Clustering via Gaussian Mixture Models

Carlos Fernandez-Granda

These slides are based on the book Probability and Statistics for Data Science by Carlos Fernandez-Granda, available for purchase here. A free preprint, videos, code, slides and solutions to exercises are available at https://www.ps4ds.net

Goal

Explain how to use Gaussian mixture models for clustering

Clustering

Separate data into classes

Isn't this classification?

Classification is supervised: we have training labels

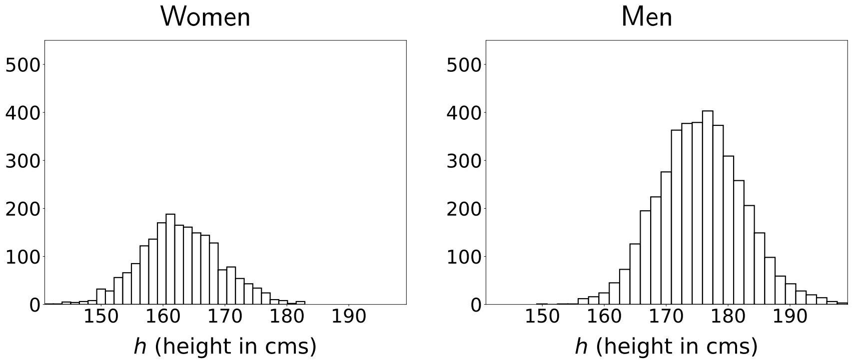
Clustering is unsupervised: we do not have training labels

Example

Clustering people according to height

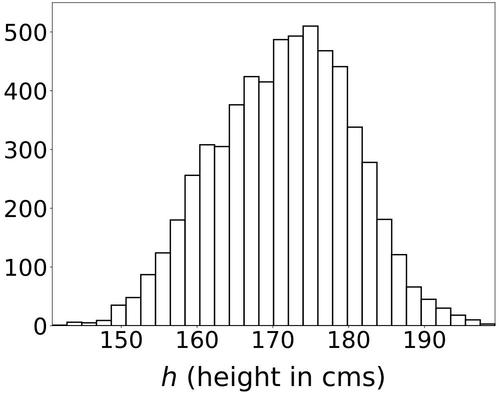
Classification

Supervised learning: We have training labels



Clustering

Unsupervised learning: No training labels



Strategy: Fit mixture model to cluster the data

Parametric mixture model

Assumptions:

(1) Data can be divided into classes

(2) Distribution of features given class y is parametric

Example

Gaussian mixture model for height

Height: Continuous random variable $\tilde{h}$

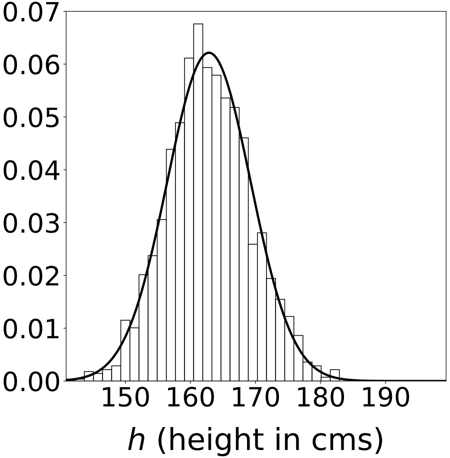
Sex: Discrete random variable $\tilde{s}$

Assumption: Conditional distribution of $\tilde{h}$ given $\tilde{s} = s$ is Gaussian with parameters that depend on s

If we have labels, fitting the model is easy!

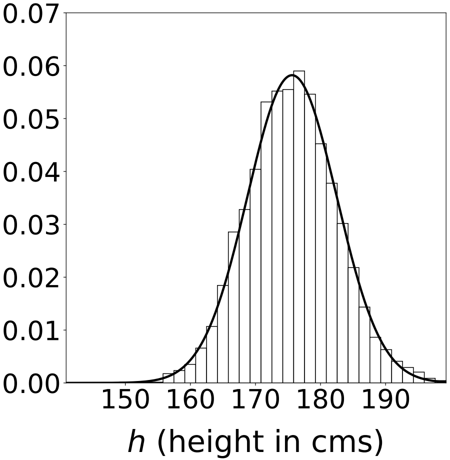
Conditional distribution of $\tilde{h}$ given $\tilde{s}$ = woman

Gaussian with $\mu_{\text{women}} = 163$ cm and $\sigma_{\text{women}} = 6.4$ cm



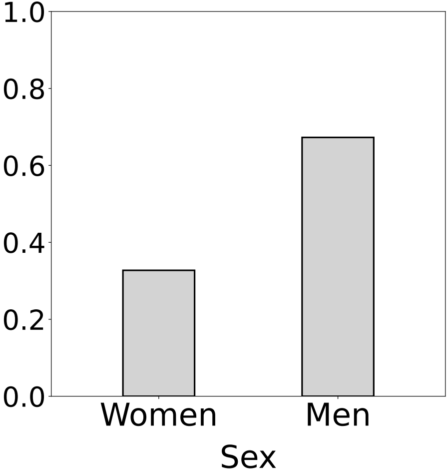
Conditional distribution of $\tilde{h}$ given $\tilde{s}$ = man

Gaussian with $\mu_{\text{men}} = 176$ cm and $\sigma_{\text{men}} = 6.9$ cm



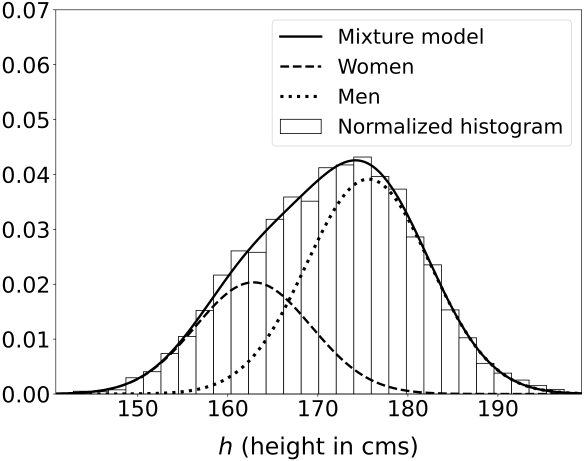
Marginal distribution of $\tilde{s}$

1,986 women and 4,082 men



Gaussian mixture model

$$f_{\tilde{h}}(h) = p_{\tilde{s}}(\text{woman})\, f_{\tilde{h}\,|\,\tilde{s}}(h\,|\,\text{woman}) + p_{\tilde{s}}(\text{man})\, f_{\tilde{h}\,|\,\tilde{s}}(h\,|\,\text{man})$$



Goal: Fit the model without labels

Gaussian mixture model

Data: $x_1, \ldots, x_n$

Modeled as samples from a continuous random vector $\tilde{x}$ dependent on a latent random variable $\tilde{k}$

Conditional distribution of $\tilde{x}$ given $\tilde{k} = k$: Gaussian with parameters μ_k, Σ_k

Challenge: How to fit the model if we don't observe $\tilde{k}$?

Parameters

Number of clusters m is assumed known

- Pmf of $\tilde{k}$: α_k, $1 \leq k \leq m$
- Mean μ_k and covariance matrix Σ_k, $1 \leq k \leq m$

How do we use the model for clustering?

$$\hat{k}_i := \arg\max_k p_{\tilde{k} \mid \tilde{x}}(k \mid x_i)$$
$$= \arg\max_k \gamma_{i,k}$$

$\gamma_{i,k} := p_{\tilde{k} \mid \tilde{x}}(k \mid x_i)$ are the membership probabilities of data point i

Conditional likelihood

Assuming i.i.d. samples

$$\begin{aligned}\mathcal{L}_{i,k}(\mu_k, \Sigma_k) &:= f_{\tilde{x} \mid \tilde{k}}(x_i \mid k) \\ &= \frac{1}{\sqrt{(2\pi)^d |\Sigma_k|}} \exp\left(-\frac{1}{2}(x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k)\right)\end{aligned}$$

Likelihood

Assuming i.i.d. samples

$$
\begin{aligned}
\mathcal{L}(\theta) &:= \prod_{i=1}^{n} f_{\tilde{x}}(x_i) \\
&= \prod_{i=1}^{n} \sum_{k=1}^{m} p_{\tilde{k}}(k) f_{\tilde{x} \mid \tilde{k}}(x_i \mid k) \\
&= \prod_{i=1}^{n} \sum_{k=1}^{m} \alpha_k \mathcal{L}_{i,k}(\mu_k, \Sigma_k)
\end{aligned}
$$

$$
\log \mathcal{L}(\theta) = \sum_{i=1}^{n} \log \sum_{k=1}^{m} \alpha_k \mathcal{L}_{i,k}(\mu_k, \Sigma_k)
$$

No closed-form maximizer!

Expectation maximization

Idea: Jointly estimate parameters and membership probabilities

Initialize model parameters, then repeatedly update

1. Membership probabilities *assuming fixed model parameters*
2. Model parameters *assuming fixed membership probabilities*

Update 1

$$
\begin{aligned}
\gamma_{i,k} &:= p_{\tilde{k} \,|\, \tilde{x}}\,(k \,|\, x_i) \\
&= \frac{p_{\tilde{k}}\,(k)\, f_{\tilde{x}_i \,|\, \tilde{k}}\,(x_i \,|\, k)}{\sum_{l=1}^{m} p_{\tilde{k}}\,(l)\, f_{\tilde{x}_i \,|\, \tilde{k}}\,(x_i \,|\, l)} \\
&= \frac{\alpha_k \mathcal{L}_{i,k}(\mu_k, \Sigma_k)}{\sum_{k=1}^{m} \alpha_l \mathcal{L}_{i,l}(\mu_l, \Sigma_l)}
\end{aligned}
$$

Update 2

α_k is the probability of belonging to cluster k

Effective number of points in cluster k?

$$n_k := \sum_{i=1}^{n} \gamma_{i,k}$$

$$\alpha_k := \frac{n_k}{n}$$

Update 2

Ideally

$$\mu_k := \frac{1}{\text{number of data in cluster } k} \sum_{x_i \text{ in cluster } k} x_i$$

But we don't know cluster assignments...

Idea: Use membership probabilities as *soft* assignments

$$\mu_k := \frac{1}{n_k} \sum_{i=1}^{n} \gamma_{i,k} x_i$$

Update 2

Ideally

$$\Sigma_k := \frac{1}{\text{number of data in cluster } k} \sum_{x_i \text{ in cluster } k} (x_i - \mu_k)(x_i - \mu_k)^T$$

Using membership probabilities as *soft* assignments

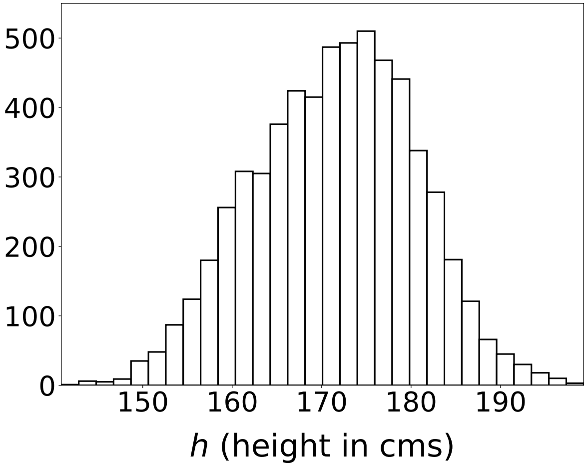
$$\Sigma_k := \frac{1}{n_k} \sum_{i=1}^{n} \gamma_{i,k} (x_i - \mu_k)(x_i - \mu_k)^T$$

Expectation maximization

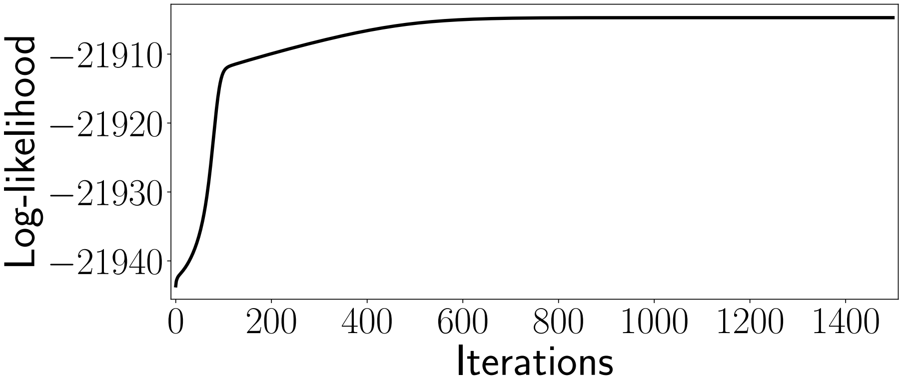
Initialize model parameters, then repeatedly update

1. Membership probabilities *assuming fixed model parameters*
2. Model parameters *assuming fixed membership probabilities*

Clustering height data



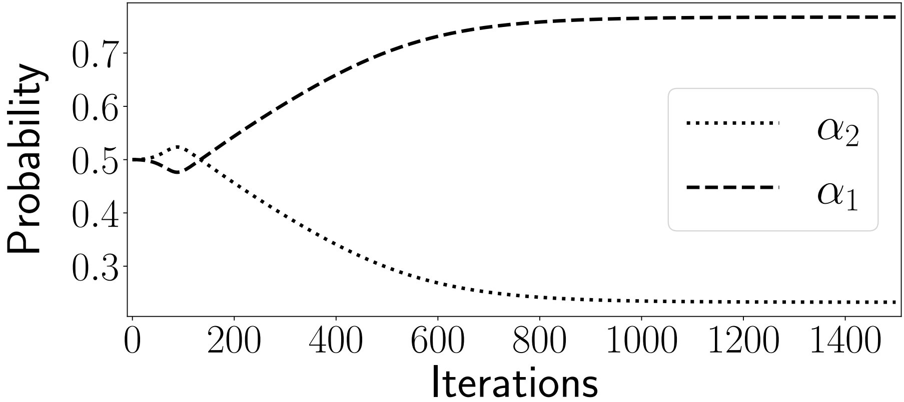
Iterations



Log-likelihood
−21910
−21920
−21930
−21940
0
200
400
600
800
1000
1200
1400
Iterations

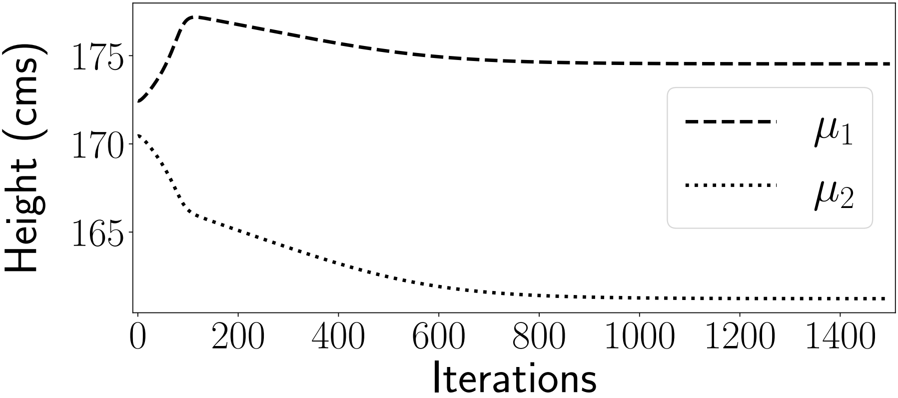
α_1, α_2

Fraction of women/men: 32.7/67.3%



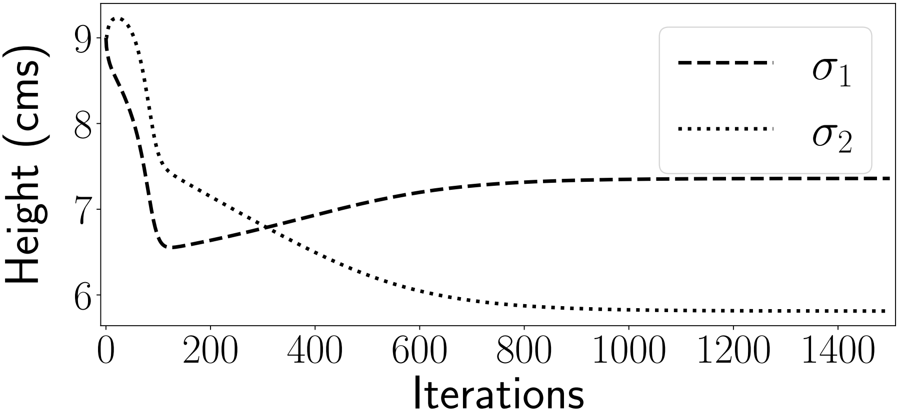
μ_1, μ_2

Mean height women/men: 163/176 cm

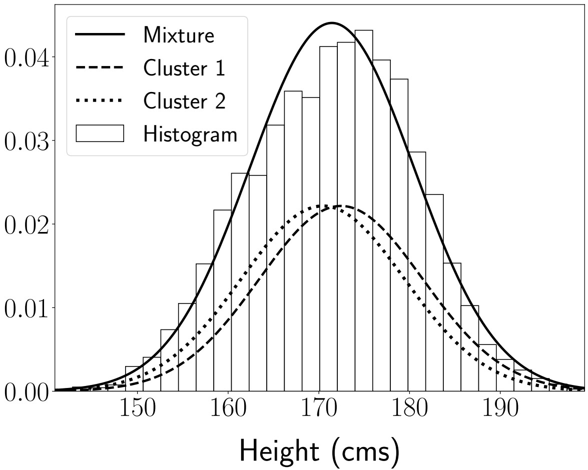


σ_1, σ_2

Standard deviation women/men: 6.4/6.9 cm

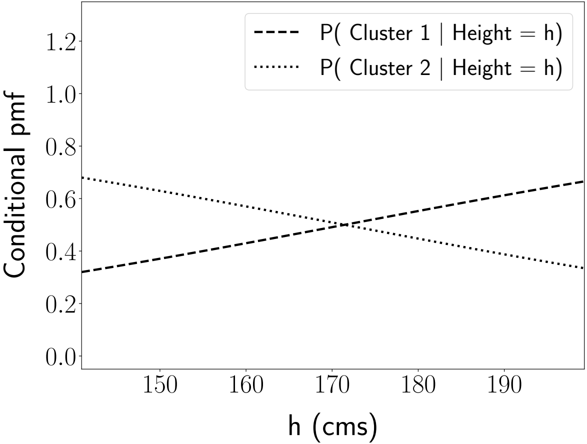


Iteration 1



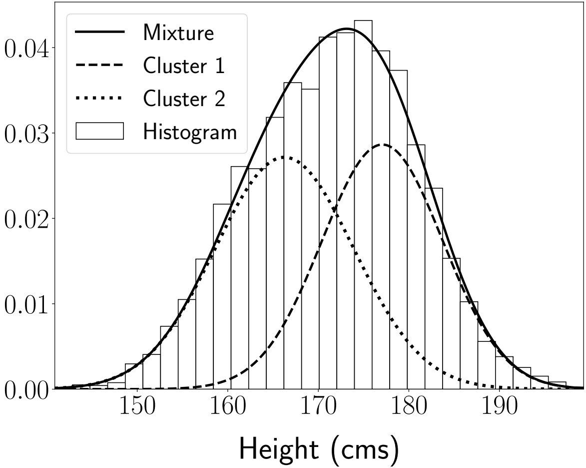
Mixture
Cluster 1
Cluster 2
Histogram
0.00
0.01
0.02
0.03
0.04
150
160
170
180
190
Height (cms)

Iteration 1

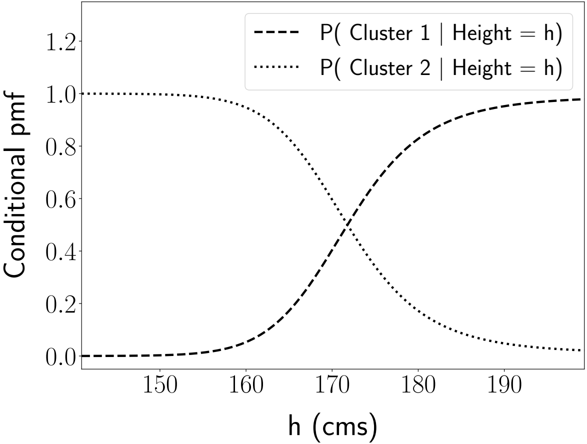


P(Cluster 1 | Height = h)
P(Cluster 2 | Height = h)
Conditional pmf
1.2
1.0
0.8
0.6
0.4
0.2
0.0
150
160
170
180
190
h (cms)

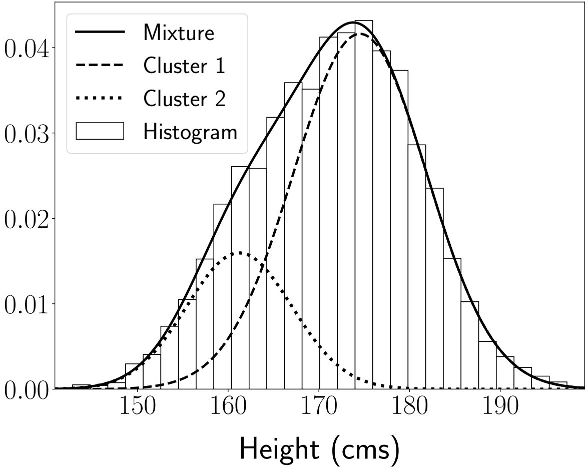
Iteration 100



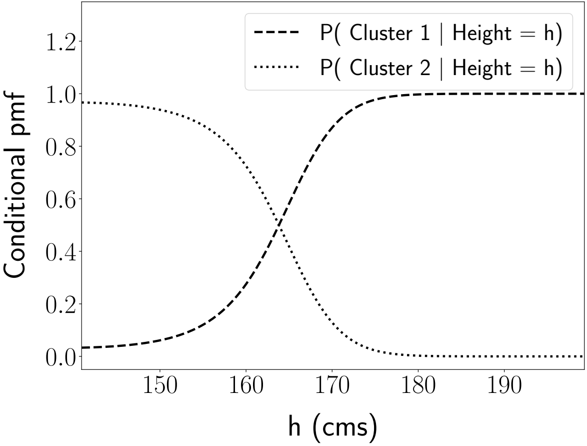
Iteration 100



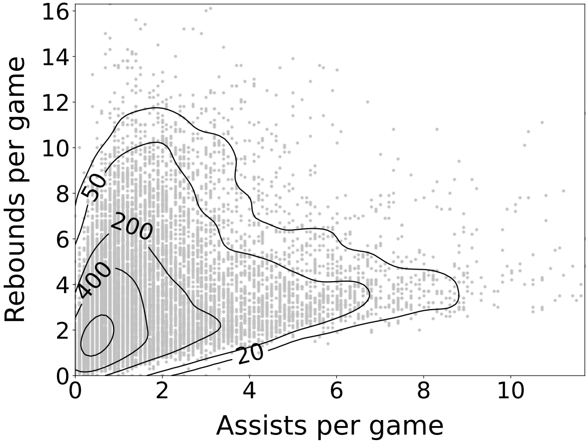
Iteration 1,500



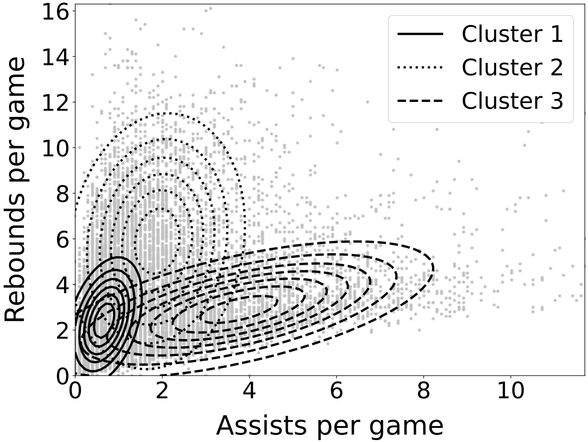
Iteration 1,500



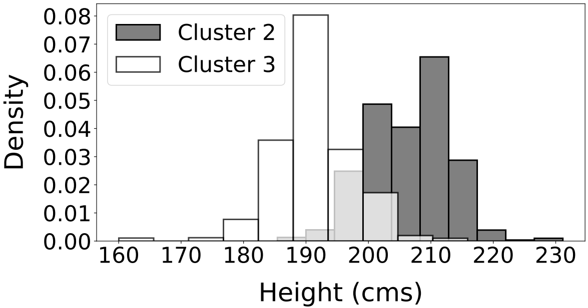
NBA players 1996 - 2019



NBA players 1996 - 2019



NBA players 1996 - 2019



What have we learned?

How to fit Gaussian mixture models using unlabeled data

Application to clustering