**Expectation-Maximization Algorithm**

The Expectation-Maximization Algorithm, or EM algorithm for short, is an approach for maximum likelihood estimation in the presence of latent variables

*A general technique for finding maximum likelihood estimators in latent variable models is the expectation-maximization (EM) algorithm.*

— Page 424, [Pattern Recognition and Machine Learning](https://amzn.to/2JwHE7I), 2006.

The EM algorithm is an iterative approach that cycles between two modes. The first model attempts to estimate the missing or latent variables, called the estimation-step or E-step. The second model attempts to optimize the parameters of the model to best explain the data, called the maximization-step or M-step.

* **E-Step**. Estimate the missing variables in the dataset.
* **M-Step**. Maximize the parameters of the model in the presence of the data.

The EM algorithm can be applied quite widely, although is perhaps most well known in machine learning for use in unsupervised learning problems, such as density estimation and clustering.

Perhaps the most discussed application of the EM algorithm is for clustering with a mixture model.

**Gaussian Mixture Model and the EM Algorithm**

A [mixture model](https://en.wikipedia.org/wiki/Mixture_model) is a model comprised of an unspecified combination of multiple probability distribution functions.

A statistical procedure or learning algorithm is used to estimate the parameters of the probability distributions to best fit the density of a given training dataset.

The Gaussian Mixture Model, or GMM for short, is a mixture model that uses a combination of Gaussian (Normal) probability distributions and requires the estimation of the mean and standard deviation parameters for each.

There are many techniques for estimating the parameters for a GMM, although a maximum likelihood estimate is perhaps the most common.

Consider the case where a dataset is comprised of many points that happen to be generated by two different processes. The points for each process have a Gaussian probability distribution, but the data is combined and the distributions are similar enough that it is not obvious to which distribution a given point may belong.

The processes used to generate the data point represents a latent variable, e.g. process 0 and process 1. It influences the data but is not observable. As such, the EM algorithm is an appropriate approach to use to estimate the parameters of the distributions.

In the EM algorithm, the estimation-step would estimate a value for the process latent variable for each data point, and the maximization step would optimize the parameters of the probability distributions in an attempt to best capture the density of the data. The process is repeated until a good set of latent values and a maximum likelihood is achieved that fits the data.

* **E-Step**. Estimate the expected value for each latent variable.
* **M-Step**. Optimize the parameters of the distribution using maximum likelihood.

We can imagine how this optimization procedure could be constrained to just the distribution means, or generalized to a mixture of many different Gaussian distributions.