

# 1 Overview of Semi-Supervised Learning Literature

## 1.1 Overview of general models

### 1.1.1 Generative Mixture Models and EM

This model works by assuming a generative model:

$$p(x, y) = p(y)p(x|y)$$

where  $p(x|y)$  is an identifiable mixture distribution such as Gaussian mixture models. The idea is then with a large amount of unlabelled data that the mixture components can be identified, and then ideally only one labelled example per component is needed to fully determine the mixture distribution. A general way of thinking of this is that the mixture components are "soft clusters".

**Gaussian distribution** The gaussian distribution is historically called the law of errors. It was by Gauss to model errors in astronomical observations, which is why it is usually referred to as the Gaussian distribution. The formula for the Gaussian distribution given mean  $\mu$  and standard deviation  $\sigma$  is given by:

$$\phi(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp - \frac{(x - \mu)^2}{2\sigma^2}$$

An important concept within mixture models in general is identifiability. In general let  $\{p_\theta\}$  be a family of distributions indexed by a parameter vector  $\theta$ .  $\theta$  is identifiable if  $\theta_1 \neq \theta_2 \Rightarrow p_{\theta_1} \neq p_{\theta_2}$ . If the model family is identifiable, in theory with infinite  $U$  one can learn  $\theta$  up to permutation of component indices. In contrast an example where the problem of an unidentifiable is shown. Given a model  $p(x|y)$  that is uniform for

## 2 Convergence

Convergence in the context of Machine Learning when the training-validation error is so close to local/global minimum or you can see it having a performance so close to local/global minimum that usually no significant error decrease/increase in performance anymore.

## Bibliography

- [Jer05] Xiaojin Zhu (Jerry). “Semi-Supervised Learning Literature Survey”. In: (2005). URL: <http://digital.library.wisc.edu/1793/60444>.
- [Van16] Jake VanderPlas. *Python Data Science Handbook*. 2016. URL: <https://jakevdp.github.io/PythonDataScienceHandbook/05.12-gaussian-mixtures.html>.