoulli}(w)$

$y \sim \text{Bernoulli}(w)$

$z = x + y$

Example

What is $p(x|w)$?

$$p\{x = 0|w\} = 1 - w \text{ and } p\{x = 1|w\} = w$$

$$p\{y = 0|w\} = 1 - w \text{ and } p\{y = 1|w\} = w$$

What is $p(z|x,y,w) = p(z|x,y)$?

$$p(z|x,y) = \delta\{z = x + y\}$$

What if we want to get rid of x and y? We can **marginalize** out x and y.

Lets compute $p(z|w)$.

Example

w ~ Uniform(0,1)
x ~ Bernoulli(w)
y ~ Bernoulli(w)
z = x + y

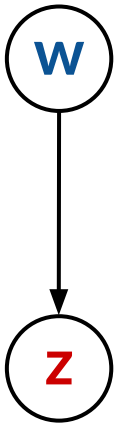
$$
\begin{aligned}
p\{z=0|w\} &= \sum_{x\in\{0,1\}} p\{z=0|x,w\}p(x|w), \\
&= \sum_{x\in\{0,1\}}\sum_{y\in\{0,1\}} p\{z=0|x,y,w\}p(x|w)p(y|w), \\
&= \sum_{x\in\{0,1\}}\sum_{y\in\{0,1\}} \delta\{x=0,y=0\}p(x|w)p(y|w), \\
&= p\{x=0|w\}p\{y=0|w\}, \\
&= (1-w)^2.
\end{aligned}
$$

Example

$w \sim \text{Uniform}(0,1)$
$x \sim \text{Bernoulli}(w)$
$y \sim \text{Bernoulli}(w)$
$z = x + y$

Repeating the process for the other possible values of z:

$$p\{z = 0|w\} = (1 - w)^2$$
$$p\{z = 1|w\} = 2w(1 - w)$$
$$p\{z = 2|w\} = w^2$$

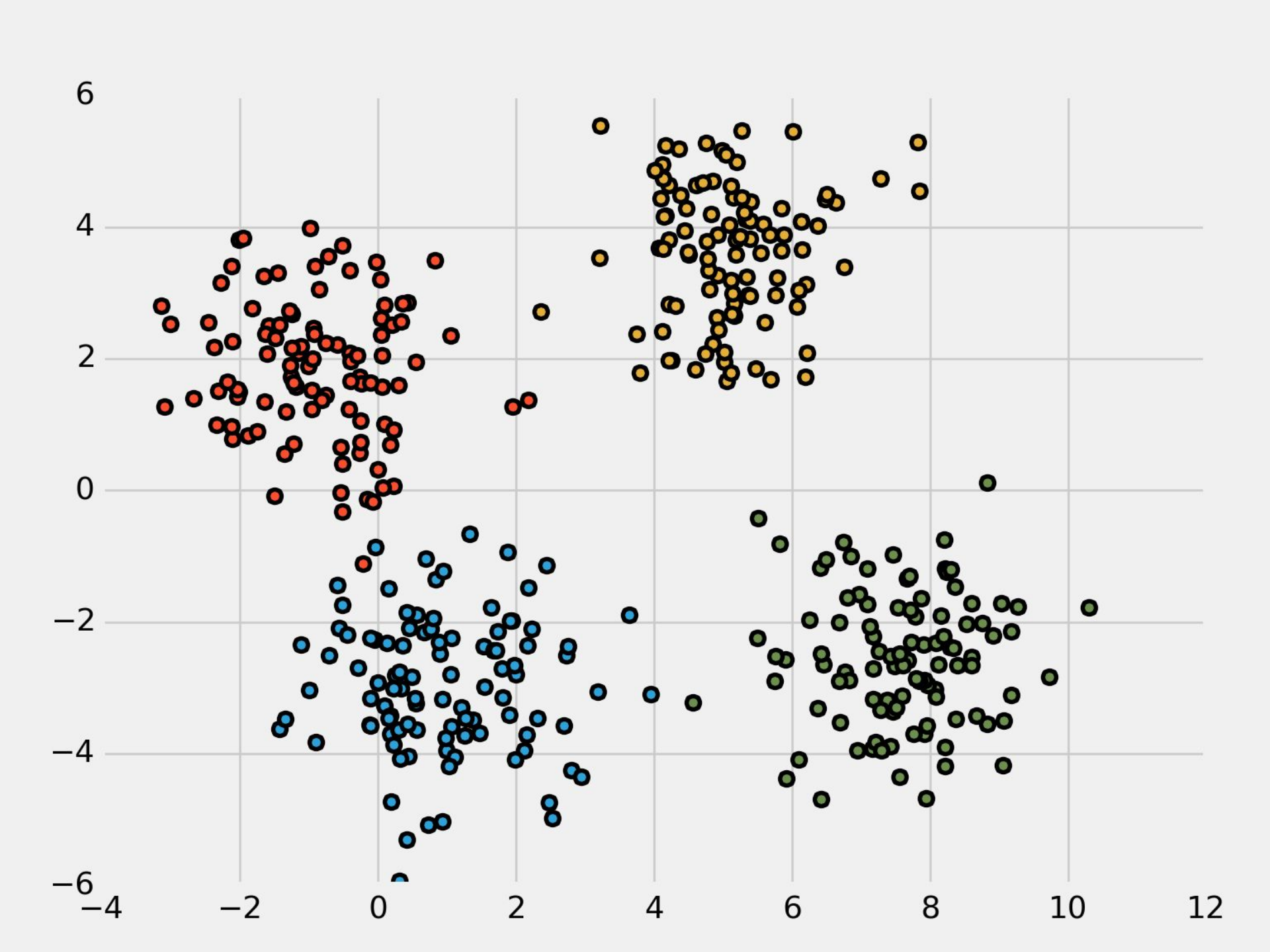


This is a new model:

$w \sim \text{Uniform}(0,1)$

$z \sim$ the above discrete distribution

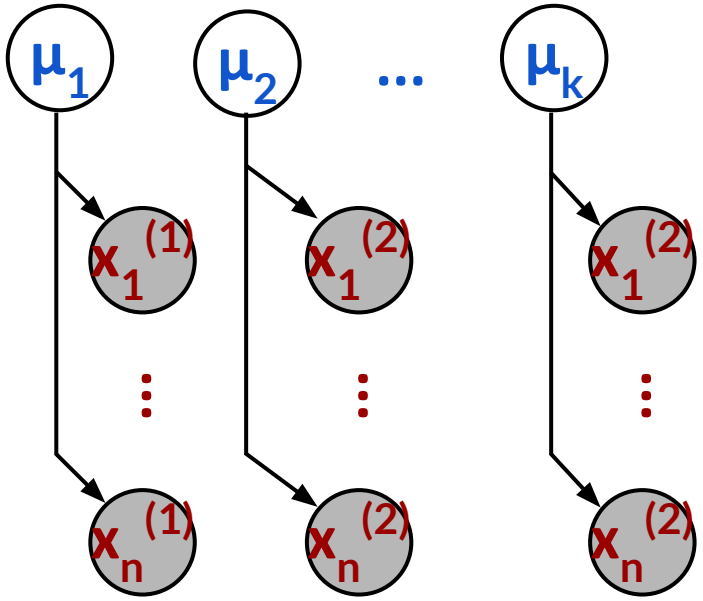
The reasoning and intuition behind variable elimination and belief propagation is the same.


6
4
2
0
−2
−4
−6
−4
−2
0
2
4
6
8
10
12

What is a good probabilistic model for this data?

The points look like they came from four normal distributions.

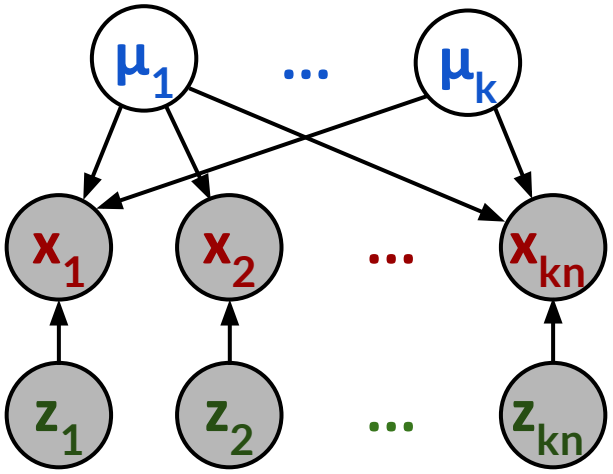
$$\mu_1, \ldots, \mu_k \sim \mathcal{N}(0, 10I),$$
$$x_1^{(1)}, \ldots, x_n^{(1)} \sim \mathcal{N}(\mu_1, I),$$
$$x_1^{(2)}, \ldots, x_n^{(2)} \sim \mathcal{N}(\mu_2, I),$$
$$\text{and so on} \ldots \text{for k clusters}$$

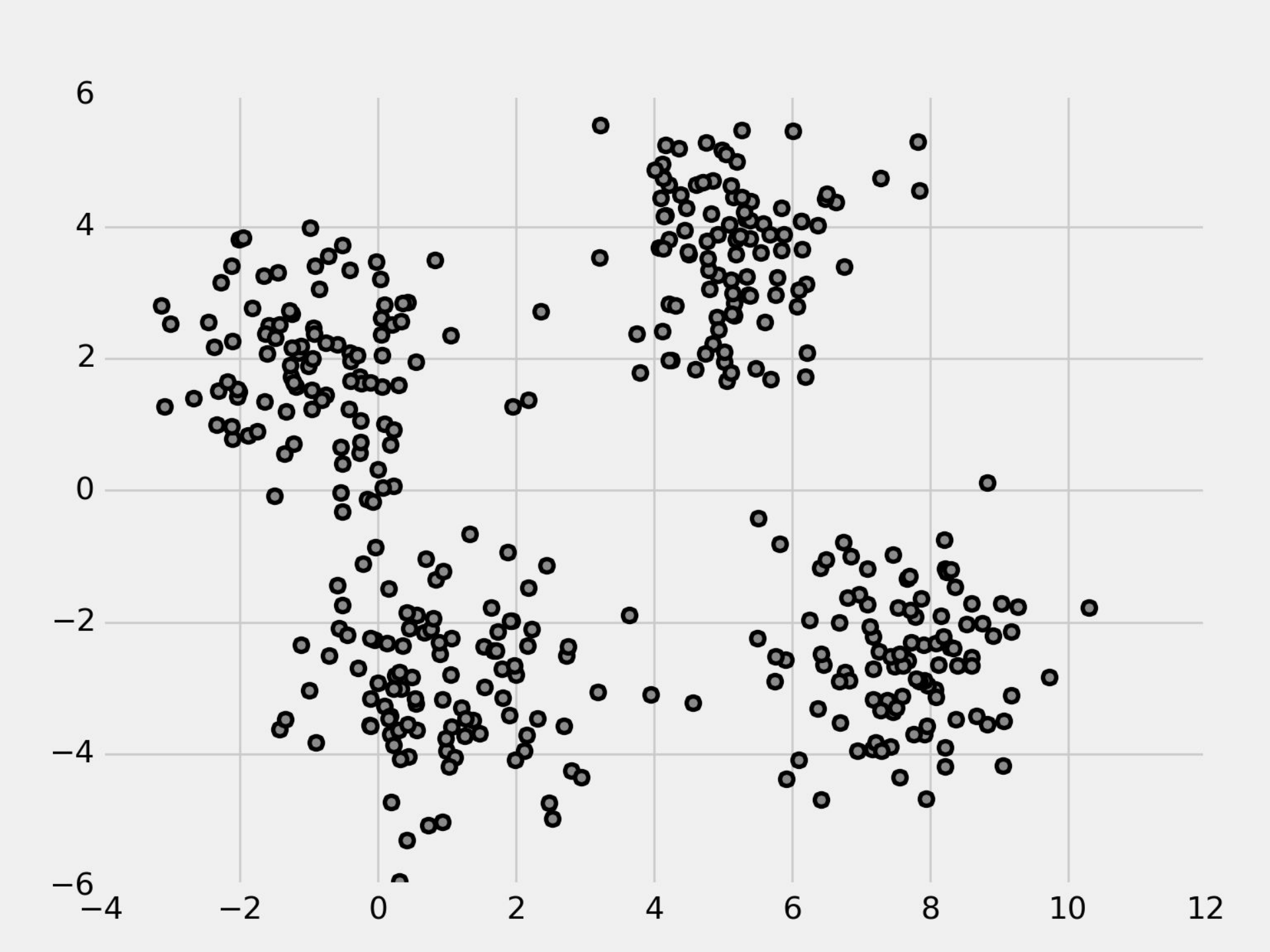


For simplicity, we fixed the covariances to the identity. If desired, you could make them unknown variables and use, for instance, an inverse-Wishart prior (see last week's recitation resources).

We can re-write this model:

$$\mu_1, \ldots, \mu_k \sim \mathcal{N}(0, 10I),$$
$$z_1, \ldots, z_n = 1,$$
$$z_{n+1}, \ldots, z_{2n} = 2,$$
$$\text{etc...}$$
$$x_i \sim \mathcal{N}(\mu_{z_i}, I) \text{ i.i.d. } i = 1, \ldots, kn.$$

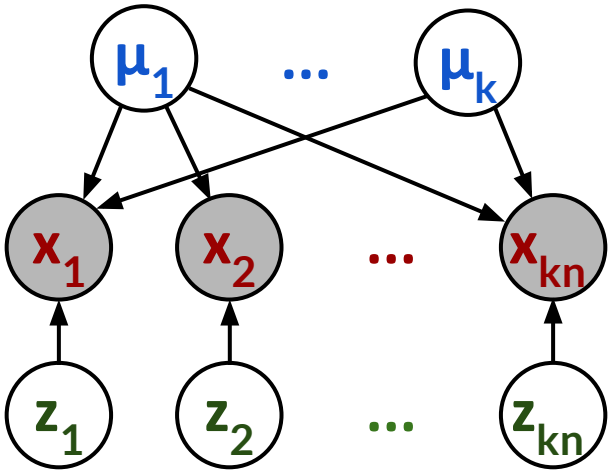


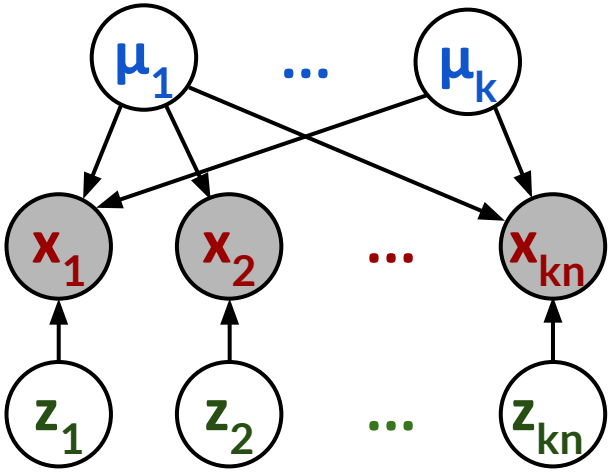

6
4
2
0
−2
−4
−6
−4
−2
0
2
4
6
8
10
12

What is a good model for this data?

We no longer have knowledge of the class assignments **z**.

$$\mu_1, \ldots, \mu_k \sim \mathcal{N}(0, 10I),$$
$$z_i = \text{Categorical}(\pi),$$
$$x_i \sim \mathcal{N}(\mu_{z_i}, I) \text{ i.i.d. } i = 1, \ldots, kn.$$





$$\mu_1, \ldots, \mu_k \sim \mathcal{N}(0, 10I),$$
$$z_i = \text{Categorical}(\pi),$$
$$x_i \sim \mathcal{N}(\mu_{z_i}, I) \text{ i.i.d. } i = 1, \ldots, kn.$$

Let's try to do inference using MAP in this model. Write the log-posterior: log p(μ_1,...,μ_k,z_1,...,z_{kn}|x_1,...,x_{kn}).

$$\log p(\boldsymbol{\mu}, \boldsymbol{z}|\boldsymbol{x}) = \log p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{z}) + \log p(\boldsymbol{\mu}) + \log p(\boldsymbol{z}) + C,$$
$$= \sum_{i=1}^{kn} \log p(x_i|\boldsymbol{\mu}, z_i) + \sum_{j=1}^{k} \log p(\mu_j) + \sum_{i=1}^{kn} \log p(z_i),$$
$$= -\frac{1}{2} \sum_{i=1}^{kn} (x_i - \mu_{z_i})^\top (x_i - \mu_{z_i}) - \frac{1}{20} \sum_{j=1}^{k} \mu_j^\top \mu_j + \sum_{i=1}^{kn} \log \pi_{z_i}.$$

Expectation maximization (EM)

- If we knew either μ or z, then MLE/MAP would be easier.
- EM is an inference algorithm for computing MLE or MAP.
- Given any probabilistic model with observations x and hidden variables θ, we first subdivide the hidden variables θ into two classes: z and μ.
- Start with a guess for μ.

E step. compute the expectation using our current estimate of μ:

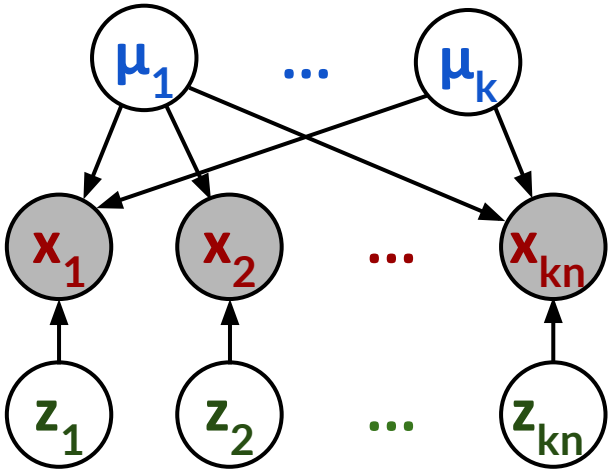
$$q(\mu) = E_{p(z|x,\mu^*)}[\log p(\mu,z|x)]$$

M step. update the estimate of μ^* by maximizing $q(\mu)$:

$$\mu^* = \arg\max q(\mu)$$

Repeat until convergence.

Expectation maximization

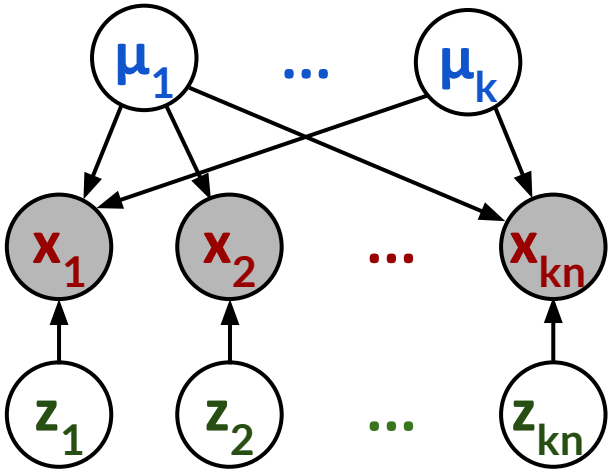


- We need to compute p(z|x,μ) to do the E step.

$$
\begin{aligned}
p(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\mu})p(\boldsymbol{x},\boldsymbol{\mu}) &= p(\boldsymbol{x},\boldsymbol{z},\boldsymbol{\mu}),\\
p(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\mu}) &= p(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{\mu})p(\boldsymbol{\mu})p(\boldsymbol{z})/p(\boldsymbol{x},\boldsymbol{\mu}),\\
\log p(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\mu}) &= \log p(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{\mu}) + \log p(\boldsymbol{z}) + C,\\
\log p\{z_i = j|\boldsymbol{x},\boldsymbol{\mu}\} &= \log p(x_i|z_i = j,\mu_j) + \log p\{z_i = j\} + C,\\
&= -\frac{1}{2}(x_i - \mu_j)^\top(x_i - \mu_j) + \log \pi_j + C,
\end{aligned}
$$

$$
p\{z_i = j|\boldsymbol{x},\boldsymbol{\mu}\} \propto \pi_j \exp\left\{-\frac{1}{2}(x_i - \mu_j)^\top(x_i - \mu_j)\right\}.
$$

Expectation maximization

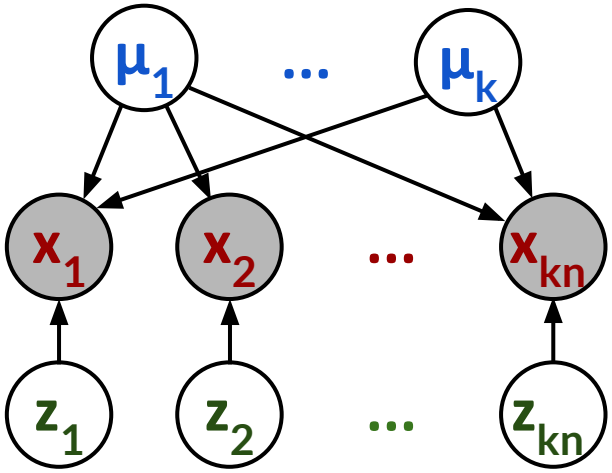


- We can use p(z|x,μ) to compute q.

$$
\begin{aligned}
q(\boldsymbol{\mu}) &= \mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\mu}^*)}[\log p(\boldsymbol{\mu}, \boldsymbol{z}|\boldsymbol{x})], \\
&= \mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\mu}^*)}[\log p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{z})] + \log p(\boldsymbol{\mu}) + C, \\
&= \sum_{i=1}^{kn} \mathbb{E}_{p(z_i|\boldsymbol{x},\boldsymbol{\mu}^*)}[\log p(x_i|\boldsymbol{\mu}, z_i)] + \log p(\boldsymbol{\mu}) + C, \\
&= -\frac{1}{2} \sum_{i=1}^{kn} \mathbb{E}_{p(z_i|\boldsymbol{x},\boldsymbol{\mu}^*)}[(x_i - \mu_{z_i})^\top (x_i - \mu_{z_i})] + \log p(\boldsymbol{\mu}) + C, \\
&= -\frac{1}{2} \sum_{i=1}^{kn} \sum_{j=1}^{k} p\{z_i = j|\boldsymbol{x}, \boldsymbol{\mu}^*\}(x_i - \mu_j)^\top (x_i - \mu_j) + \log p(\boldsymbol{\mu}) + C,
\end{aligned}
$$

$$
q(\mu_j) = -\frac{1}{2} \sum_{i=1}^{kn} p\{z_i = j|\boldsymbol{x}, \boldsymbol{\mu}^*\}(x_i - \mu_j)^\top (x_i - \mu_j) - \frac{1}{20} \mu_j^\top \mu_j + C.
$$

Expectation maximization



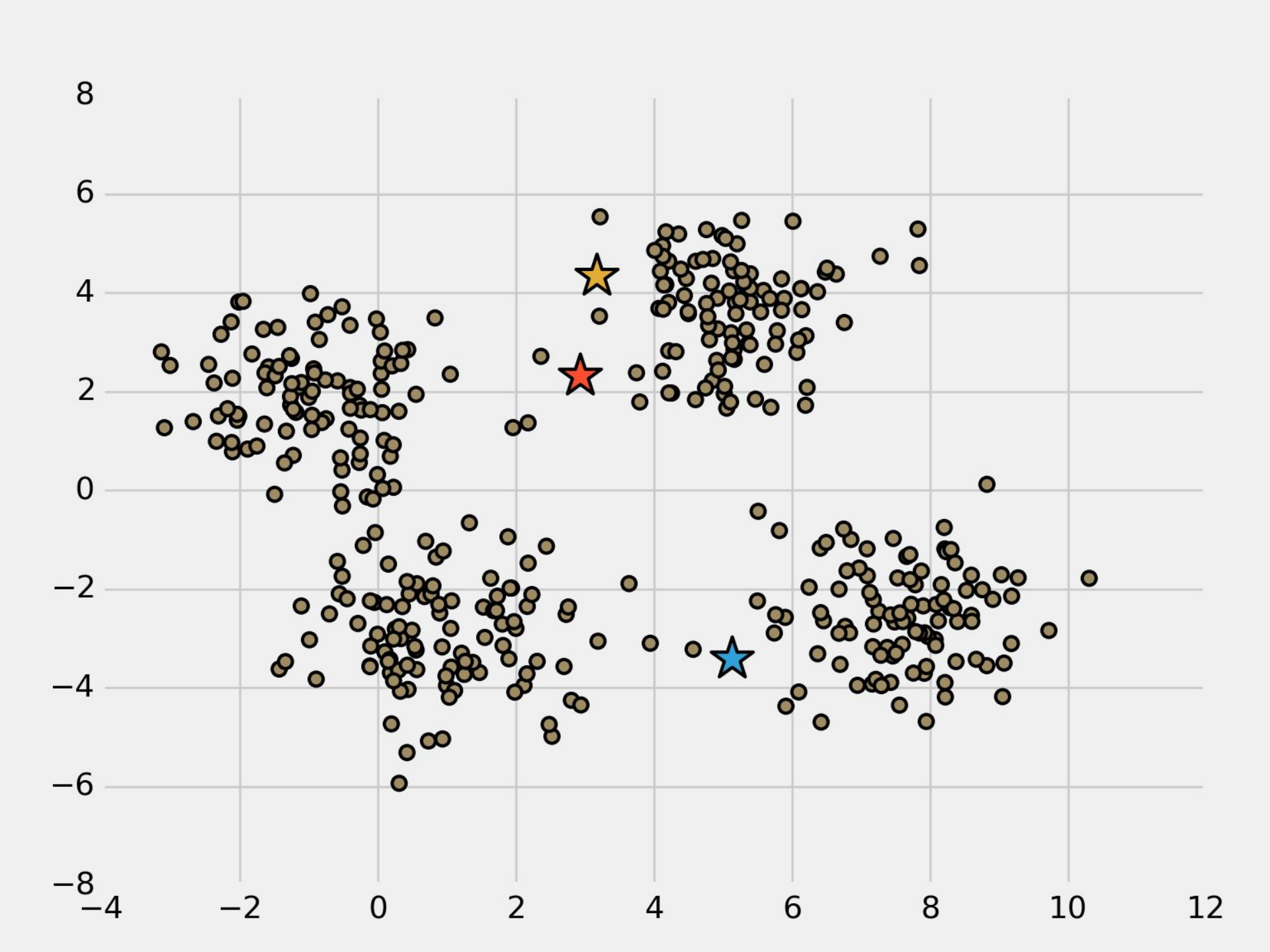
- Now, for the M step, we just maximize q.

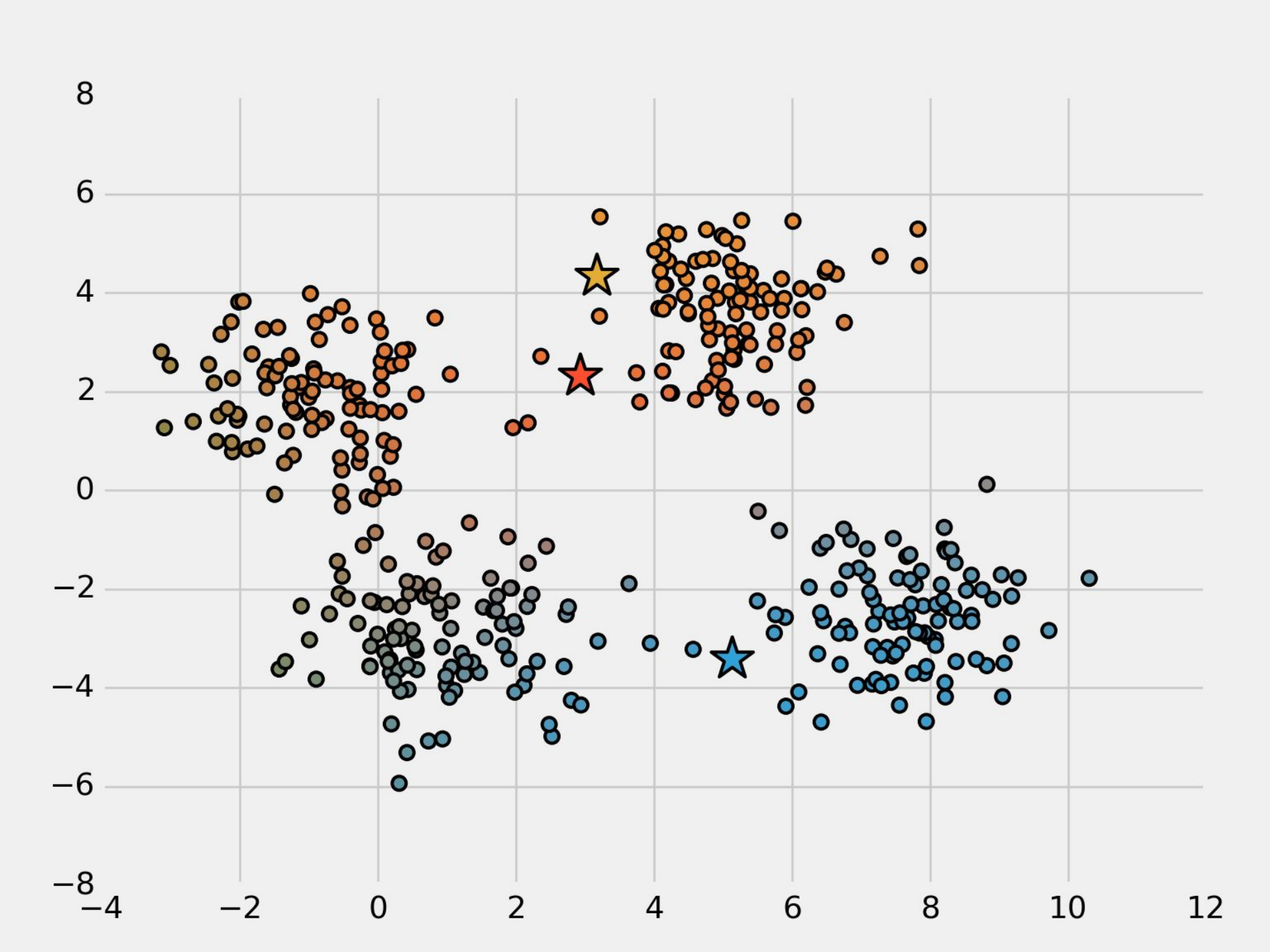
$$\mu_j^* = \arg\max_{\mu_j} q(\mu_j),$$

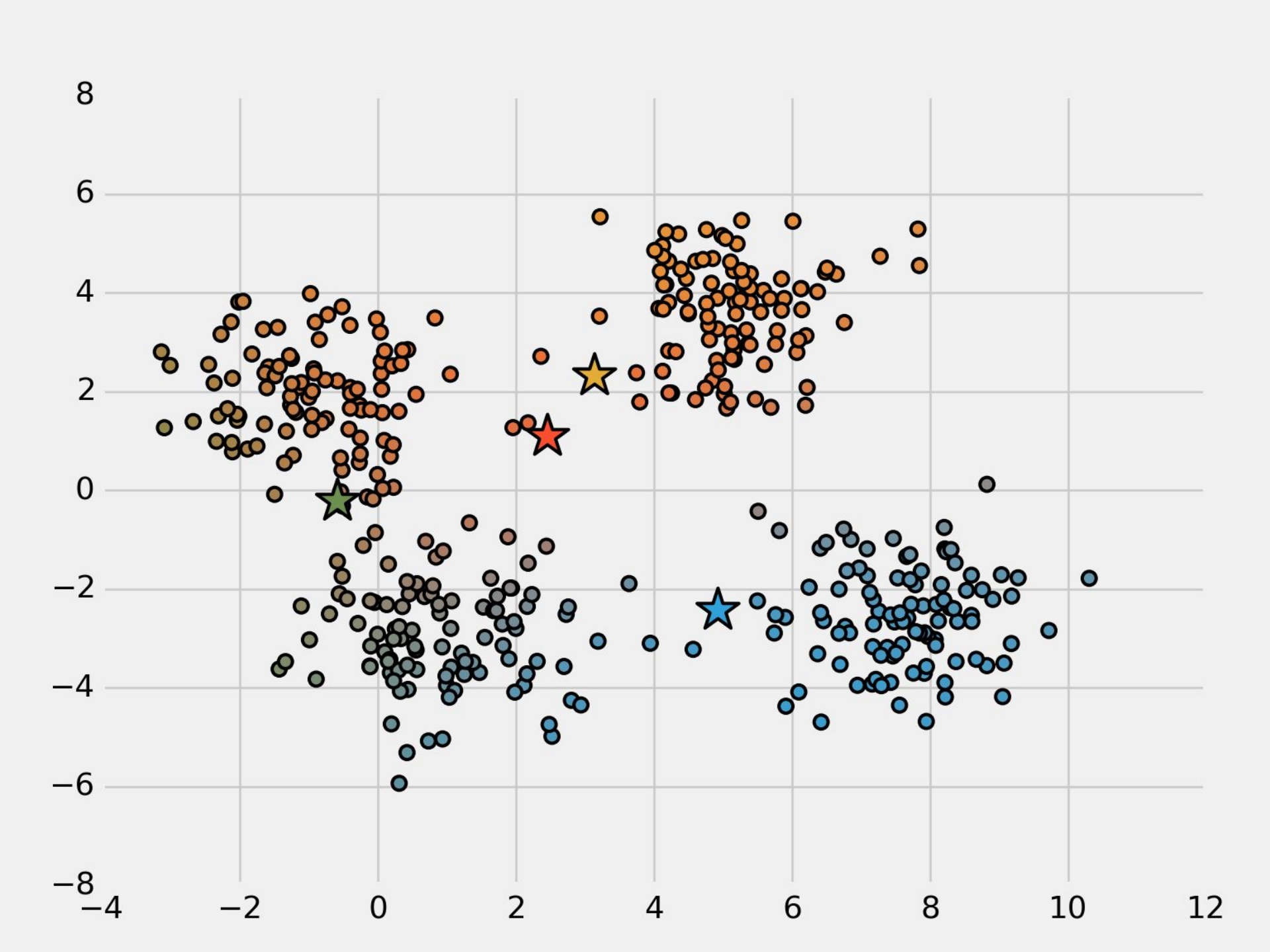
$$\frac{\partial q}{\partial \mu_j} = \sum_{i=1}^{kn} p\{z_i = j|\boldsymbol{x}, \boldsymbol{\mu}^*\}(x_i - \mu_j) - \frac{1}{10}\mu_j,$$

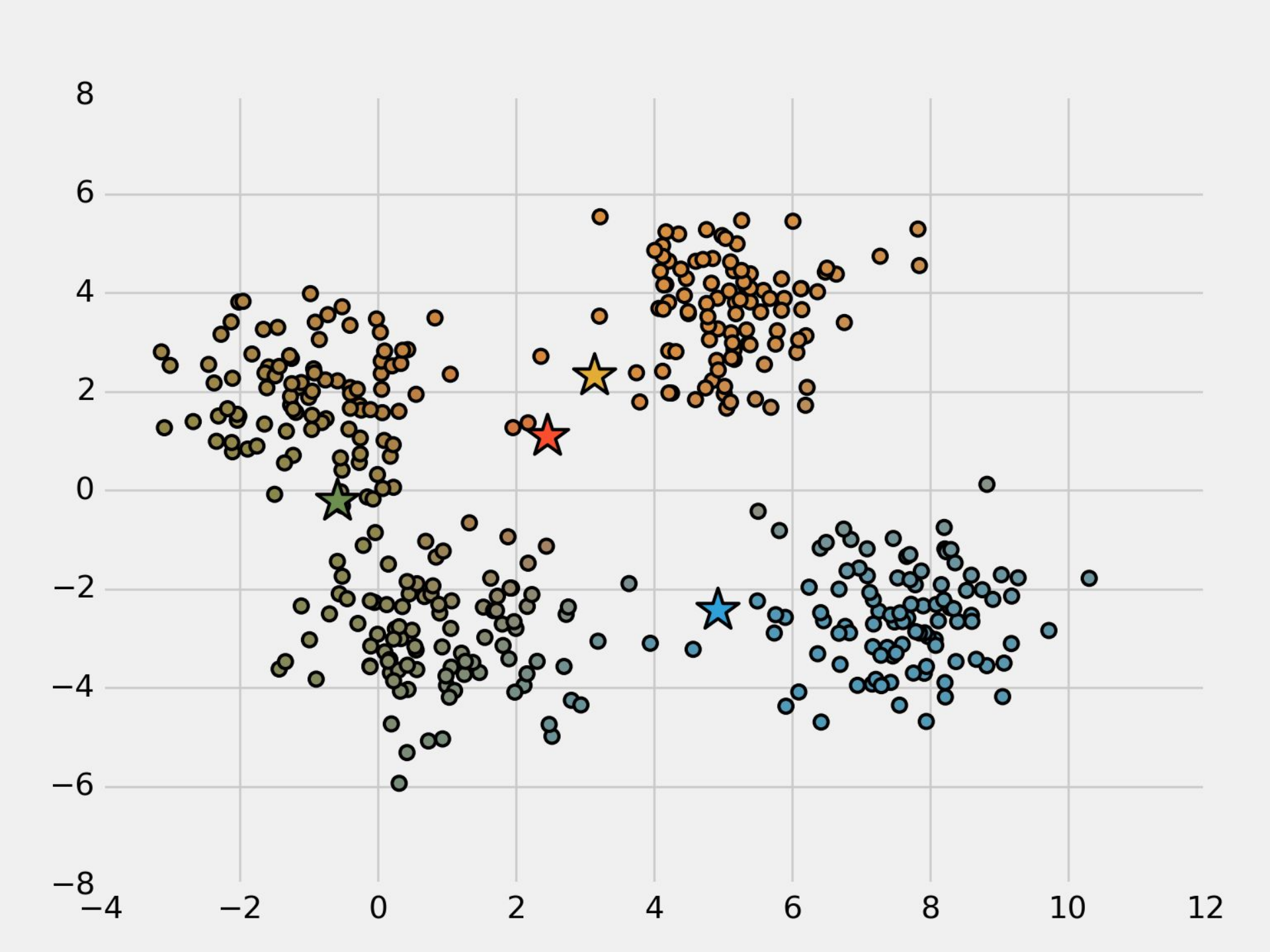
$$0 = \sum_{i=1}^{kn} p\{z_i = j|\boldsymbol{x}, \boldsymbol{\mu}^*\}x_i - \mu_j^* \left(\frac{1}{10} + \sum_{i=1}^{kn} p\{z_i = j|\boldsymbol{x}, \boldsymbol{\mu}^*\}\right),$$

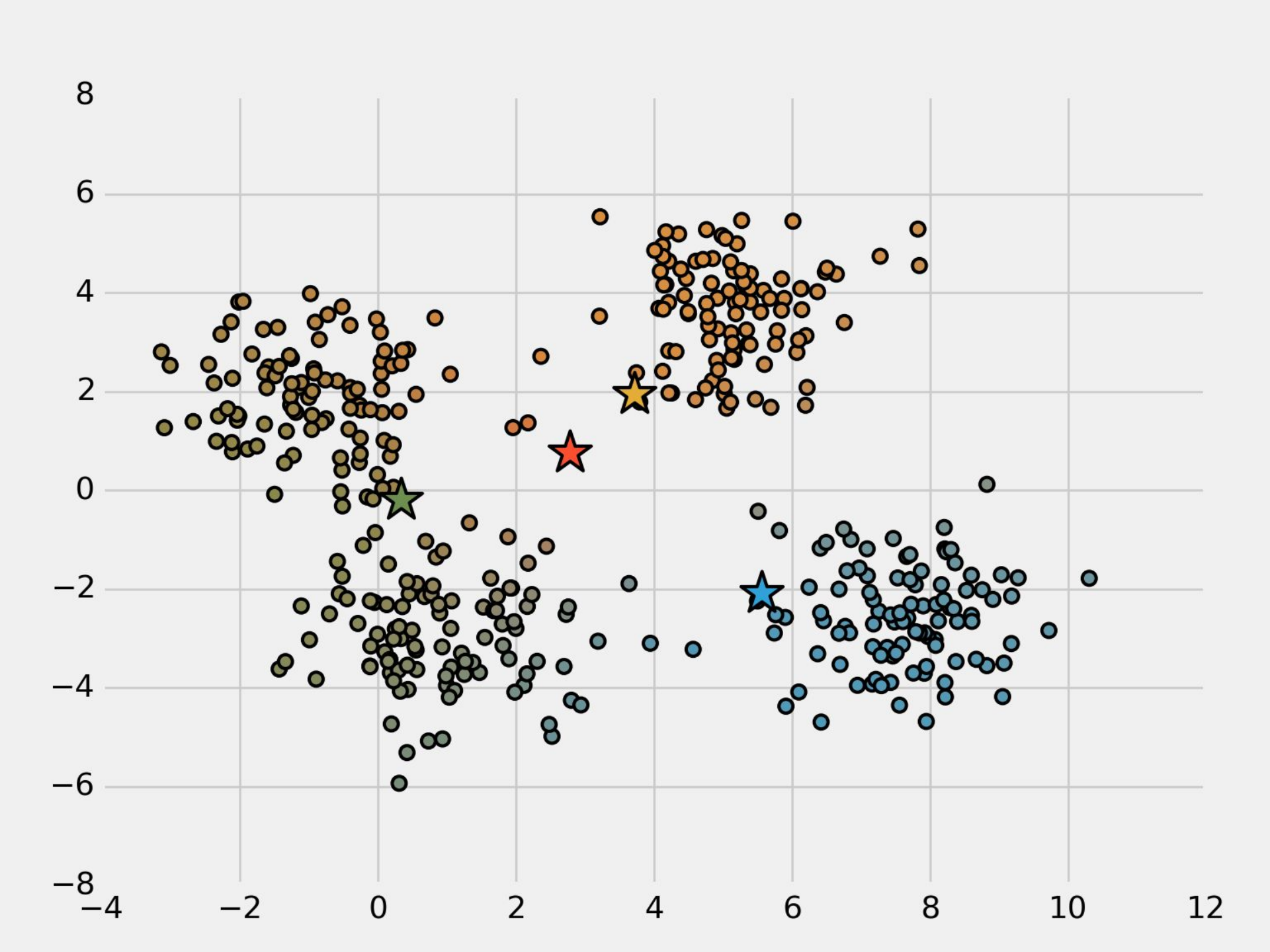
$$\mu_j^* = \left(\frac{1}{10} + \sum_{i=1}^{kn} p\{z_i = j|\boldsymbol{x}, \boldsymbol{\mu}^*\}\right)^{-1} \sum_{i=1}^{kn} p\{z_i = j|\boldsymbol{x}, \boldsymbol{\mu}^*\}x_i$$

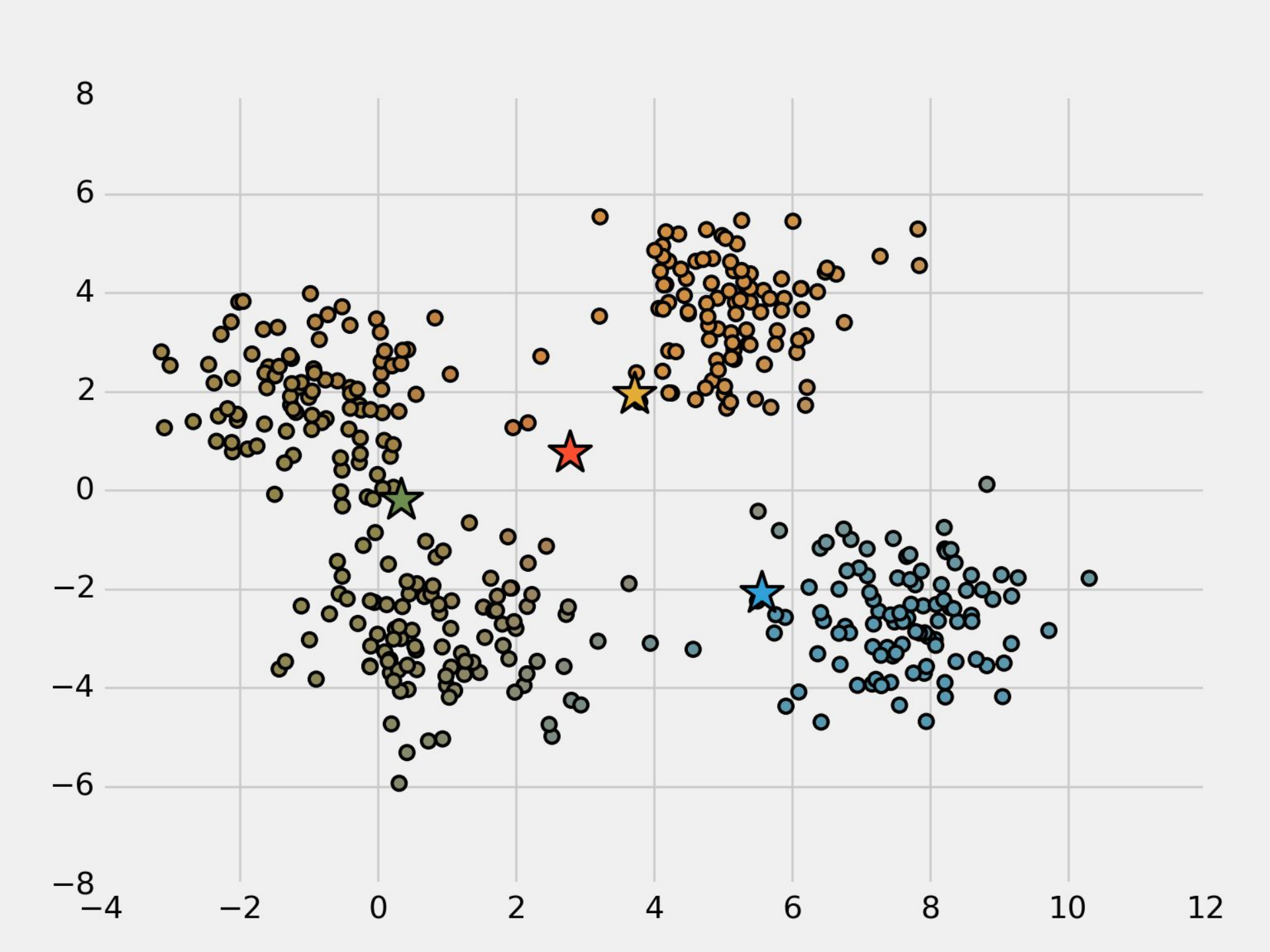

8
6
4
2
0
−2
−4
−6
−8
−4
−2
0
2
4
6
8
10
12

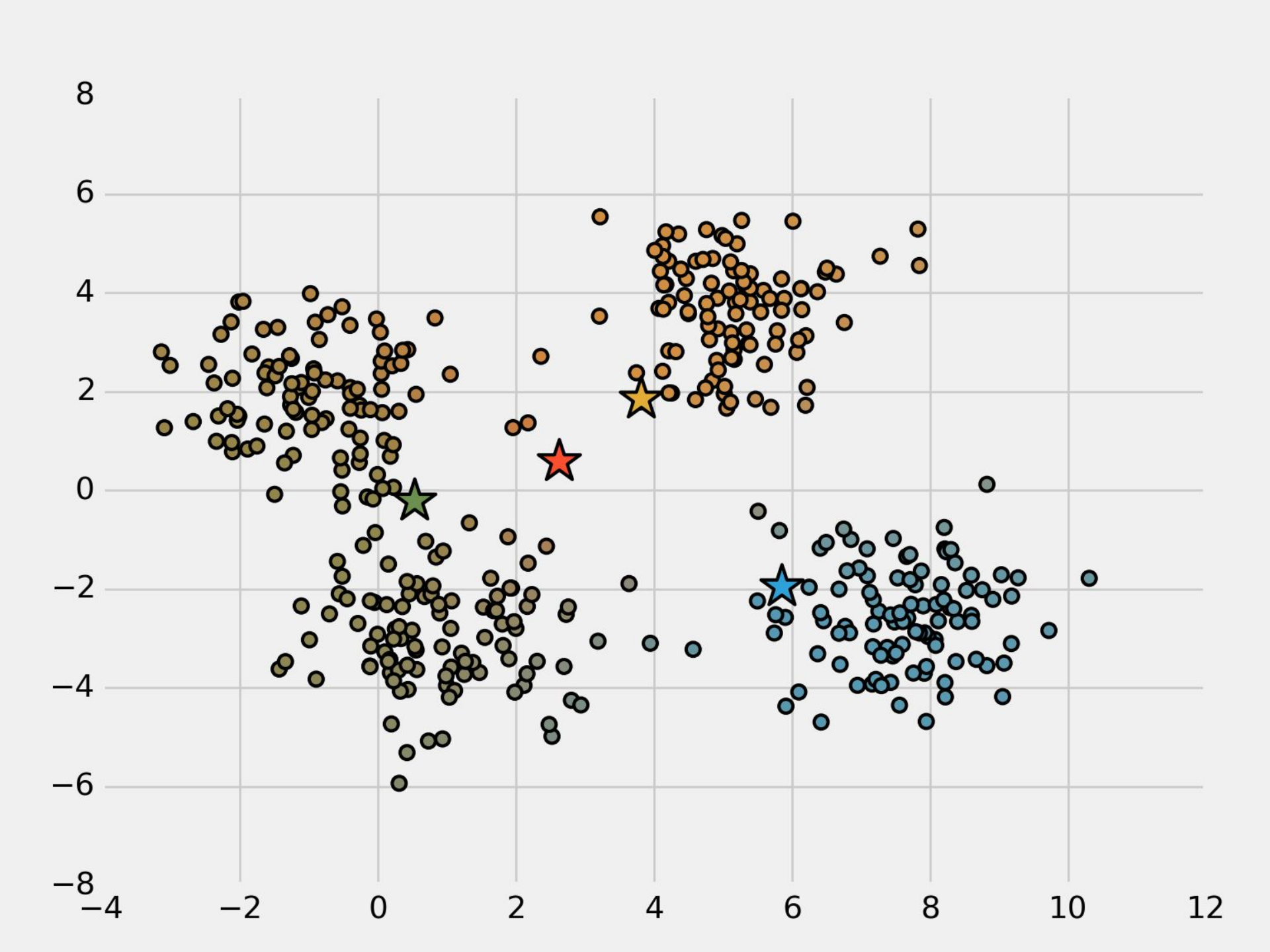

8
6
4
2
0
−2
−4
−6
−8
−4
−2
0
2
4
6
8
10
12

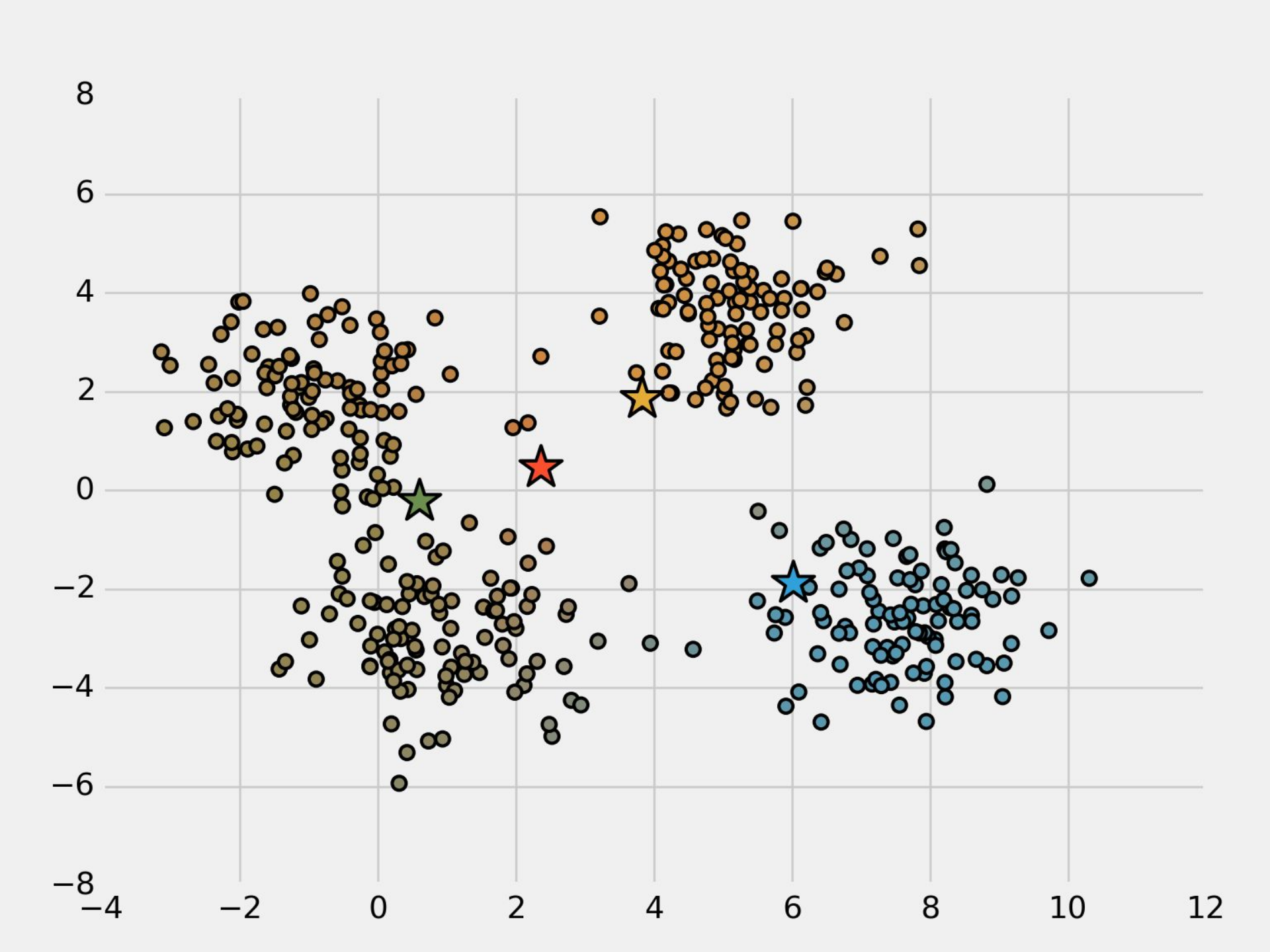

8
6
4
2
0
−2
−4
−6
−8
−4
−2
0
2
4
6
8
10
12

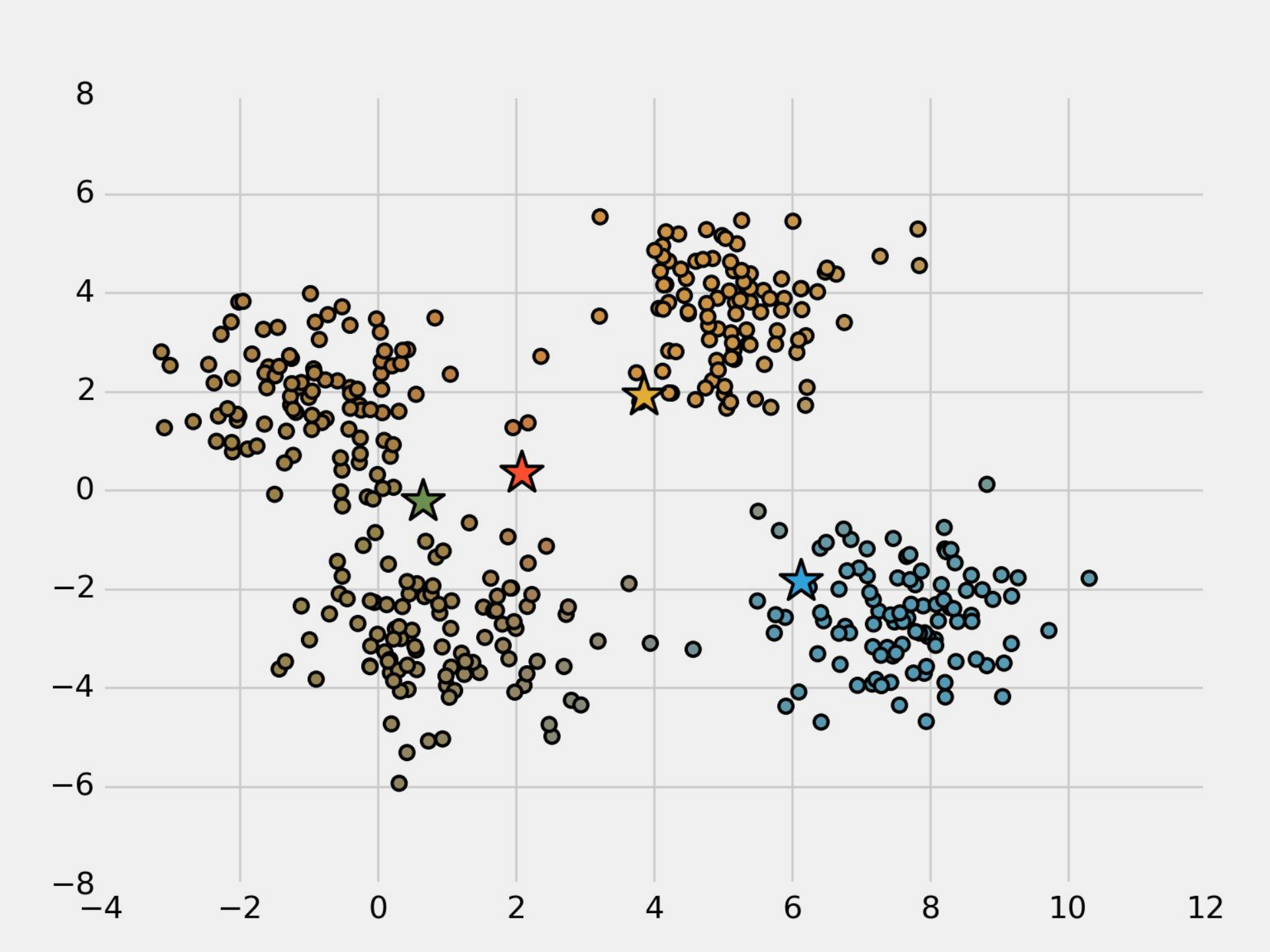

8
6
4
2
0
−2
−4
−6
−8
−4
−2
0
2
4
6
8
10
12

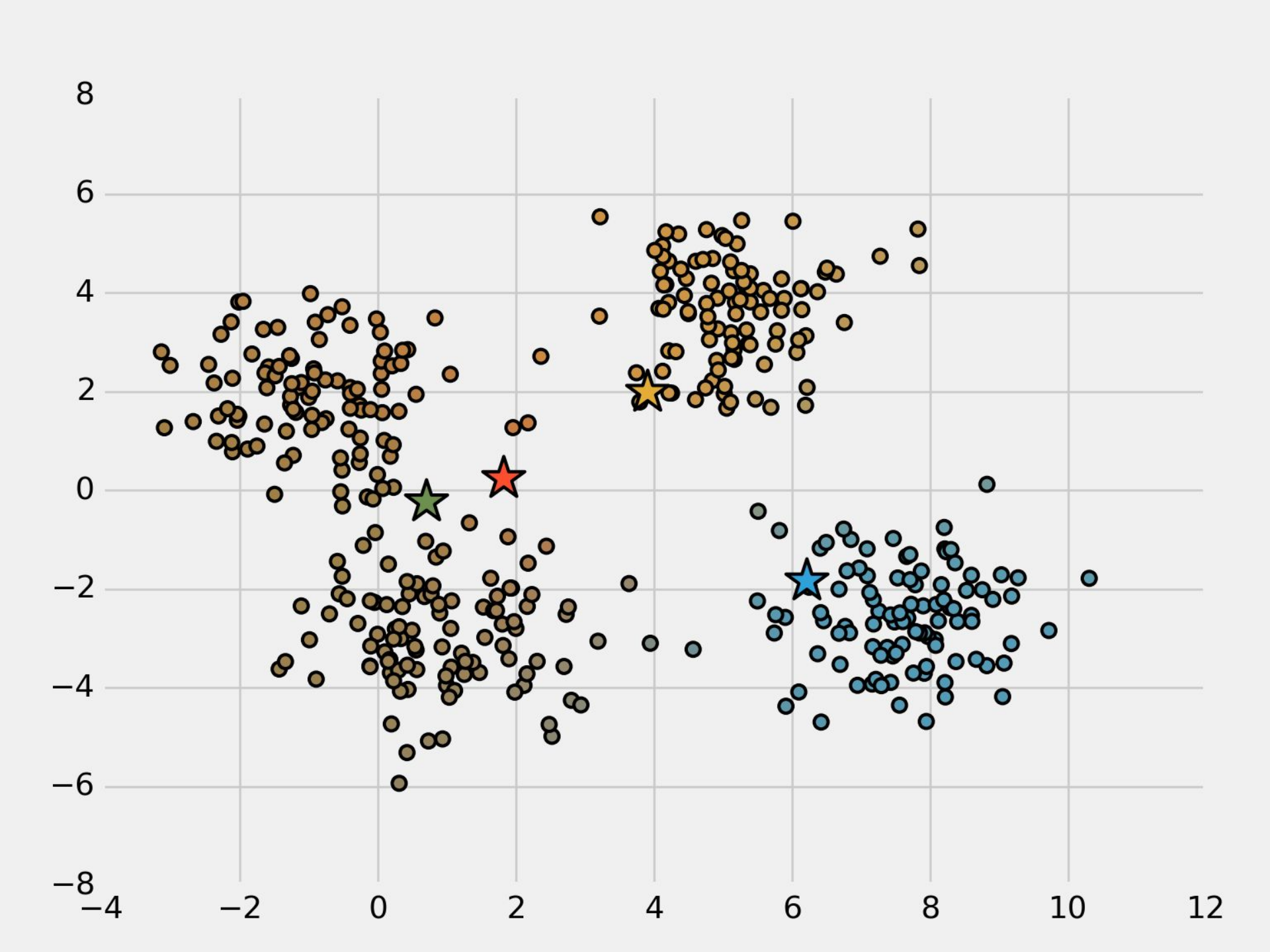

8
6
4
2
0
−2
−4
−6
−8
−4
−2
0
2
4
6
8
10
12

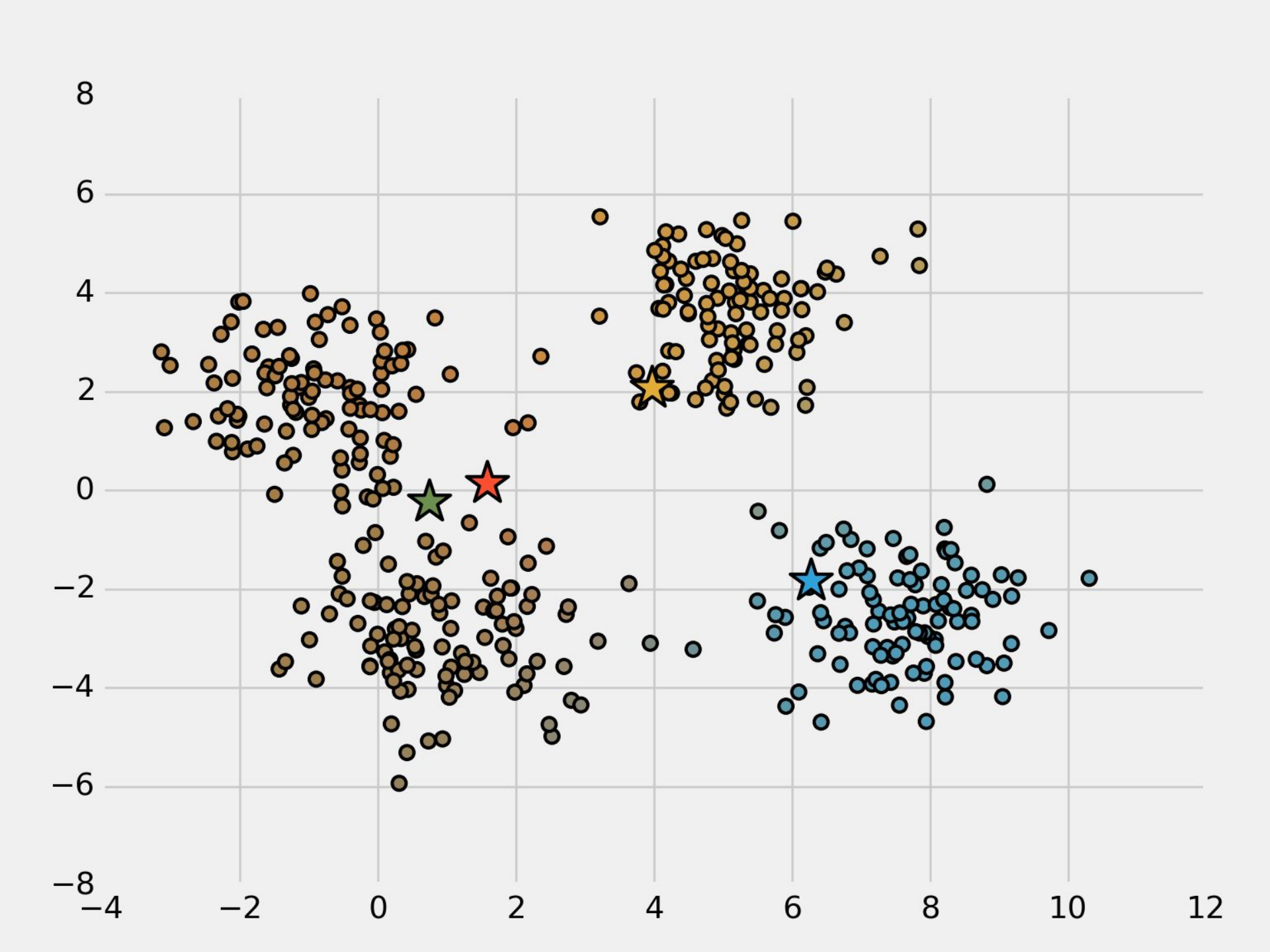

8
6
4
2
0
−2
−4
−6
−8
−4
−2
0
2
4
6
8
10
12

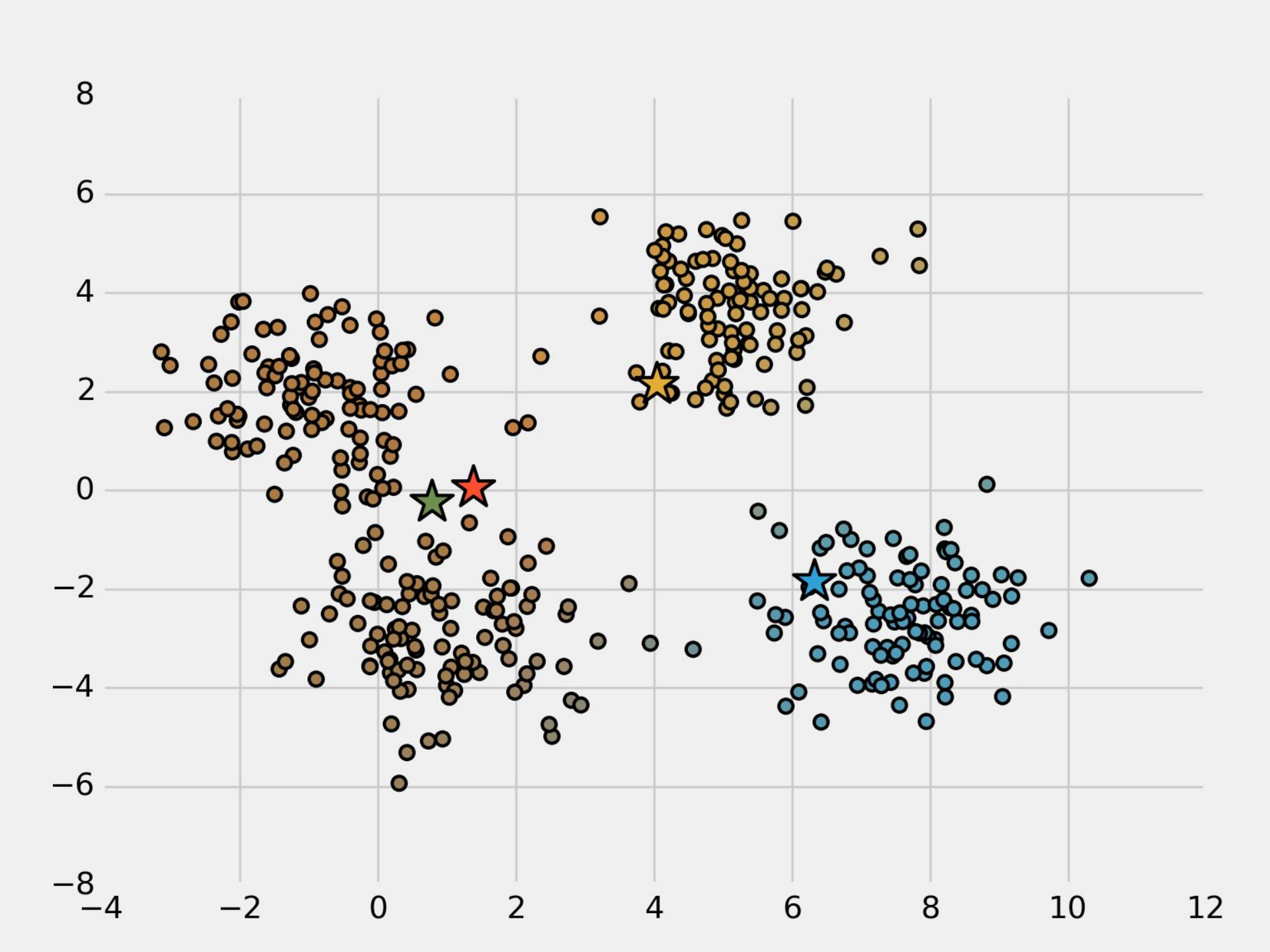

8
6
4
2
0
−2
−4
−6
−8
−4
−2
0
2
4
6
8
10
12

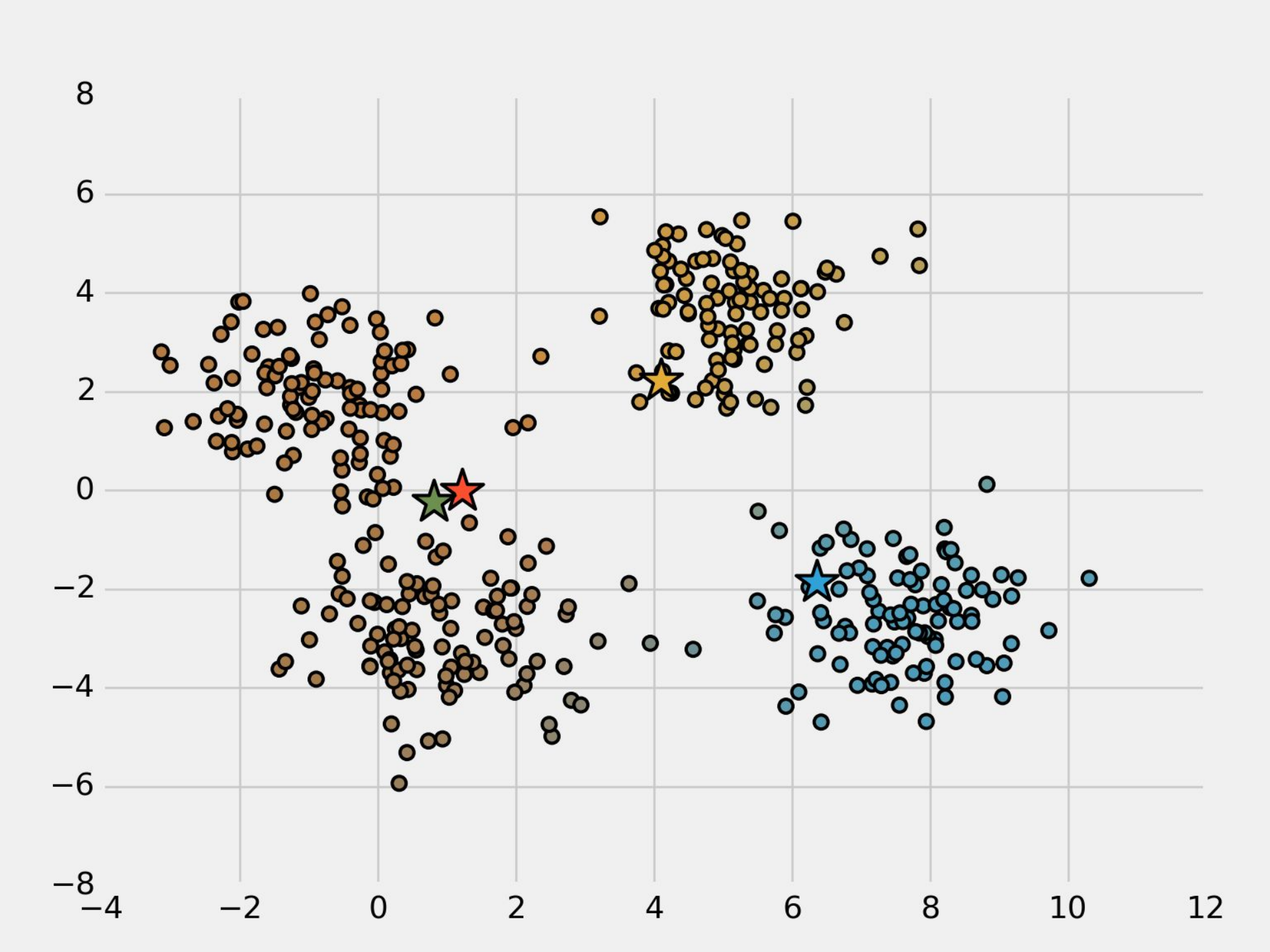

8
6
4
2
0
−2
−4
−6
−8
−4
−2
0
2
4
6
8
10
12

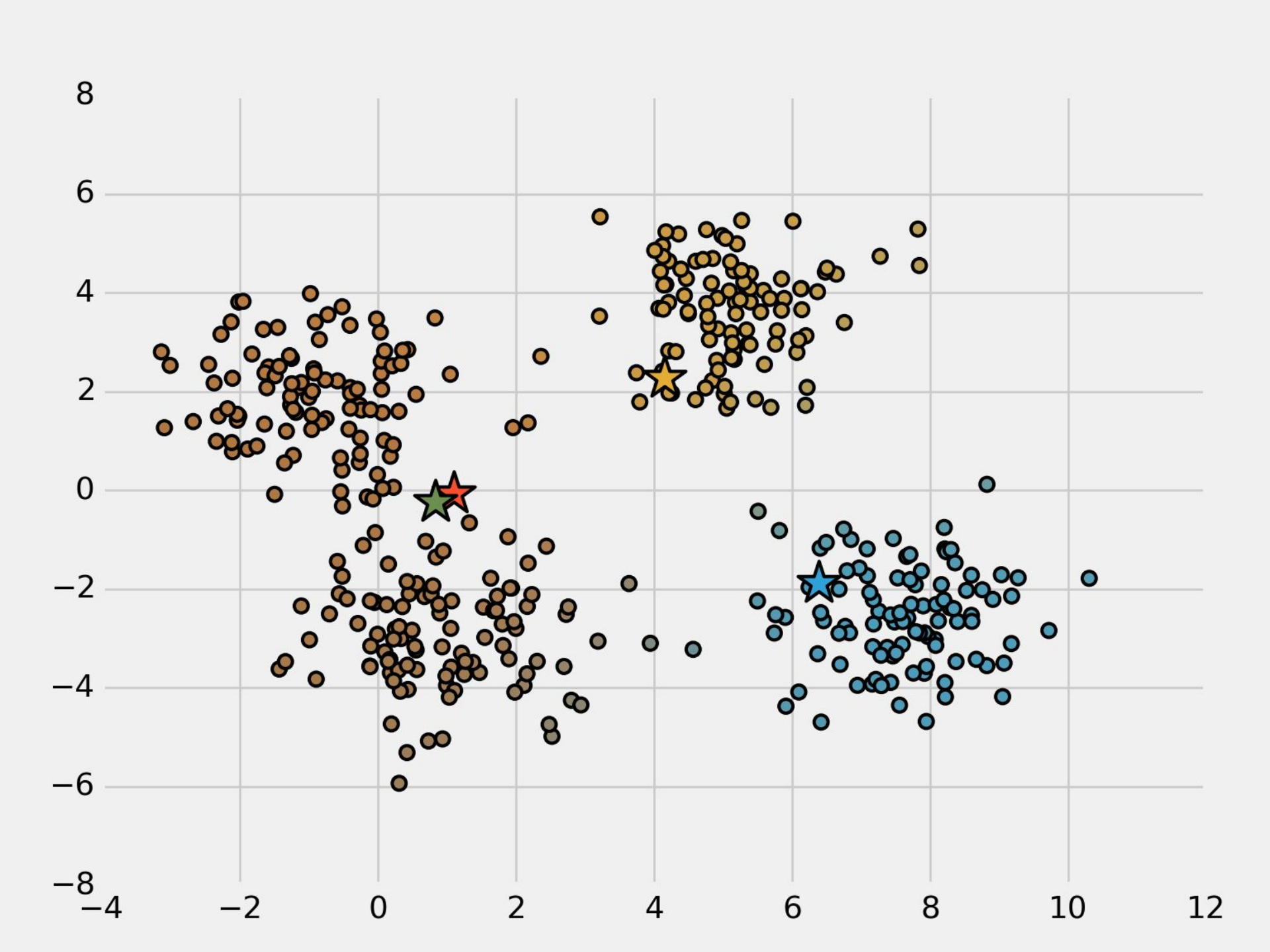

8
6
4
2
0
−2
−4
−6
−8
−4
−2
0
2
4
6
8
10
12

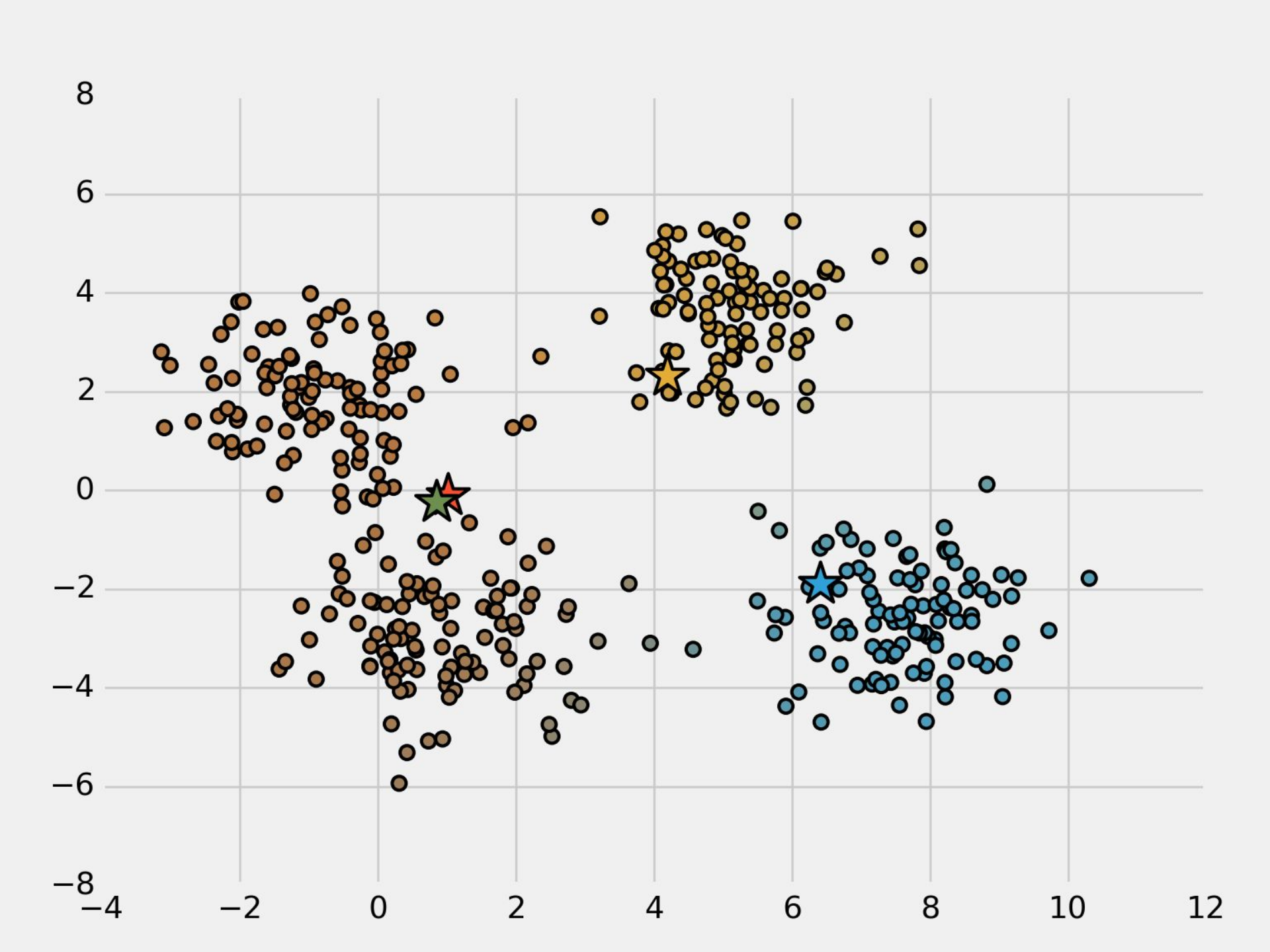

8
6
4
2
0
−2
−4
−6
−8
−4
−2
0
2
4
6
8
10
12

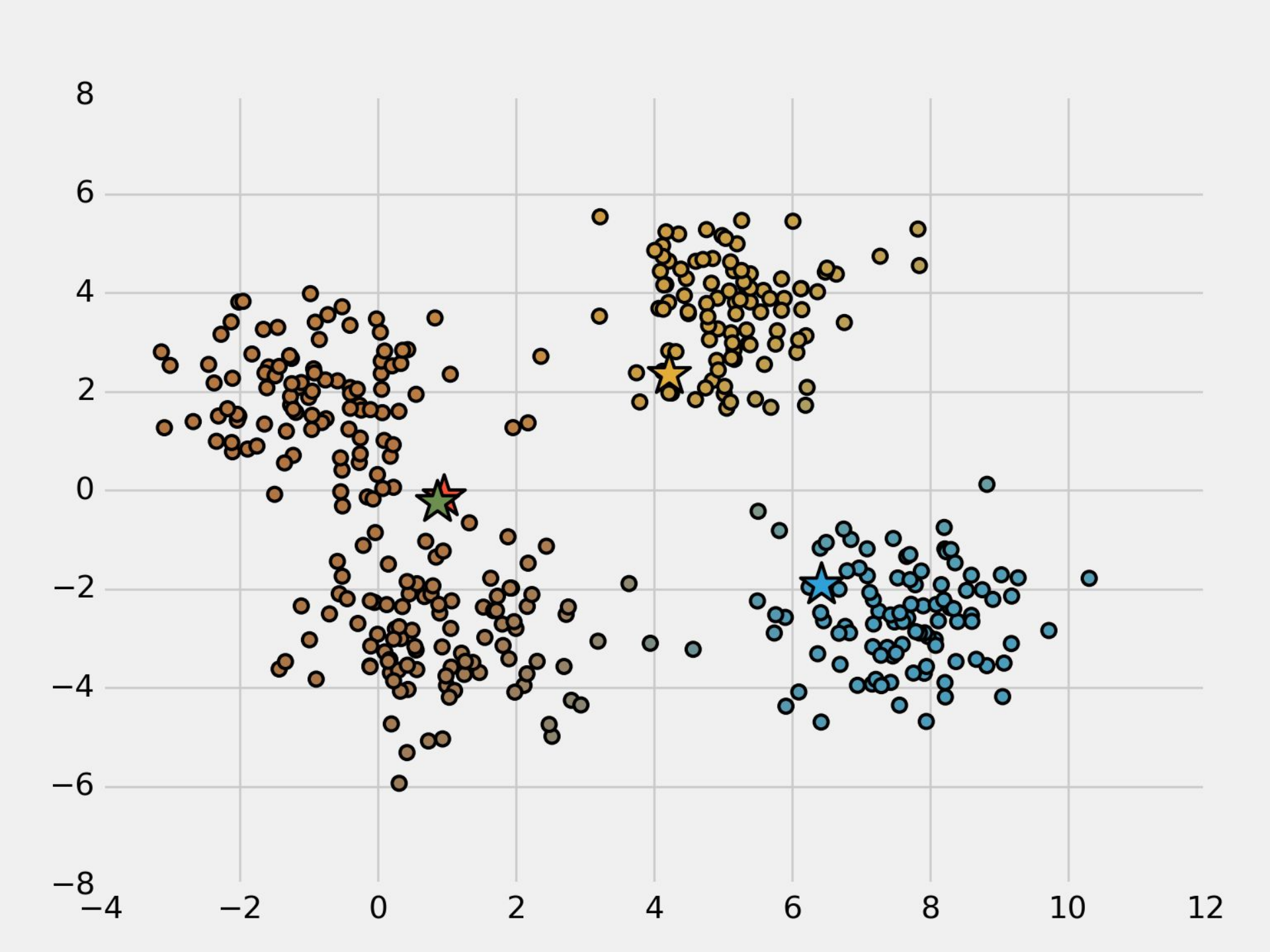

8
6
4
2
0
−2
−4
−6
−8
−4
−2
0
2
4
6
8
10
12

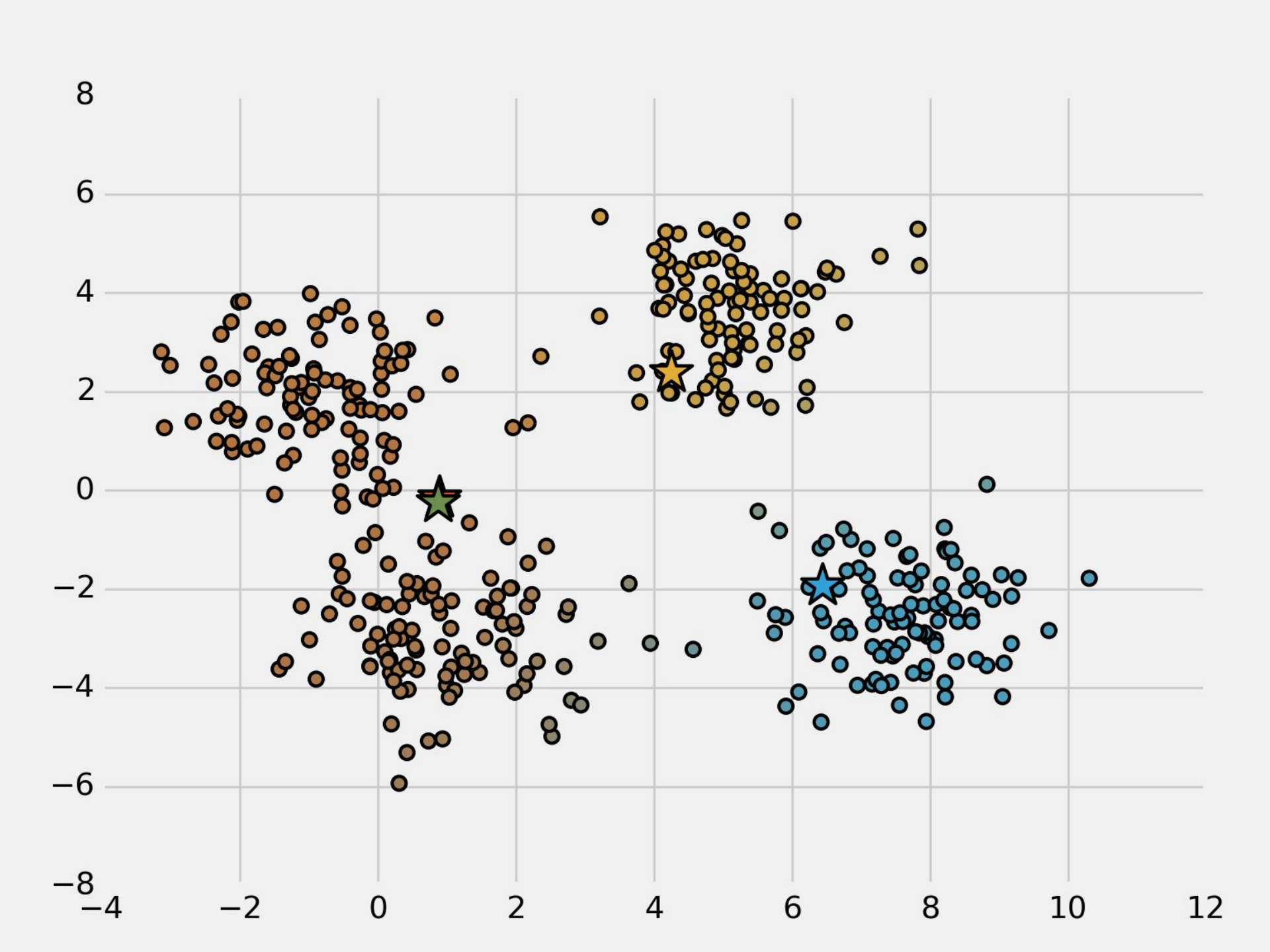

8
6
4
2
0
−2
−4
−6
−8
−4
−2
0
2
4
6
8
10
12

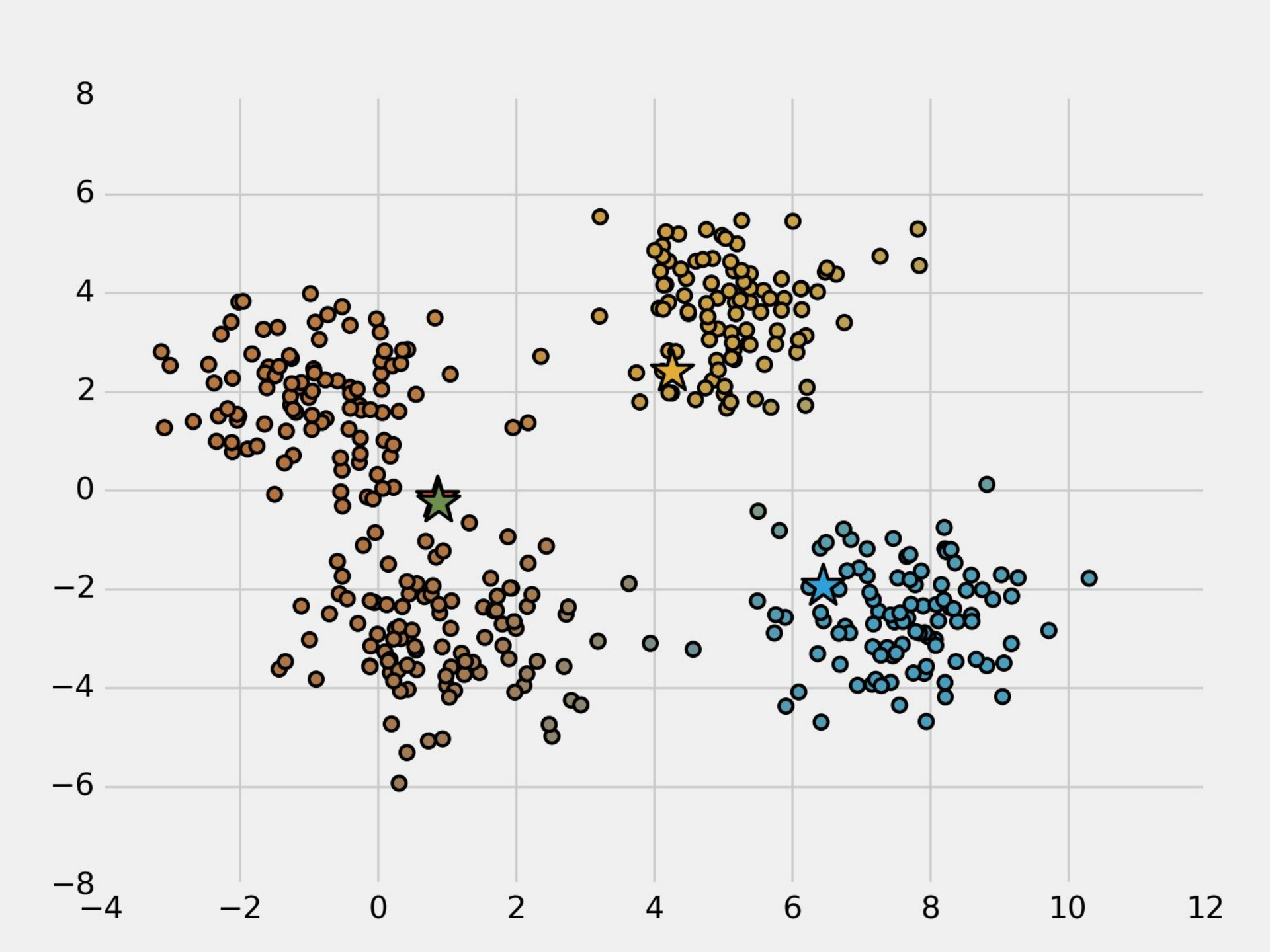

8
6
4
2
0
−2
−4
−6
−8
−4
−2
0
2
4
6
8
10
12

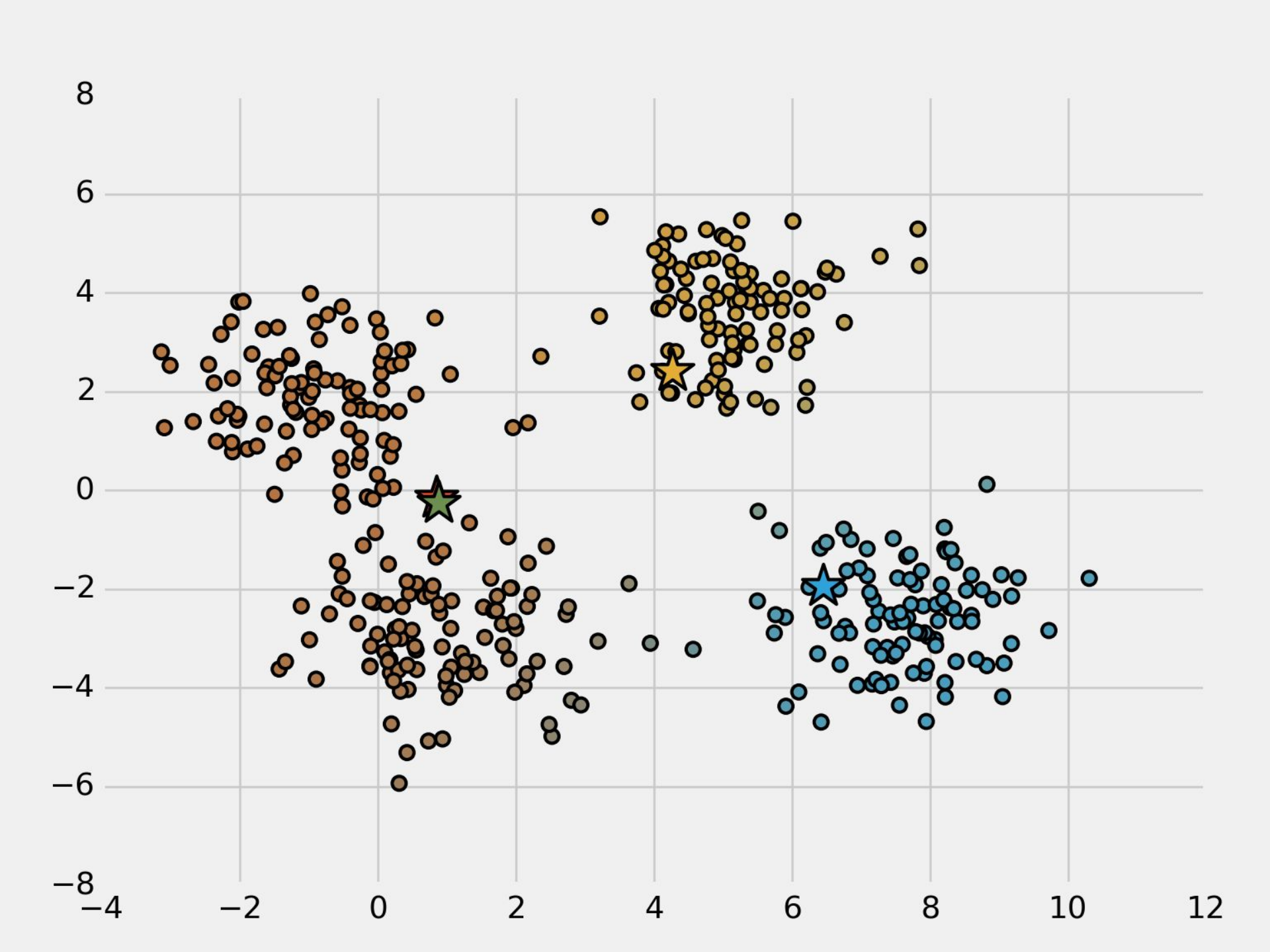

8
6
4
2
0
−2
−4
−6
−8
−4
−2
0
2
4
6
8
10
12

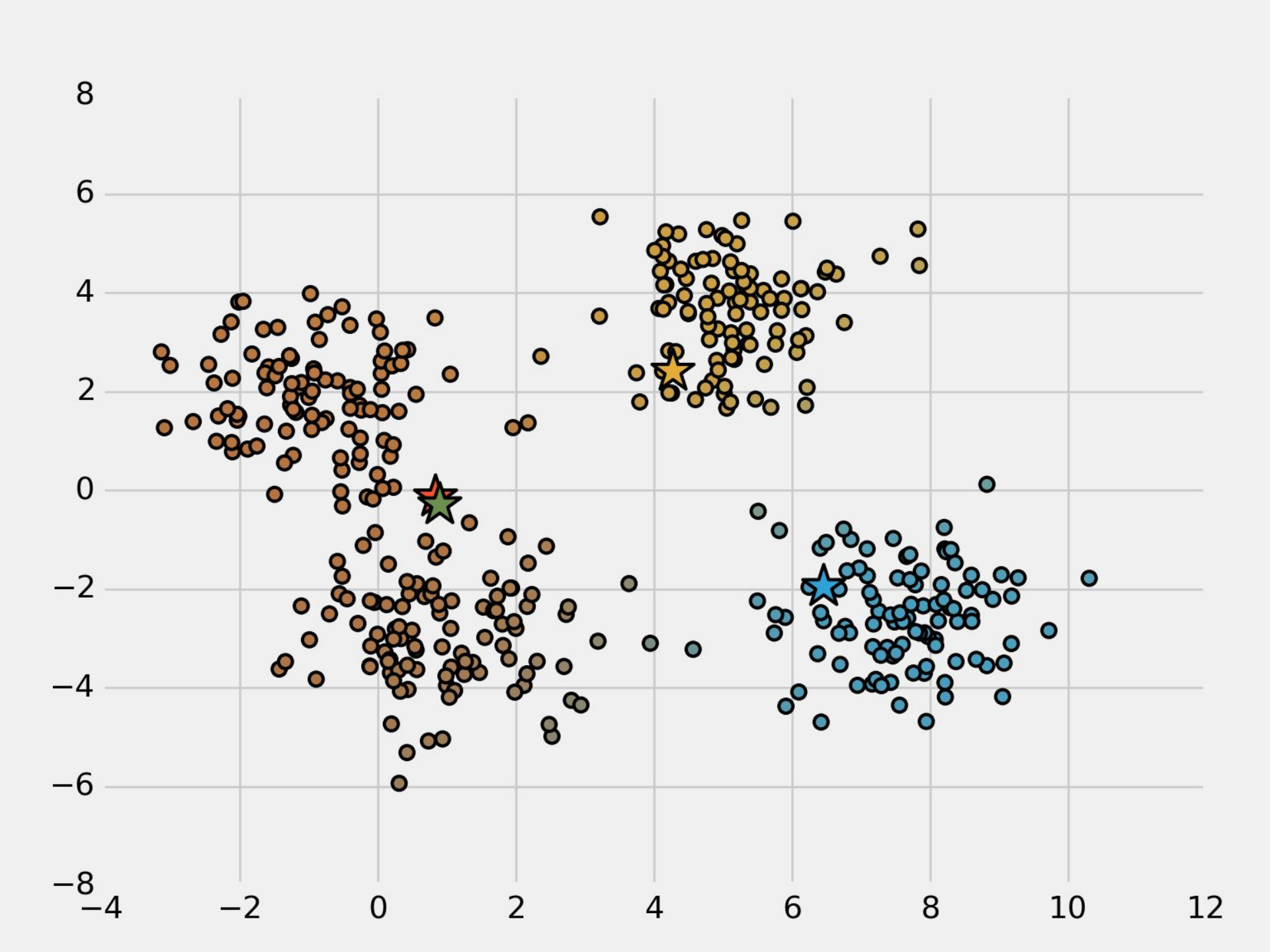

8
6
4
2
0
−2
−4
−6
−8
−4
−2
0
2
4
6
8
10
12

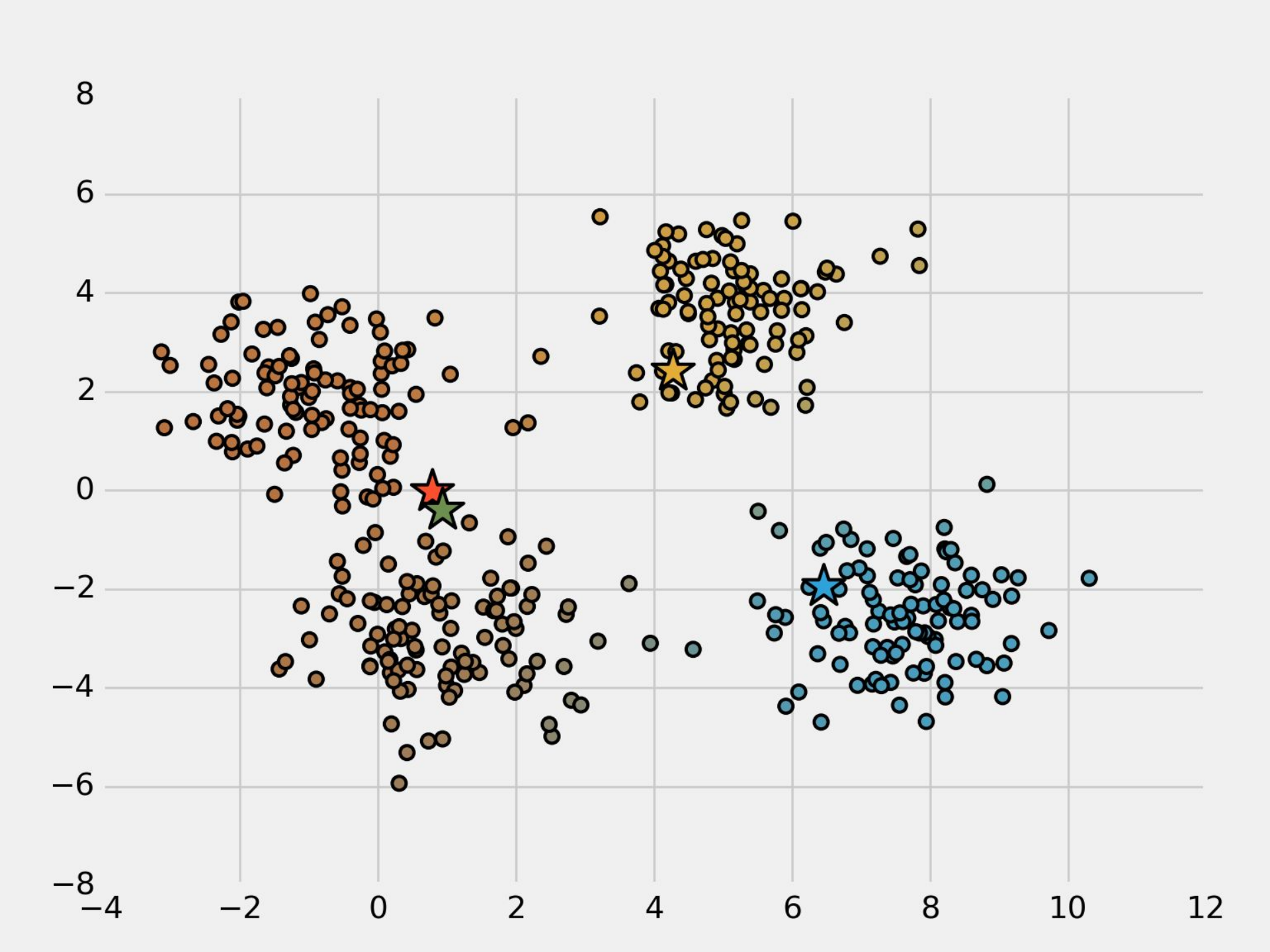

8
6
4
2
0
−2
−4
−6
−8
−4
−2
0
2
4
6
8
10
12

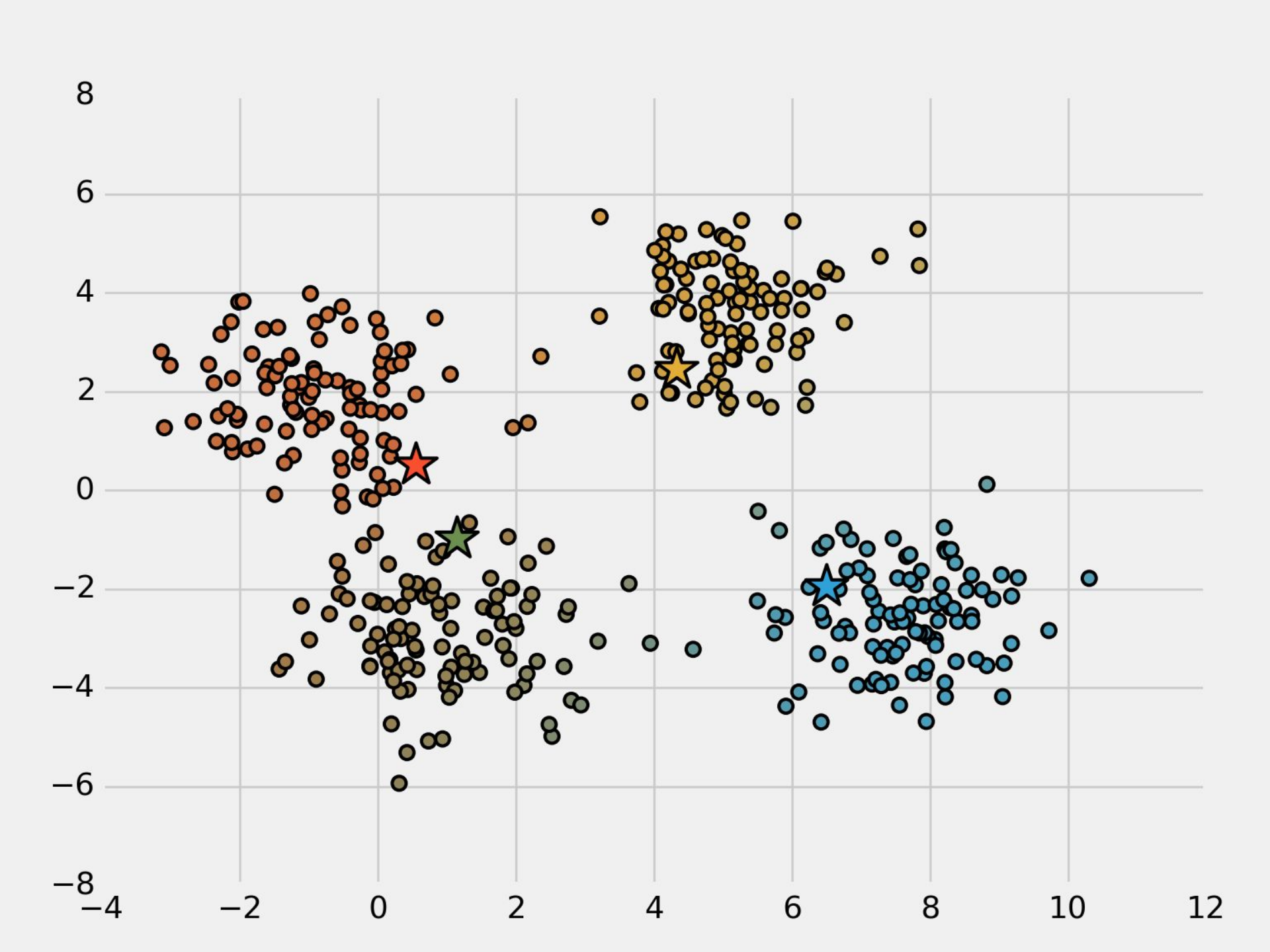

8
6
4
2
0
−2
−4
−6
−8
−4
−2
0
2
4
6
8
10
12

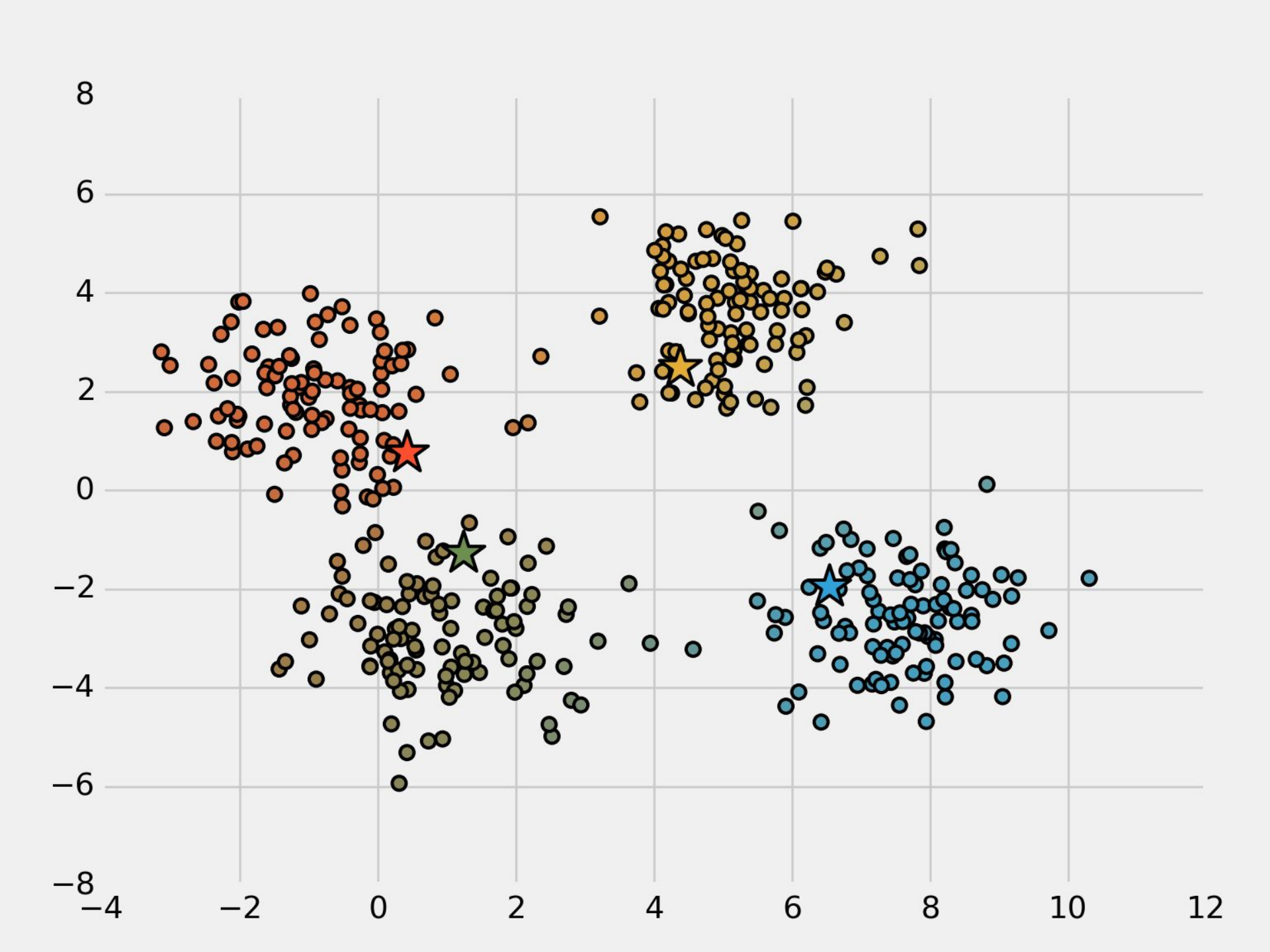

8
6
4
2
0
−2
−4
−6
−8
−4
−2
0
2
4
6
8
10
12

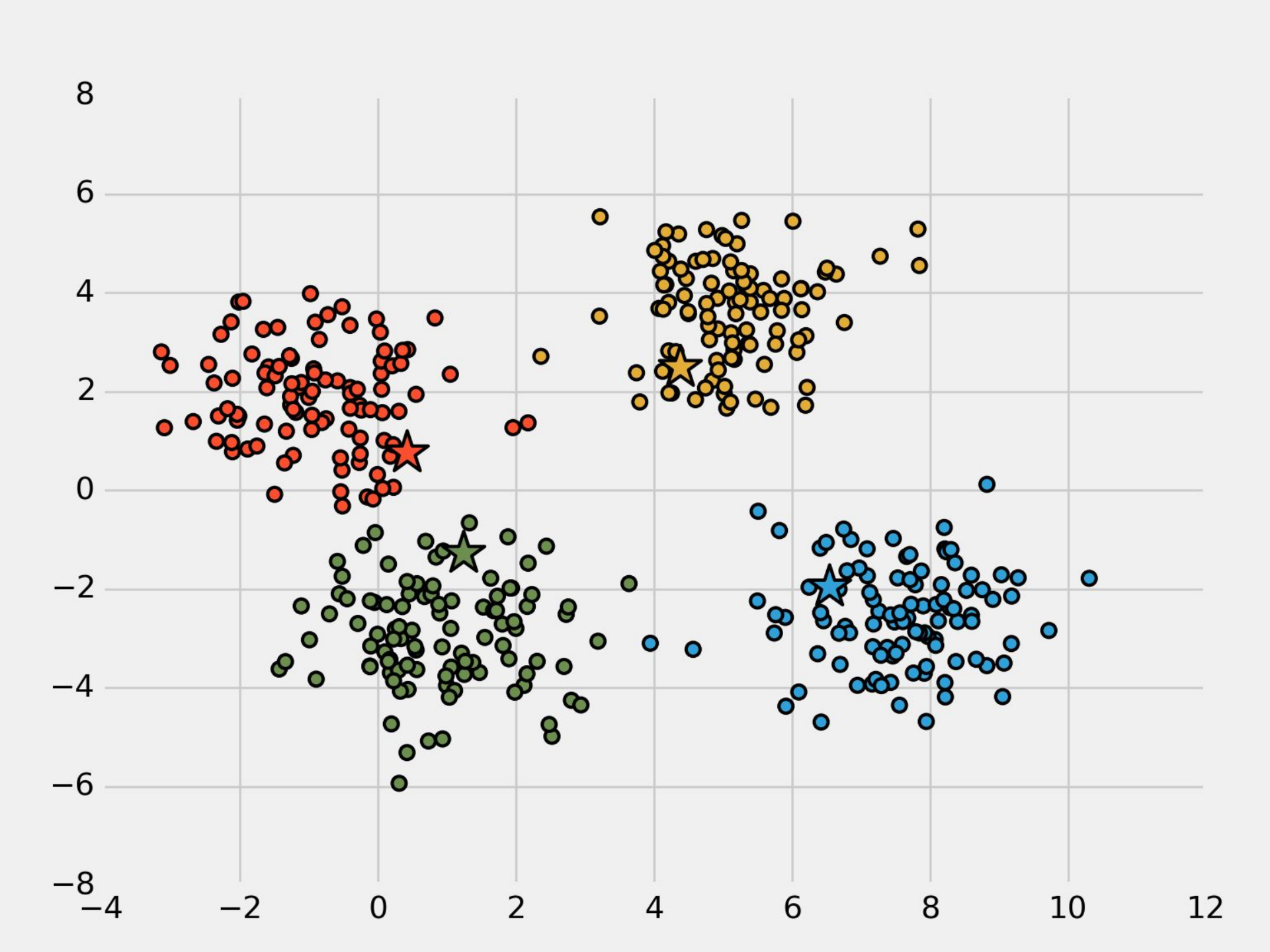

8
6
4
2
0
−2
−4
−6
−8
−4
−2
0
2
4
6
8
10
12


8
6
4
2
0
−2
−4
−6
−8
−4
−2
0
2
4
6
8
10
12


8
6
4
2
0
−2
−4
−6
−8
−4
−2
0
2
4
6
8
10
12


8
6
4
2
0
−2
−4
−6
−8
−4
−2
0
2
4
6
8
10
12


8
6
4
2
0
−2
−4
−6
−8
−4
−2
0
2
4
6
8
10
12


8
6
4
2
0
−2
−4
−6
−8
−4
−2
0
2
4
6
8
10
12