Gaussian Mixture Models

Discriminative and Generative Modeling

Discriminative models seek to learn conditional probability:

$$P(y|x)$$

How likely is each label y, given some set of features x

Generative models seek to learn probability distribution:

$$P(x)$$

How likely each data point *x* is

Generative Modeling

Use cases of generative models:

1. Generating new data: If the underlying distribution $P(x)$ is known, then you can sample from that distribution new data points with high probability.

   Language modeling models the distribution of next tokens (words) given some initial set of words

   Most image generators model the probability distribution of images and sample new images from this distribution

2. Anomaly Detection: identify events (data points) that were very unlikely given the underlying distribution $P(x)$

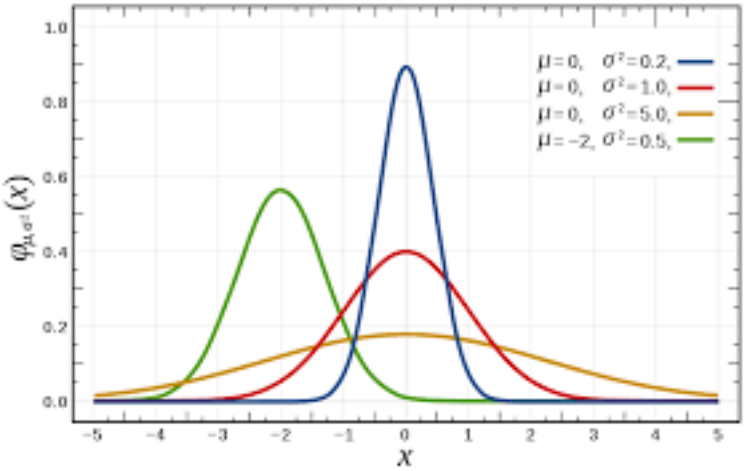
   Credit card fraud detection requires identifying transactions that individuals were *unlikely* to make

Gaussian (Normal) Distributions

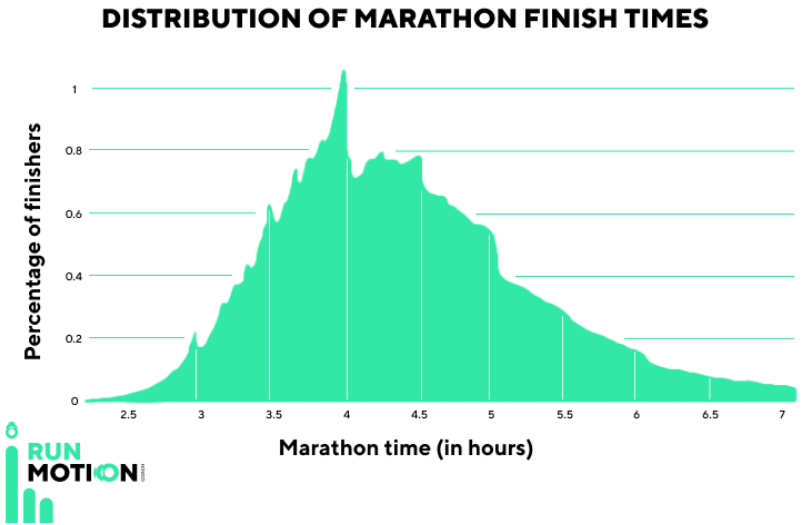
- Defined by mean μ and standard deviation σ
- Probability Density Function: $p(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$

For the gaussian distribution $\mathcal{N}(\mu = 0, \sigma^2 = 1)$, what is the probability of sampling a point $x = 1$?

$$p(2;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(2-\mu)^2}{2\sigma^2}}$$
$$p(2;0,1) = \frac{1}{\sqrt{2\pi 1^2}} e^{\frac{-(2-0)^2}{2 1^2}}$$
$$p(2) = 0.053$$



Some fun gaussian(ish) distributions in the wild



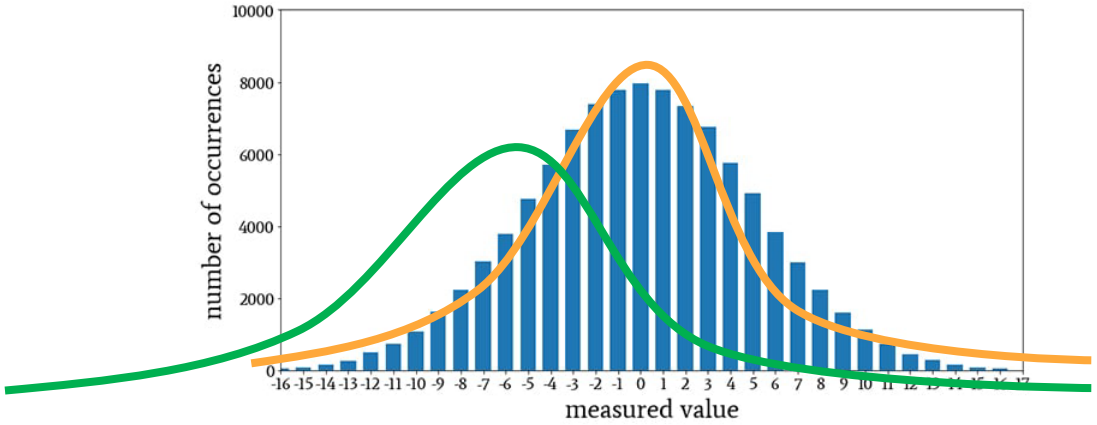
Likelihood

The likelihood of a dataset X and distribution D, is the probability of dataset X being drawn from D.

Intuition: Which Distribution (orange or green) has a higher likelihood of generating X?

Assume each data point x_i is drawn independently from D

Likelihood $\ell(X, D) = \Pi_{i=1}^{n} P(x_i; D)$



Maximum Likelihood Estimation

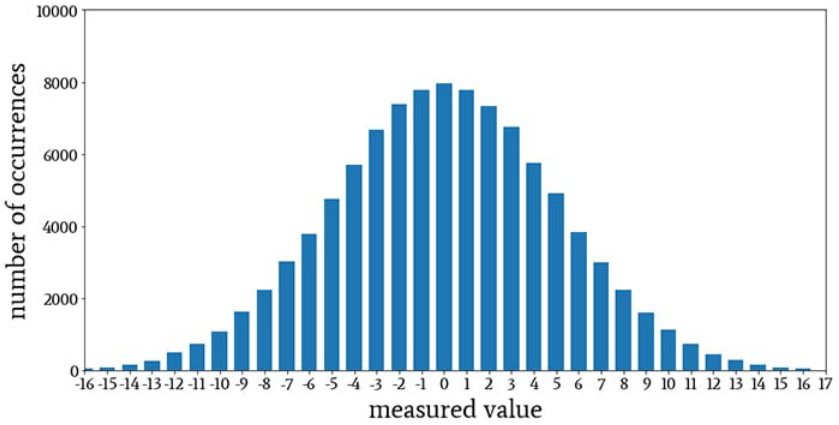
Given some dataset X, what is the most likely distribution D?

1. Assume distribution type (model class). (e.g., assume D is gaussian, categorical, or binomial)
2. Estimate model parameters θ.

Given $X = \{x_1, x_2, \dots, x_n\}$, what values of μ, σ will maximize likelihood?

Likelihood $\ell(X, \mathcal{N}(\mu, \sigma)) = \Pi_{i=1}^{n} P(x_i; \mathcal{N}(\mu, \sigma))$

$$\ell(X, \mathcal{N}(\mu, \sigma)) = \Pi_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$



Maximum Likelihood Estimation (MLE)

How to maximize: $\ell(X, \mathcal{N}(\mu,\sigma)) = \Pi_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}$

Maximizing $\log f(x)$ also maximizes $f(x)$!

$$\ln \ell(X, \mathcal{N}(\mu,\sigma)) = \ln \Pi_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}$$

$$\ln \ell(X, \mathcal{N}(\mu,\sigma)) = \sum_{i=1}^{n} \ln \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}$$

$$\ln \ell(X, \mathcal{N}(\mu,\sigma)) = \sum_{i=1}^{n} \ln \frac{1}{\sqrt{2\pi\sigma^2}} + \ln e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}$$

$$\ln \ell(X, \mathcal{N}(\mu,\sigma)) = \sum_{i=1}^{n} \ln \frac{1}{\sqrt{2\pi\sigma^2}} + -\frac{(x_i-\mu)^2}{2\sigma^2}$$

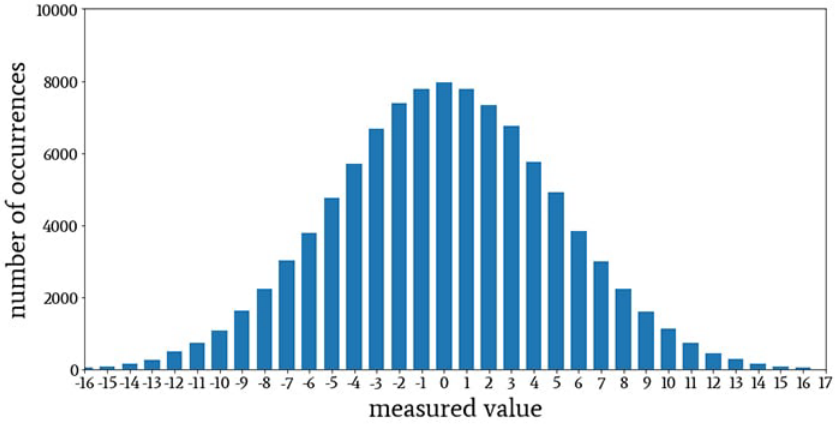
$$\frac{dLL}{d\mu} = \sum_{i=1}^{n} \frac{x_i - \mu}{\sigma^2}$$

$$\frac{dLL}{d\mu} = \frac{1}{\sigma^2} \sum_{i=1}^{n} x_i - \mu$$

$$0 = \sum_{i=1}^{n} x_i - \mu$$

$$0 = n\mu + \sum_{i=1}^{n} x_i$$

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n}$$

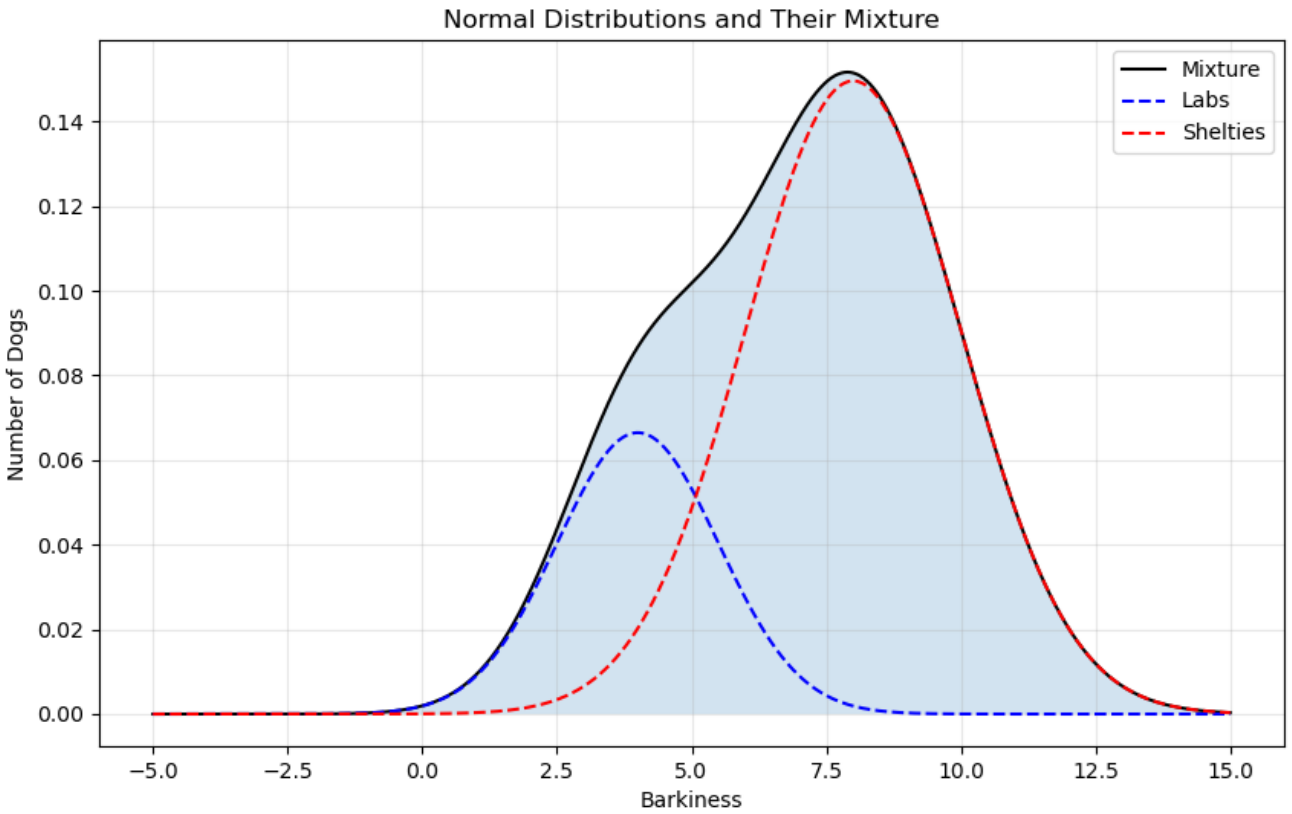


MLE for Gaussian Distribution

Any guess on what MLE will give for σ?

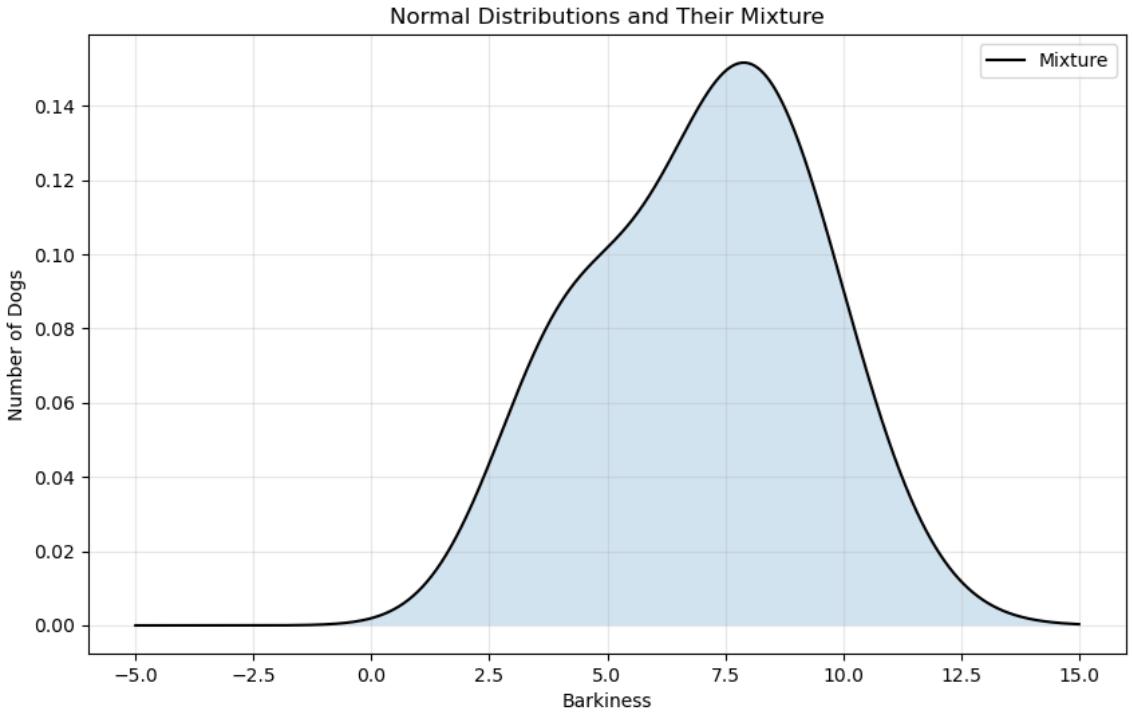
$$\sigma = \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{n}$$

Gaussian Mixture Distribution



Gaussian Mixture Modeling

Goal: Find parameters of K gaussian distributions and their mixing coefficients w_i



Useful variables and definitions

μ_k, σ_k: parameters of k'th gaussian distribution
w_k: How much the k'th gaussian distribution contributes to the overall distribution
$\sum_{k=1}^{K} w_k = 1$
$\theta = \{\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{w}\}$: Parameters of mixture model

$X = \{x_1, x_2, \dots, x_n\}$, observed dataset
z_i: Which gaussian generated example x_i
Probability of a single example x_i

$$P(x_i|\theta) = \sum_{k=1}^{K} P(x_i|z_i = k, \theta)\, \mathrm{P}(z_i = k|\theta)$$

Gaussian Mixture Model Likelihood:

$$\ell(X;\theta) = \prod_{i=1}^{n} P(x_i|\theta) = \prod_{i=1}^{n} \sum_{k=1}^{K} P(x_i|z_i = k, \theta)\mathrm{P}(z_i = k|\theta)$$

Goal #1:
Assume θ is known, compute:
$p(z_i = k|x_i, \theta)$

Goal #2:
Assume z is known, compute θ that maximizes likelihood

An Algorithm for GMMs

Initialize θ randomly

Repeat:

Assume θ is known, compute: $p(z_i = k|x_i, \theta)$:

$$\mathrm{P}(z_i = k|x_i, \theta) = \frac{1}{\sqrt{2\pi{\sigma_k}^2}} e^{\frac{-(x_i-\mu_k)^2}{2{\sigma_k}^2}}$$

Assume z is known, compute θ that maximizes likelihood:

Mixing Parameters: $w_k = \sum_j^{\mathrm{n}} z_j \,/\mathrm{K}$

Means: $\mu_k = \frac{\sum_i^n w_k x_i}{\sum_i^n w_k}$

Standard Deviations: $\sigma_k = \frac{\sum_i^n w_{\mathrm{k}}(x_i-\mu_k)^2}{\sum_i^n w_{\mathrm{k}}}$

Does this feel familiar?

Expectation Maximization (EM) Algorithm

Used for Maximum Likelihood Estimation problems with equations that cannot be solved directly

In the case of GMMs, we have a latent variable z (which Gaussian produced a sample) in addition to the parameters of the model (mixture coefficients and μ, σ)

EM alternates between an *Expectation* step and a *Maximization* step

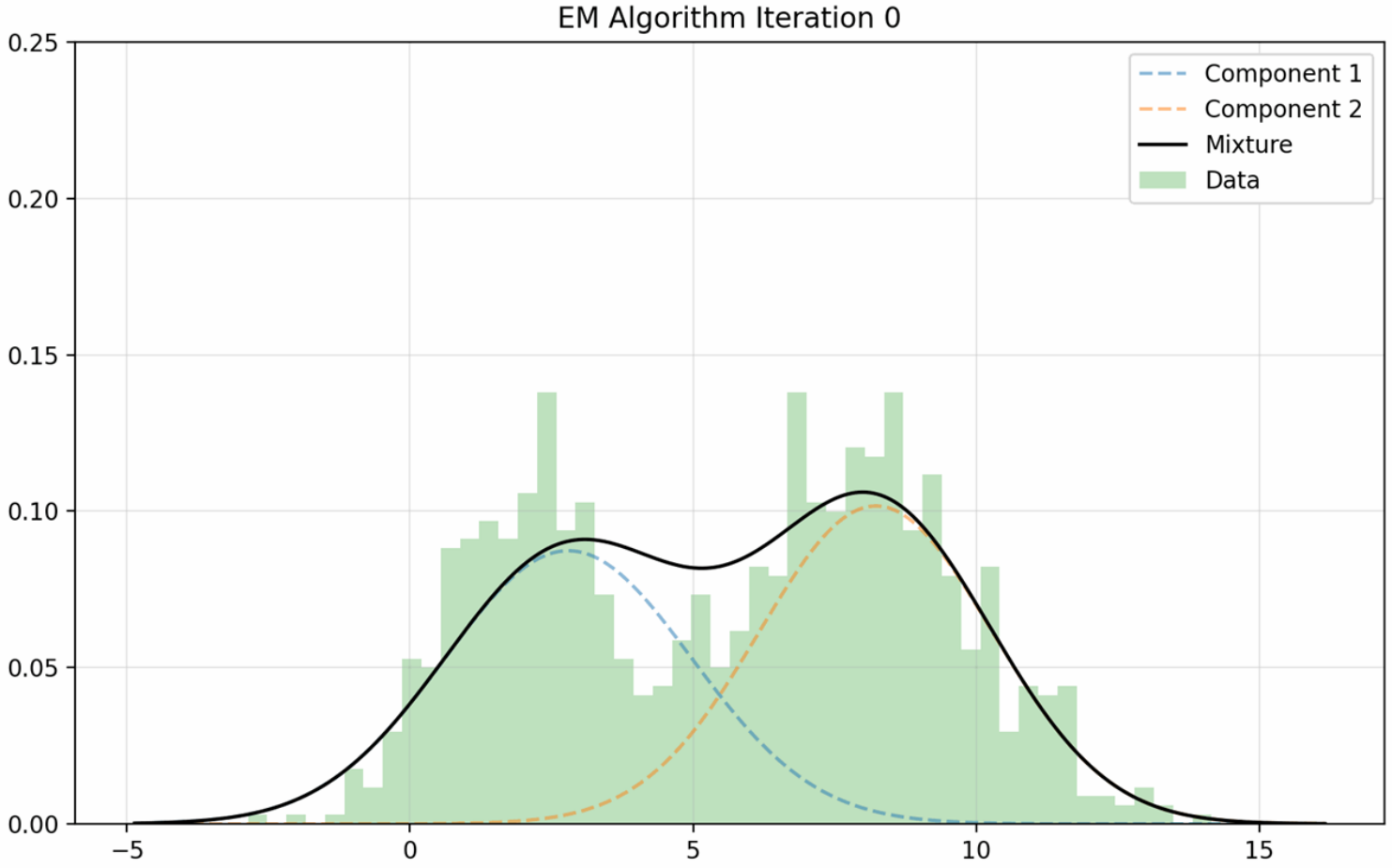
The EM Algorithm

Step 1: Initialize model parameters to be random values

Repeat:

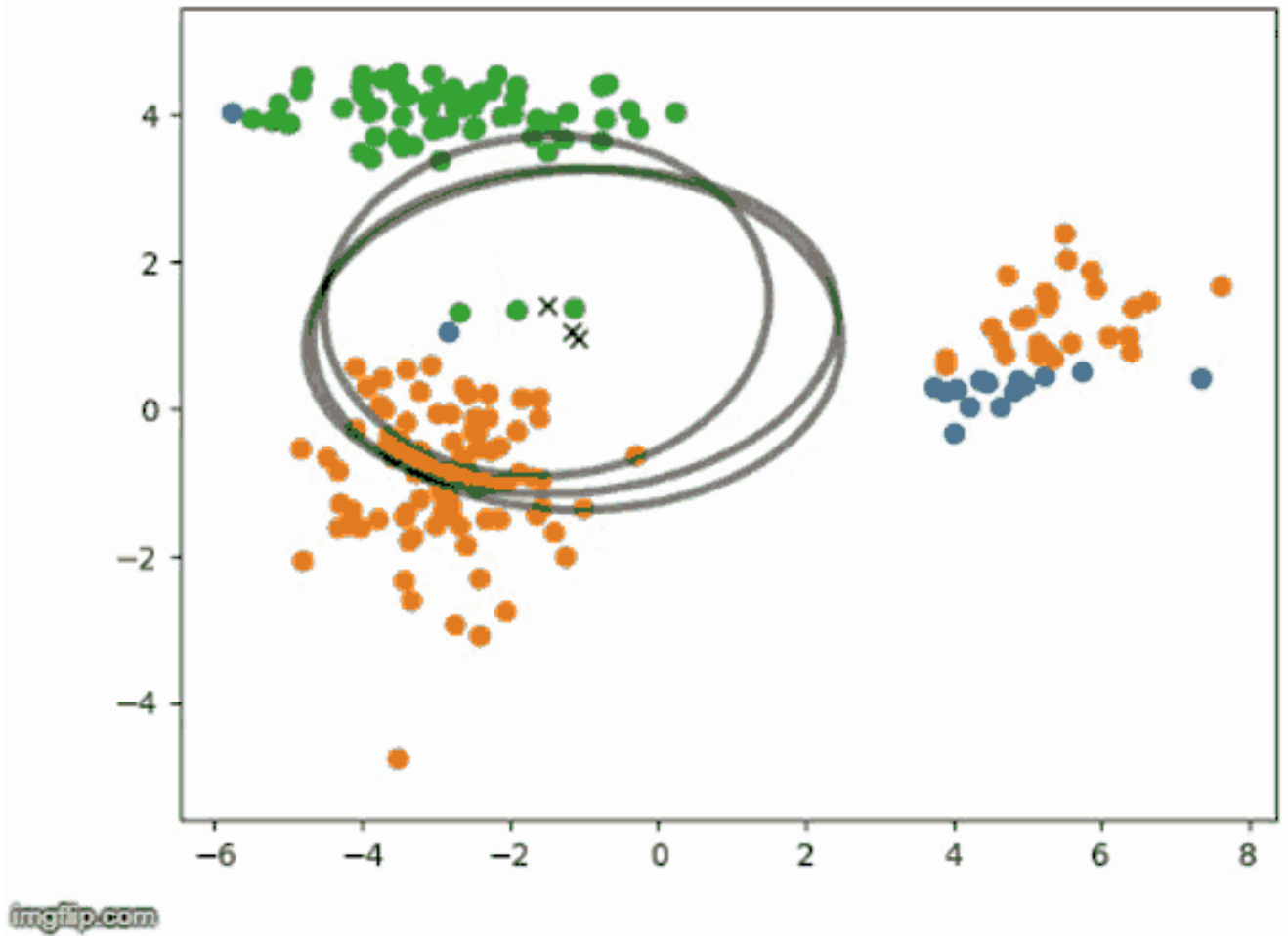
Expectation Step: given model parameters, compute expected latent variables z.

Maximization Step: Given latent variables z, compute model parameters most likely to result in z


EM Algorithm Iteration 0
Component 1
Component 2
Mixture
Data
0.25
0.20
0.15
0.10
0.05
0.00
−5
0
5
10
15

Gaussian Mixture Model (K=3)

Source: https://tenor.com/view/gaussian-mixture-models-em-method-math-gauss-computer-science-nerd-gif-15288262



Credit Card Transactions