Gaussian Mixture Models

Table of contents

- Univariate Gaussian Distribution
- Multivariate Gaussian Distribution
- Gaussian Mixture Models
    - Mixture Models
    - Defining Gaussian Mixture Models (GMMs)

Univariate Gaussian Distribution

A Gaussian distribution is a continuous probability distribution that is characterized by its symmetrical bell shape.

A Gaussian distribution of a single random variable is defined as follows:

$$p(x \mid \mu, \sigma^2) = \mathcal{N}(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Here, μ and σ are scalars representing the mean and standard deviation of the distribution.

We like Gaussians because they have several nice properties, for e.g., marginals and conditionals of Gaussians are still Gaussians.

Multivariate Gaussian Distribution

When dealing with multiple random variables, we need a version of the Gaussian distribution that is able to capture the uncertainty of all them simultaneously.

As is most natural when generalizing to higher dimensions, we turn to linear algebra for this task.

Firstly, D multiple random variables can be collectively described as a vector $\boldsymbol{x}$ in $\mathbb{R}^D$

$$\boldsymbol{x} = [x_1, x_2, x_3, \ldots, x_D]$$

Using this vector representation, the multivariate Gaussian distribution is defined as follows:

$$p(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{D}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right)$$

Here, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are not scalars but a vector of means and a matrix of variance-covariances respectively.

The value $\left|\boldsymbol{\Sigma}\right|$ is the determinant of $\boldsymbol{\Sigma}$ and D is the number of dimensions $\boldsymbol{x} \in \mathbb{R}^D$ that depends on the number of random variables in consideration.

The D-dimensional vector of means $\boldsymbol{\mu}$ consists of the means of the corresponding random variables

$$\boldsymbol{\mu} = [\mu_1, \mu_2, \mu_3, \ldots, \mu_D]$$

The variance-covariance matrix is a symmetric, positive semidefinite, $D \times D$ matrix that is defined in index notation as

$$\boldsymbol{\Sigma}_{ij} = \begin{cases} \mathrm{Var}(x_i) & i = j \\ \mathrm{Cov}(x_i, x_j) & i \neq j \end{cases}$$

Since $\mathrm{Cov}(x, y) = \mathrm{Cov}(y, x)$ is generally true, the symmetric nature of this matrix $\mathbf{\Sigma}$ follows as a natural consequence.

The positive semidefiniteness of $\mathbf{\Sigma}$ means that for any $\boldsymbol{x} \in \mathbb{R}^D$,

$$\boldsymbol{x}^{\mathrm{T}} \, \mathbf{\Sigma} \, \boldsymbol{x} \geq 0$$

We know that the inverse of a positive semidefinite matrix is also positive semidefinite, thus,

$$\boldsymbol{x}^{\mathrm{T}} \, \mathbf{\Sigma}^{-1} \, \boldsymbol{x} \geq 0$$

This implies that in the definition of the multivariate Gaussian distribution, the term

$$(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \geq 0$$

Which is consistent with what we see in the analogous term in the univariate Gaussian distribution, namely that

$$\frac{(x-\mu)^2}{2\sigma^2} \geq 0$$

The $\left|\boldsymbol{\Sigma}\right|^{-\frac{1}{2}}$ term is analogous to the $\frac{1}{\sqrt{\sigma^2}}$ term in the univariate distribution and for each random variable we have, we multiplicatively pick up a factor of $(2\pi)^{-\frac{1}{2}}$, resulting in the $(2\pi)^{-\frac{D}{2}}$ term we see.

Furthermore, observe that the term in the exponent

$$(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$$

Is an inner product with $\boldsymbol{\Sigma}^{-1}$ as the inner product matrix. This can be proven by working out the dimensions

$$(1\times D)\cdot(D\times D)\cdot(D\times 1)\equiv(1\times 1)$$

Which is a scalar.

Finally, we would like to point out that interpolating over $\boldsymbol{\mu}$ has the effect of shifting the multidimensional Gaussian on the D-dimensional (hyper)plane, whereas changing the matrix $\mathbf{\Sigma}$ has the effect of changing the multidimensional shape of the Gaussian.

Gaussian Mixture Models

In many real world scenarios, when it comes to trying to model some data using a statistical distribution, there are uncertainties and mechanisms that cannot be captured solely by a simplistic model such as a single Gaussian distribution. A practical example where this is true is in the method of density estimation.

It is more often than not the case that the data in question is likely to have been produced by multiple underlying sub-distributions associated with some hidden or *latent* variables.

We want to find a way to represent this presence of sub-populations within the overall population when trying to model our data. In such situations, we turn to more comprehensive approaches such as mixture models.

Mixture Models

Suppose we assume that each data point x_n has been produced by a latent variable z, and express this causal relation as $z \to x$. A straightforward approach would be to assume that z is a categorical distribution representing K underlying distributions.

In this scenario, each data point can be mapped to a specific distribution by considering z as a one-hot vector that identifies the membership of that data point to a component distributiion.

For instance,

$$\boldsymbol{z} = [z_1, z_2, z_3]^{\mathrm{T}} = [0, 1, 0]^{\mathrm{T}}$$

means that the data point belongs to the second component distribution.

This corresponds to a *hard assignment* of each point to its generative distribution. In reality, we do not have access to this one-hot vector of sorts, and we instead impose a distribution over z representing a soft assignment:

$$p(\boldsymbol{z}) = \boldsymbol{\pi} = [\pi_1, \ldots, \pi_K]^{\mathrm{T}}$$

Here, π_k is the probability that a particular data point belongs to the k-th component distribution, so we have the constraints that

$$0 \leq \pi_k \leq 1$$

$$\sum_{k=1}^{K} \pi_k = 1$$

Now, each data point does not exclusively belong to a certain component distribution, but to all of them with different probabilities. This approach is what defines *mixture models*

Defining Gaussian Mixture Models (GMMs)

We now consider mixture models where the underlying component distributions are Gaussians. This is convenient for a number of reasons as outlined before (marginals and conditionals of Gaussians are still Gaussians).

The way this works is that each Gaussian would have its own mean and variance and we could mix them by adjusting the proportional probability coefficients π_k

This is analogous to a Fourier series decomposition or say mixing different sounds by using sliders on a console.

This approach is called *Gaussian Mixture Models*

Deriving the likelihood of a GMM from our latent variable framework is quite straightforward.

We first collect the parameters of all the component Gaussians into a vector $\boldsymbol{\theta}$, defined as

$$\boldsymbol{\theta} = \{\mu_k, \sigma_k, \pi_k\}_{k=1}^{K}$$

The likelihood $p(x|\boldsymbol{\theta})$ is obtained through the marginalization of the latent variable z, which in our case consists of summing out all the latent variables from the joint distribution $p(x, z)$. By the product rule of probability distributions, this can be written down as

$$p(x|\boldsymbol{\theta}) = \sum_{z} p(x|\boldsymbol{\theta}, z)p(z|\boldsymbol{\theta})$$

Now, recall that $p(x|\boldsymbol{\theta}, z_k)$ is a Gaussian distribution $\mathcal{N}(x|\mu_k, \sigma_k)$ with z consisting of K such Gaussian components,

$$p(x|\boldsymbol{\theta}, z_k) = \mathcal{N}(x|\mu_k, \sigma_k)$$

A specific weight π_k represents the probability of the k-th component being the generating distribution $p(z_k = 1|\boldsymbol{\theta})$

$$p(z_k = 1|\boldsymbol{\theta}) = \pi_k$$

By turning the sum over z in the marginalization step into a sum over k as z is after all indexed by k, and using the knowledge in the previous two statements, it follows that a GMM with K *univariate* Gaussian components can be defined as

$$\mathcal{N}(\mu_k, \sigma_k) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \sigma_k)$$

and under the obvious constraints of

$$0 \leq \pi_k \leq 1$$

$$\sum_{k=1}^{K} \pi_k = 1$$

Similarly, we can define a GMM for the *multivariate* case:

$$\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Under identical constraints for π.

GMMs are more expressive than simple Gaussians and they are often able to capture subtle differences in data more effectively.