Maximum Likelihood Estimation

CS 446

Maximum likelihood: abstract formulation

We've had one main **"meta-algorithm"** this semester:

- (Regularized) ERM principle: pick the model that minimizes an average loss over training data.

Maximum likelihood: abstract formulation

We've had one main **"meta-algorithm"** this semester:

- (Regularized) ERM principle: pick the model that minimizes an average loss over training data.

We've also discussed another:
the "Maximum likelihood estimation (MLE)" principle:

- Pick a set of probability models for your data: $\mathcal{P} := \{p_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$.
  - $p_{\boldsymbol{\theta}}$ will denote *both* densities *and* masses; the literature is similarly inconsistent.
  - Given samples $(\boldsymbol{z}_i)_{i=1}^n$, pick the model that maximized the likelihood

$$\max_{\theta \in \Theta} \mathcal{L}(\theta) = \max_{\theta \in \Theta} \ln \prod_{i=1}^{n} p_\theta(\boldsymbol{z}_i) = \max_{\theta \in \Theta} \sum_{i=1}^{n} \ln p_\theta(\boldsymbol{z}_i),$$

where the $\ln(\cdot)$ is for mathematical convenience,
and $\boldsymbol{z}_i$ can be a labeled pair $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ or just $\boldsymbol{x}_i$.

Connections between ERM and MLE

- We can often derive and justify many basic methods with either (e.g., least squares, logistic regression, k-means, . . .).
  - MLE ideas were used to verive VAEs, which we'll cover next week!

Connections between ERM and MLE

- We can often derive and justify many basic methods with either (e.g., least squares, logistic regression, k-means, . . .).
  - MLE ideas were used to verive VAEs, which we'll cover next week!
- Each perspective suggests some different details and interpretation.

Connections between ERM and MLE

- We can often derive and justify many basic methods with either (e.g., least squares, logistic regression, k-means, . . .).
  - MLE ideas were used to verive VAEs, which we'll cover next week!
- Each perspective suggests some different details and interpretation.
- Both approaches rely upon seemingly arbitrary assumptions and choices.

Connections between ERM and MLE

- We can often derive and justify many basic methods with either (e.g., least squares, logistic regression, k-means, . . .).
  - MLE ideas were used to verive VAEs, which we'll cover next week!
- Each perspective suggests some different details and interpretation.
- Both approaches rely upon seemingly arbitrary assumptions and choices.
- The success of MLE seems to often hinge upon an astute choice of model.
  - Applied scientists often like MLE and its ilk due to interpretability and "usability": they can easily encode domain knowledge. We'll return to this.

Example 1: coin flips.

- We flip a coin of bias $\theta \in [0, 1]$.
- Write down $x_i = 0$ for tails, $x_i = 1$ for heads; then

$$p_\theta(x_i) = x_i\theta + (1 - x_i)(1 - \theta),$$

or alternatively

$$p_\theta(x_i) = \theta^{x_i}(1 - \theta)^{1-x_i}.$$

The second form will be more convenient.

Example 1: coin flips.

- We flip a coin of bias $\theta \in [0, 1]$.
- Write down $x_i = 0$ for tails, $x_i = 1$ for heads; then

$$p_\theta(x_i) = x_i\theta + (1 - x_i)(1 - \theta),$$

or alternatively

$$p_\theta(x_i) = \theta^{x_i}(1 - \theta)^{1-x_i}.$$

The second form will be more convenient.

- Writing $H := \sum_i x_i$ and $T := \sum_i (1 - x_i) = n - H$ for convenience,

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} \big(x_i \ln\theta + (1 - x_i)\ln(1 - \theta)\big) = H \ln\theta + T \ln(1 - \theta).$$

Example 1: coin flips.

- We flip a coin of bias $\theta \in [0, 1]$.
- Write down $x_i = 0$ for tails, $x_i = 1$ for heads;
  then
  $$p_\theta(x_i) = x_i\theta + (1 - x_i)(1 - \theta),$$
  or alternatively
  $$p_\theta(x_i) = \theta^{x_i}(1 - \theta)^{1-x_i}.$$
  The second form will be more convenient.
- Writing $H := \sum_i x_i$ and $T := \sum_i (1 - x_i) = n - H$ for convenience,
  $$\mathcal{L}(\theta) = \sum_{i=1}^{n} \big(x_i \ln\theta + (1 - x_i)\ln(1 - \theta)\big) = H \ln\theta + T \ln(1 - \theta).$$
  Differentiating and setting to 0,
  $$0 = \frac{H}{\theta} - \frac{T}{1 - \theta},$$
  which gives $\theta = \frac{H}{T+H} = \frac{H}{N}$.
- In this way, we've justified a natural algorithm.

Example 2: mean of a Gaussian

- Suppose $x_i \sim \mathcal{N}(\mu, \sigma^2)$, so $\boldsymbol{\theta} = (\mu, \sigma^2)$, and

$$\ln p_{\boldsymbol{\theta}}(x_i) = \ln \frac{\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)}{\sqrt{2\pi\sigma^2}} = -\frac{(x_i-\mu)^2}{2\sigma^2} - \frac{\ln(2\pi\sigma^2)}{2}.$$

Example 2: mean of a Gaussian

- Suppose $x_i \sim \mathcal{N}(\mu, \sigma^2)$, so $\boldsymbol{\theta} = (\mu, \sigma^2)$, and

$$\ln p_{\boldsymbol{\theta}}(x_i) = \ln \frac{\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)}{\sqrt{2\pi\sigma^2}} = -\frac{(x_i-\mu)^2}{2\sigma^2} - \frac{\ln(2\pi\sigma^2)}{2}.$$

- Therefore

$$\mathcal{L}(\boldsymbol{\theta}) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 + \text{stuff without } \mu;$$

  applying ∇_μ and setting to zero gives $\mu = \frac{1}{n}\sum_i x_i$.

- A similar derivation gives $\sigma^2 = \frac{1}{n}\sum_i (x_i-\mu)^2$.

Discussion: Bayesian vs. frequentist perspectives

Question: $\sum_{i=1}^{n} \frac{x_i}{n}$ estimates a Gaussian μ parameter; but isn't it useful more generally?

Discussion: Bayesian vs. frequentist perspectives

Question: $\sum_{i=1}^{n} \frac{x_i}{n}$ estimates a Gaussian μ parameter; but isn't it useful more generally?

Bayesian perspective: we choose a model and believe it well-approximates reality; learning its parameters determines underlying phenomena.

- Bayesian methods can handle model misspecification; LDA is an example which works well despite seemingly impractical assumptions.

Discussion: Bayesian vs. frequentist perspectives

Question: $\sum_{i=1}^{n} \frac{x_i}{n}$ estimates a Gaussian μ parameter; but isn't it useful more generally?

Bayesian perspective: we choose a model and believe it well-approximates reality; learning its parameters determines underlying phenomena.

- Bayesian methods can handle model misspecification; LDA is an example which works well despite seemingly impractical assumptions.

Frequentist perspective: we ask certain questions, and reason about the accuracy of our answers.

- For *many* distributions, $\sum_{i=1}^{n} \frac{x_i}{n}$ is a valid estimate of the mean, moreover with confidence intervals of size $1/\sqrt{n}$.
  This approach isn't free of assumptions: IID is there...

Discussion: Bayesian vs. frequentist perspectives (part 2)

- Discussion also appears in the form "generative vs discriminative ML".
- As before: both philosophies can justify/derive the *same* algorithm; they differ on some details (e.g., choosing k in k-means).
- **IMO:** it's nice having more tools
  (as mentioned before: VAE derived from MLE perspective).

Example 3: Least squares (recap)

If we assume $Y|\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{w}^\mathsf{T}\boldsymbol{X}, \sigma^2)$, then

$$\begin{aligned}\mathcal{L}(\boldsymbol{w}) &= \sum_{i=1}^{n} \ln p_{\boldsymbol{w}}(\boldsymbol{x}_i, y_i) \\ &= \sum_{i=1}^{n} \big(\ln p_{\boldsymbol{w}}(y_i|\boldsymbol{x}_i) + \ln p(\boldsymbol{x}_i)\big) \\ &= \sum_{i=1}^{n} \left(-\frac{(\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_i - y_i)^2}{2\sigma^2} + \text{terms without } \boldsymbol{w}\right).\end{aligned}$$

Therefore

$$\arg\max_{\boldsymbol{w}\in\mathbb{R}^d} \mathcal{L}(\boldsymbol{w}) = \arg\min_{\boldsymbol{w}\in\mathbb{R}^d} \sum_{i=1}^{n} (\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_i - y_i)^2.$$

Example 3: Least squares (recap)

If we assume $Y|\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{w}^\mathsf{T}\boldsymbol{X}, \sigma^2)$, then

$$\begin{aligned}
\mathcal{L}(\boldsymbol{w}) &= \sum_{i=1}^{n} \ln p_{\boldsymbol{w}}(\boldsymbol{x}_i, y_i) \\
&= \sum_{i=1}^{n} \left(\ln p_{\boldsymbol{w}}(y_i|\boldsymbol{x}_i) + \ln p(\boldsymbol{x}_i)\right) \\
&= \sum_{i=1}^{n} \left(-\frac{(\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_i - y_i)^2}{2\sigma^2} + \text{terms without } \boldsymbol{w} \right).
\end{aligned}$$

Therefore

$$\arg\max_{\boldsymbol{w}\in\mathbb{R}^d} \mathcal{L}(\boldsymbol{w}) = \arg\min_{\boldsymbol{w}\in\mathbb{R}^d} \sum_{i=1}^{n} (\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_i - y_i)^2.$$

We can derive/justify the algorithm either way, but some refinements now differ with each perspective (e.g., regularization).

Example 4: Naive Bayes

- Let's try a simple prediction setup, with (Bayes) optimal classifier

$$\arg\max_{y \in \mathcal{Y}} p(Y = y | \boldsymbol{X} = \boldsymbol{x}).$$

(We haven't discussed this concept a lot, but it's widespread in ML.)

Example 4: Naive Bayes

- Let's try a simple prediction setup, with (Bayes) optimal classifier

$$\arg\max_{y \in \mathcal{Y}} p(Y = y | \boldsymbol{X} = \boldsymbol{x}).$$

  (We haven't discussed this concept a lot, but it's widespread in ML.)

- One way to proceed is to learn $p(Y|\boldsymbol{X})$ exactly; that's a pain.

Example 4: Naive Bayes

- Let's try a simple prediction setup, with (Bayes) optimal classifier

$$\arg\max_{y \in \mathcal{Y}} p(Y = y | \boldsymbol{X} = \boldsymbol{x}).$$

(We haven't discussed this concept a lot, but it's widespread in ML.)

- One way to proceed is to learn $p(Y|\boldsymbol{X})$ exactly; that's a pain.
- Let's assume coordinates of $\boldsymbol{X} = (X_1, \ldots, X_d)$ are independent given Y:

$$p(Y = y | \boldsymbol{X} = \boldsymbol{x}) = \frac{p(Y = y, \boldsymbol{X} = \boldsymbol{x})}{p(\boldsymbol{X} = \boldsymbol{x})} = \frac{p(\boldsymbol{X} = \boldsymbol{x} | Y = y) p(Y = y)}{p(\boldsymbol{X} = \boldsymbol{x})}$$
$$= \frac{p(Y = y) \prod_{j=1}^{d} p(X_j = x_j | Y = y)}{p(\boldsymbol{X} = \boldsymbol{x})},$$

and

$$\arg\max_{y \in \mathcal{Y}} p(Y = y | \boldsymbol{X} = \boldsymbol{x}) = \arg\max_{y \in \mathcal{Y}} p(Y = y) \prod_{j=1}^{d} p(\boldsymbol{X} = \boldsymbol{x}) | Y = y).$$

Example 4: Naive Bayes (part 2)

$$\arg\max_{y \in \mathcal{Y}} p(Y = y | \boldsymbol{X} = \boldsymbol{x}) = \arg\max_{y \in \mathcal{Y}} p(Y = y) \prod_{j=1}^{d} p(\boldsymbol{X} = \boldsymbol{x}) | Y = y).$$

Example 4: Naive Bayes (part 2)

$$\arg\max_{y \in \mathcal{Y}} p(Y = y | \boldsymbol{X} = \boldsymbol{x}) = \arg\max_{y \in \mathcal{Y}} p(Y = y) \prod_{j=1}^{d} p(\boldsymbol{X} = \boldsymbol{x}) | Y = y).$$

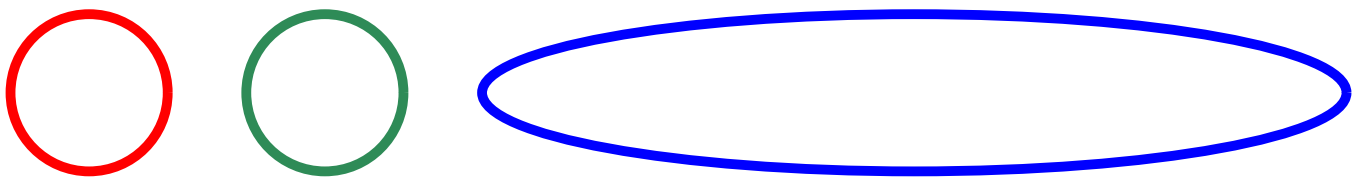
Examples where this helps:

- Suppose $\boldsymbol{X} \in \{0, 1\}^d$ has an arbitrary distribution;\ it's specified with $2^d - 1$ numbers.\ The factored form above needs d numbers. To see how this can help, suppose $\boldsymbol{x} \in \{0, 1\}^d$; instead of having to learn a probability model of 2^d possibilities, we now have to learn $d + 1$ models each with 2 possibilities (binary labels).
- HW5 will use the standard "Iris dataset".\ This data is continuous, Naive Bayes would approximate univariate distributions.

Mixtures of Gaussians.

k-means has spherical clusters?

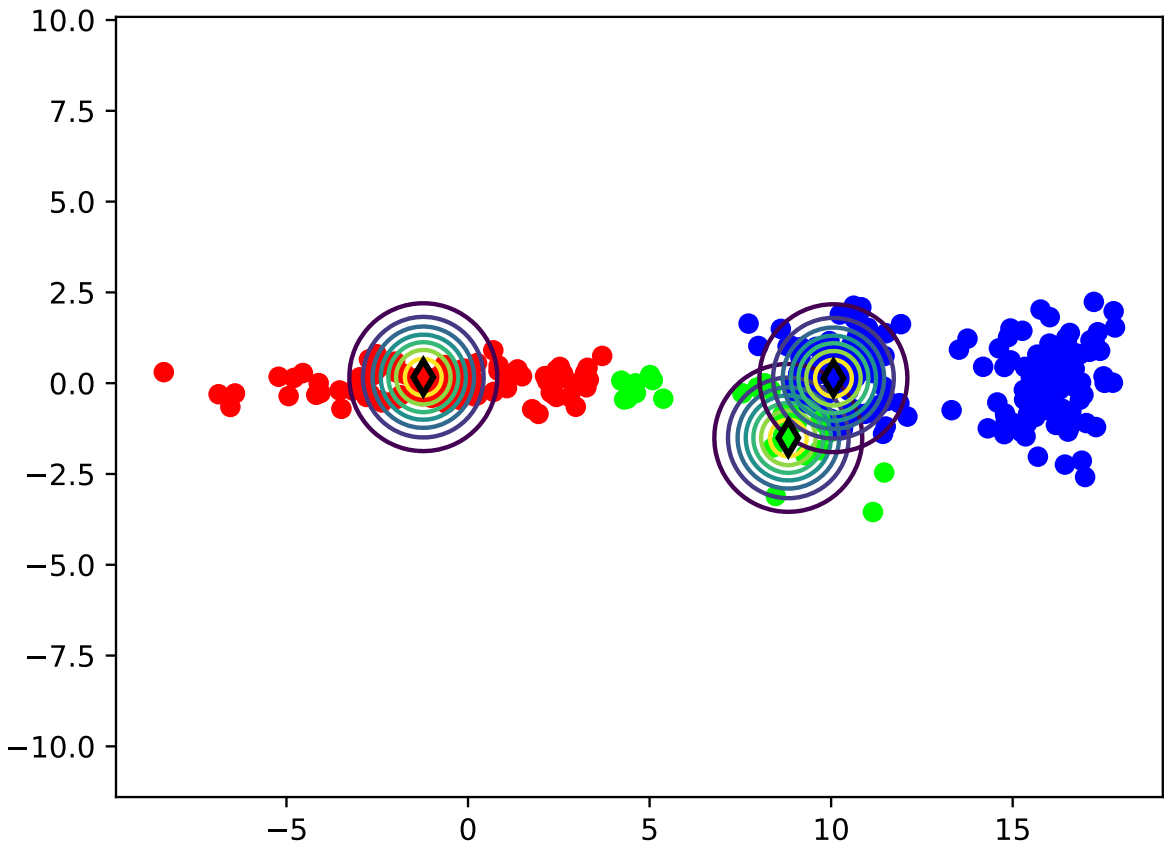
Recall that k-means baked in spherical clusters.



How about we model each cluster with a Gaussian?

k-means has spherical clusters?

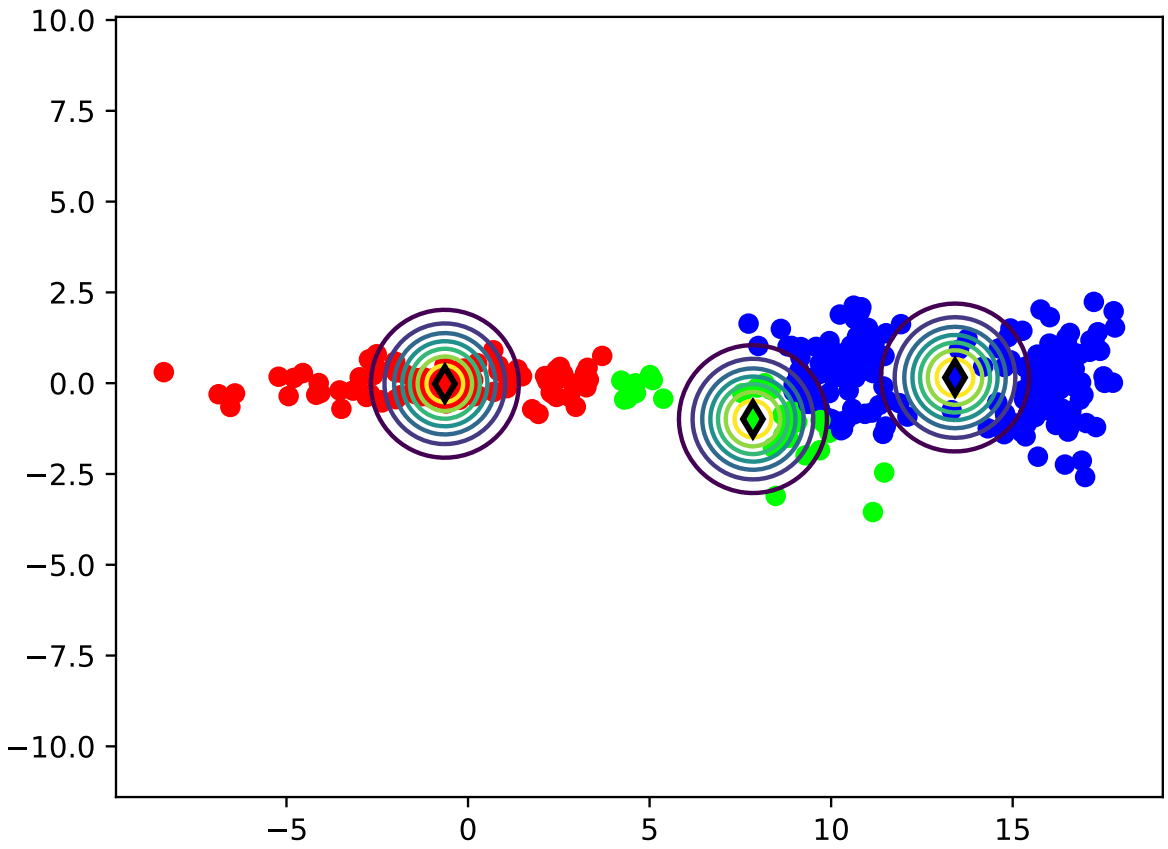
Recall that k-means baked in spherical clusters.



How about we model each cluster with a Gaussian?

k-means has spherical clusters?

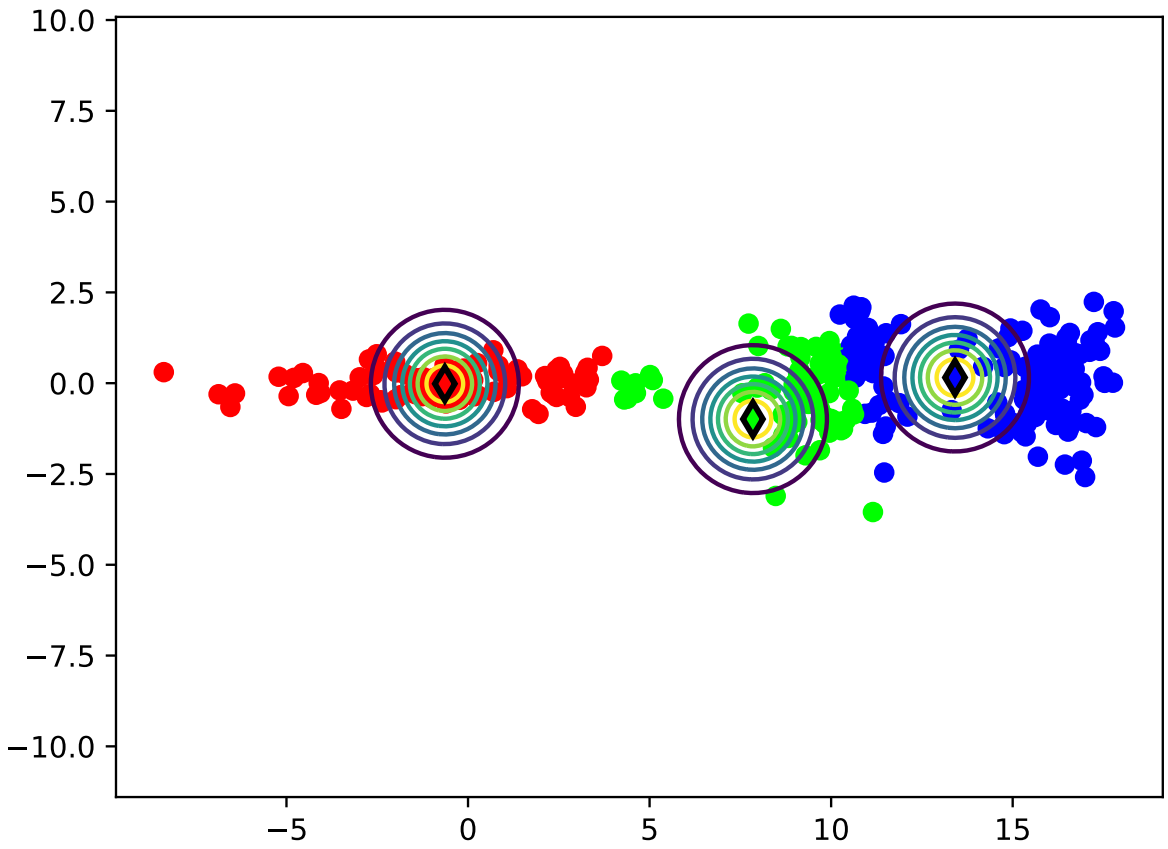
Recall that k-means baked in spherical clusters.



How about we model each cluster with a Gaussian?

k-means has spherical clusters?

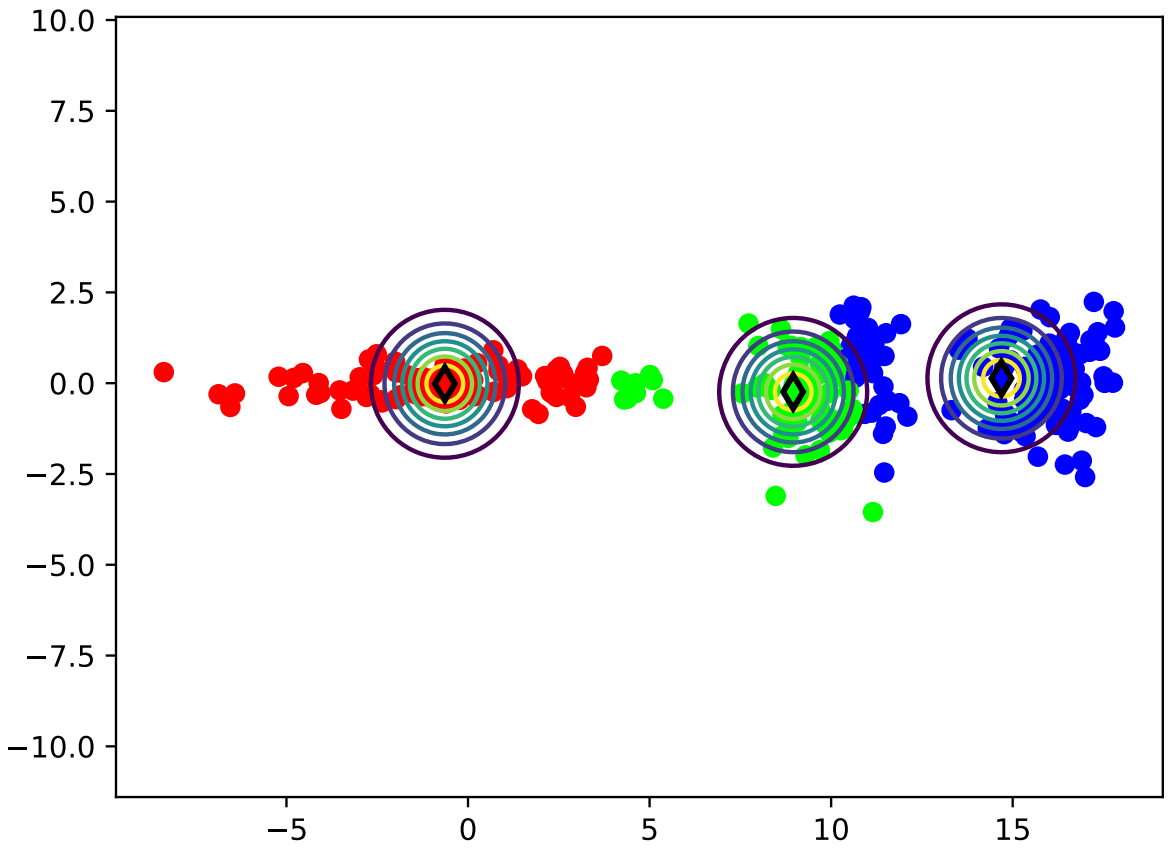
Recall that k-means baked in spherical clusters.



How about we model each cluster with a Gaussian?

k-means has spherical clusters?

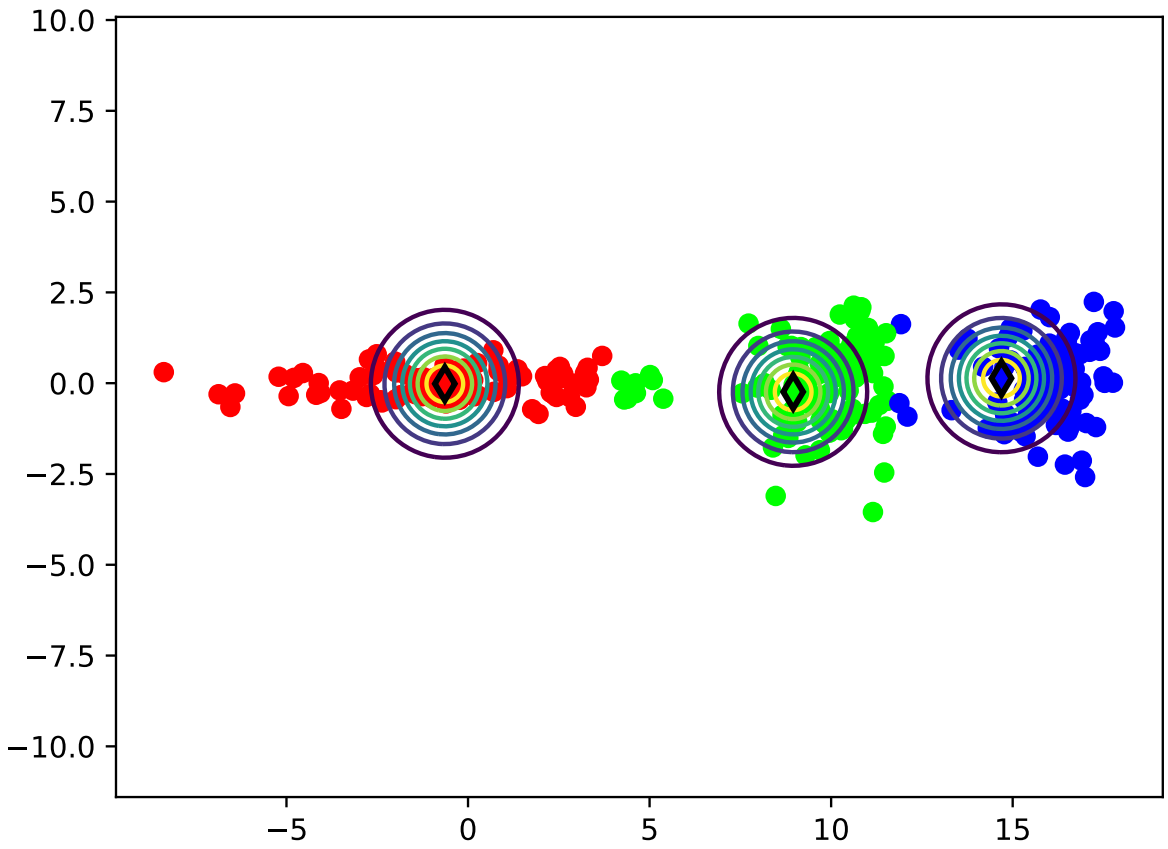
Recall that k-means baked in spherical clusters.



How about we model each cluster with a Gaussian?

k-means has spherical clusters?

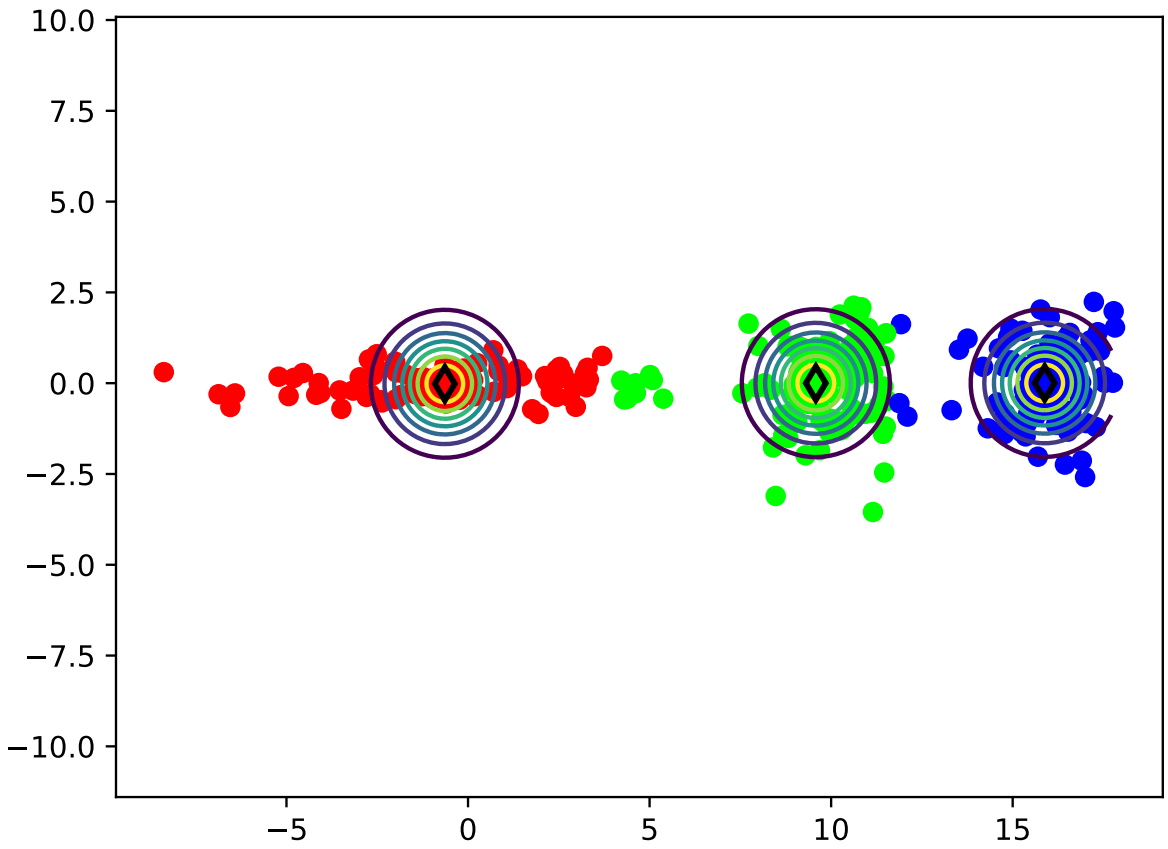
Recall that k-means baked in spherical clusters.



How about we model each cluster with a Gaussian?

k-means has spherical clusters?

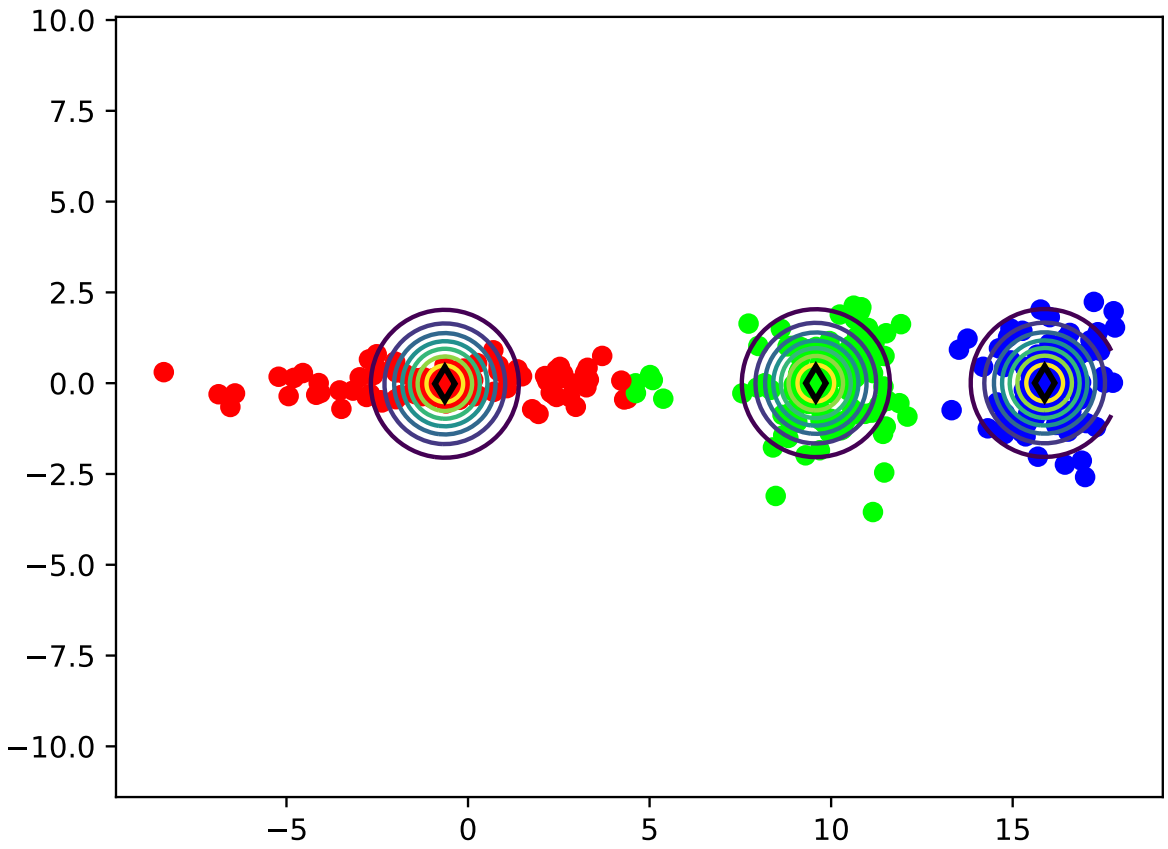
Recall that k-means baked in spherical clusters.



How about we model each cluster with a Gaussian?

k-means has spherical clusters?

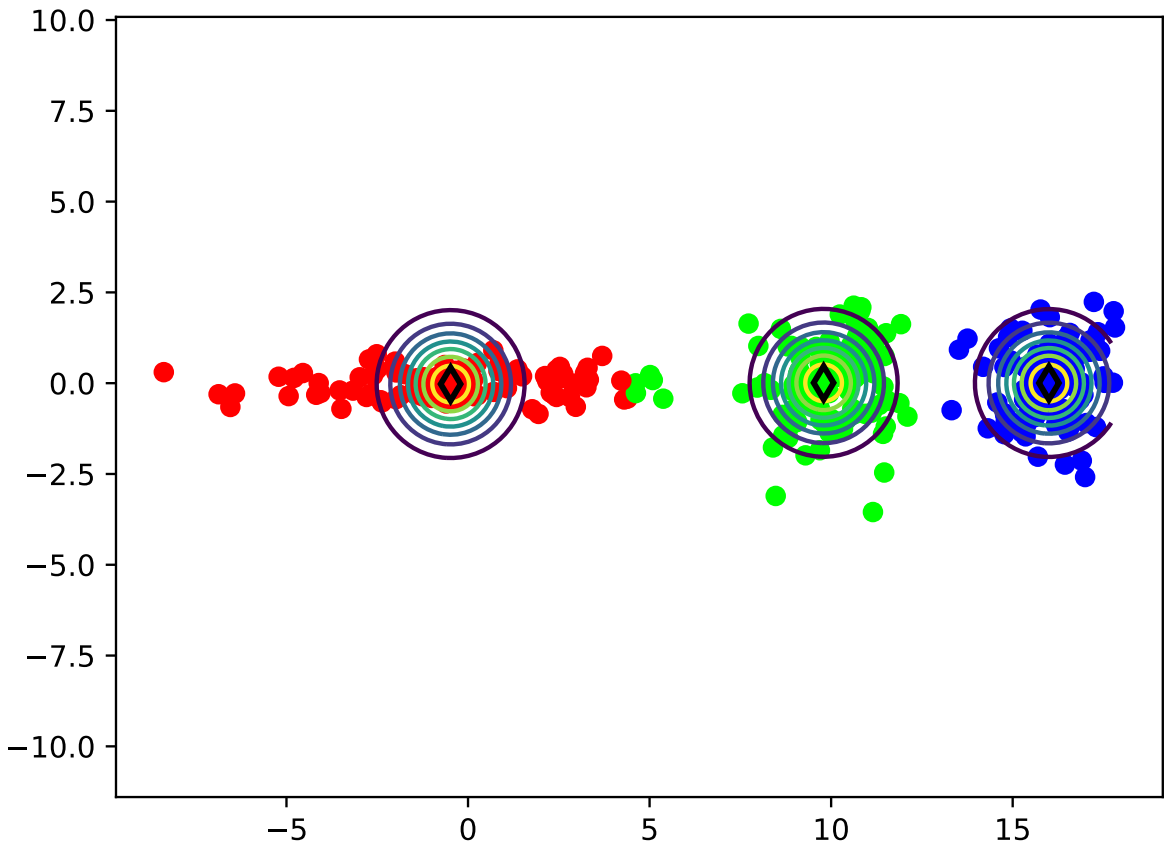
Recall that k-means baked in spherical clusters.



How about we model each cluster with a Gaussian?

k-means has spherical clusters?

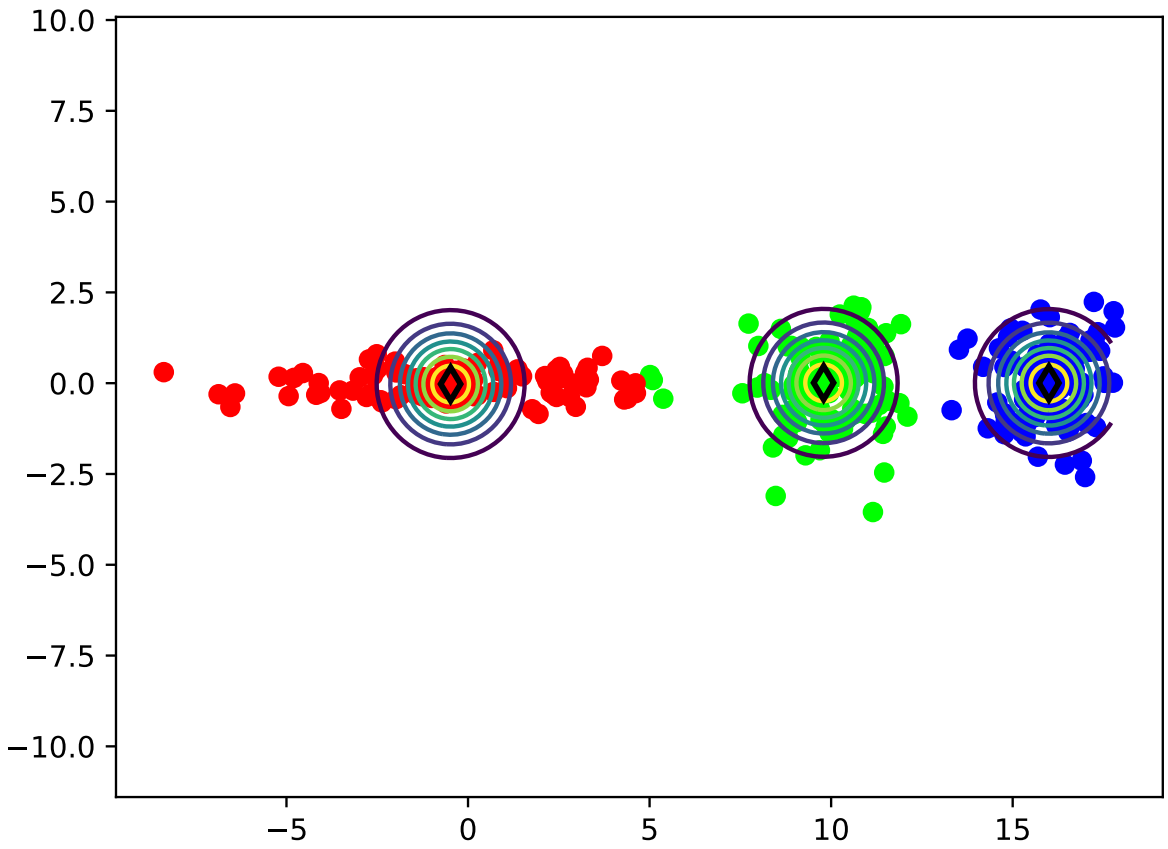
Recall that k-means baked in spherical clusters.



How about we model each cluster with a Gaussian?

k-means has spherical clusters?

Recall that k-means baked in spherical clusters.



How about we model each cluster with a Gaussian?

Gaussian Mixture Model

- Suppose data is drawn from k Gaussians, meaning

$$Y = j \sim \text{Discrete}(\pi_1, \ldots, \pi_k),$$
$$\boldsymbol{X} = \boldsymbol{x} | Y = j \sim \mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j),$$

  and the parameters are $\boldsymbol{\theta} = ((\pi_1, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \ldots, (\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k))$.
  (Note: this is a **generative** model, and we have a way to sample.)
- The probability density at a given $\boldsymbol{x}$ is

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{j=1}^{k} p_{\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j}(\boldsymbol{x} | Y = j) \pi_j,$$

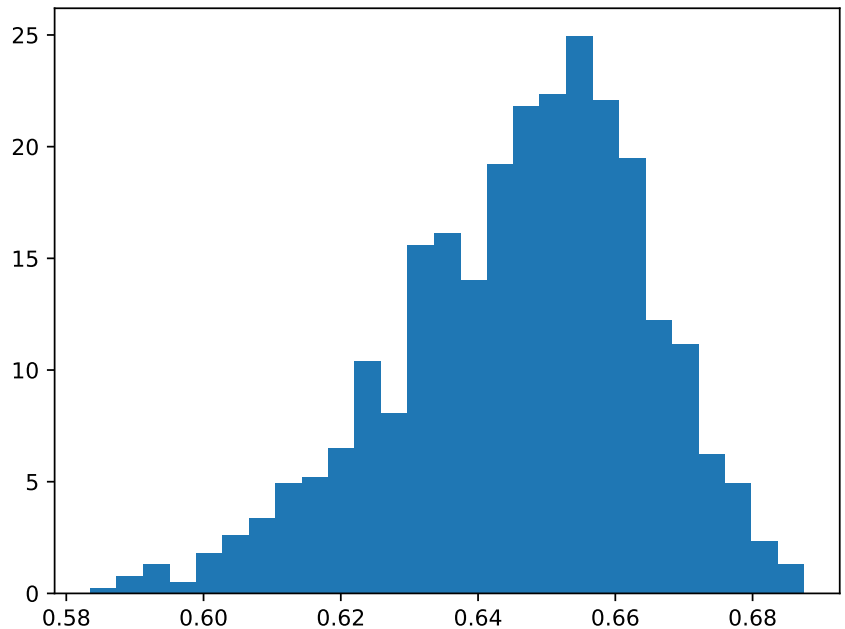
  and the maximum likelihood problem becomes

$$\mathcal{L}((\pi_j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)_{j=1}^k) = \sum_{i=1}^{n} \ln \sum_{j=1}^{k} \frac{\pi_j}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_j)^\intercal \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_j)\right)$$

  The $\ln$ and the $\exp$ are no longer next to each other; we can't just take the derivative and set the answer to 0.

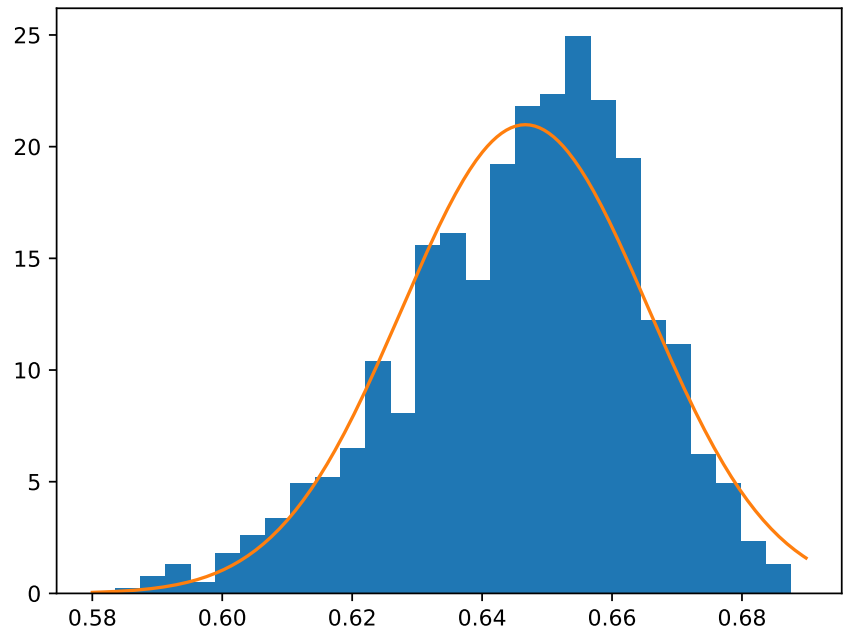
Pearson's crabs.

Statistician Karl Pearson wanted to understand the distribution of "forehead breadth to body length" for 1000 crabs



Pearson's crabs.

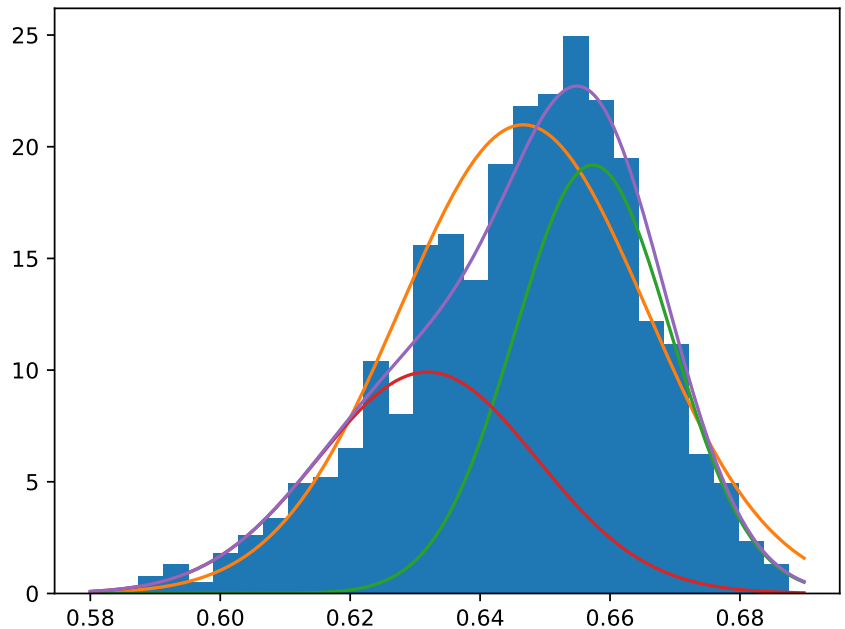
Statistician Karl Pearson wanted to understand the distribution of "forehead breadth to body length" for 1000 crabs



Doesn't look Gaussian!

Pearson's crabs.

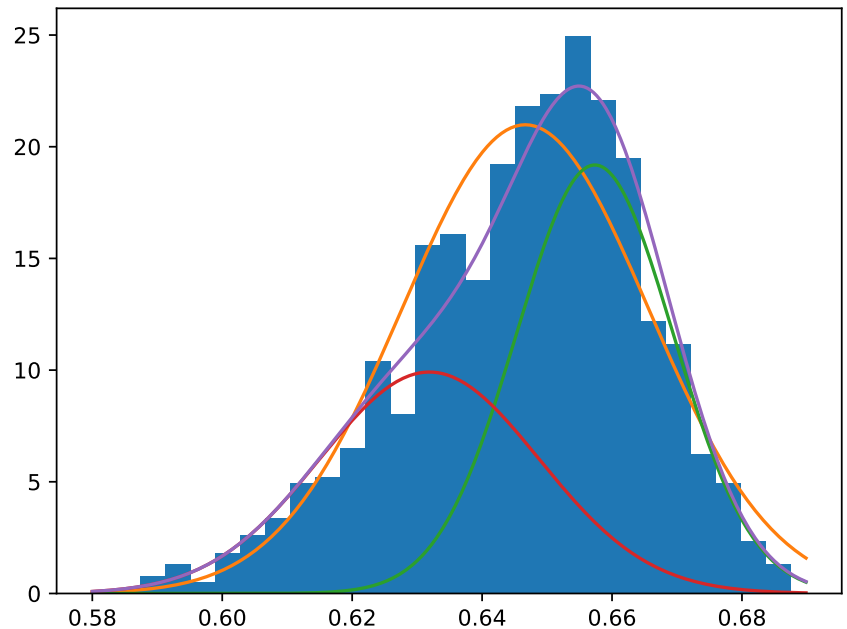
Statistician Karl Pearson wanted to understand the distribution of "forehead breadth to body length" for 1000 crabs



Pearson fit a **mixture of two Gaussians**.

Pearson's crabs.

Statistician Karl Pearson wanted to understand the distribution of "forehead breadth to body length" for 1000 crabs



Pearson fit a **mixture of two Gaussians**.

Remark. Pearson did *not* use E-M. For this he invented the "method of moments" and obtained a solution by hand.

Aside: why Gaussians at all?

- You can argue Gaussian is a good model for *single* populations thanks to the CLT (Central Limit Theorem).
- Pearson, seeing the skewed distribution, felt there are two populations.
- Treating these populations as independent, one gets a mixture of Gaussians.

Summary of part 1.

Summary (of part 1)

- MLE principle; its philosophy and when it might work well.
- Naive Bayes.
- The generative model of Gaussian Mixtures.