Machine learning L11 Semi-supervised learning

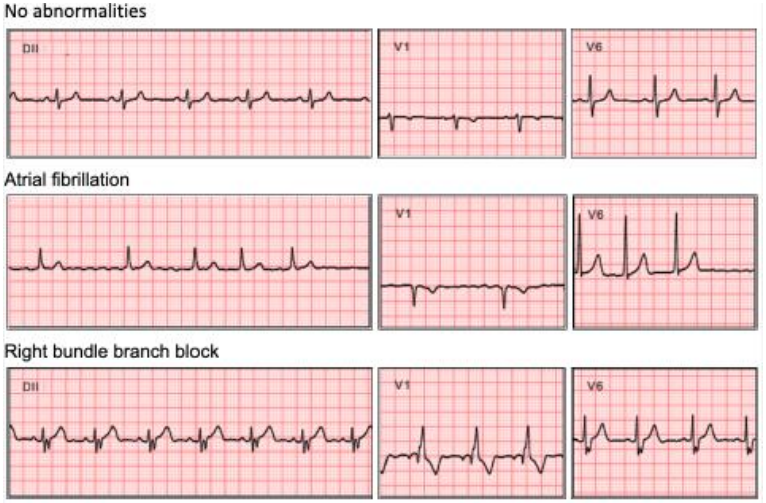
Frank Miller, Department of Statistics

May 25, 2021

Stockholm University

Labeled and unlabeled data

- In many situations, it is easy/cheap to obtain unlabeled data but difficult/expensive to obtain labeled data
- Examples:
  - ECG classification



Left picture from: Lindholm A, Wahlström N, Lindsten F, Schön TB (2021). Machine Learning – a first course for engineers and scientists. http://smlbook.org

  - Image classification




Stockholm University

Labeled and unlabeled data

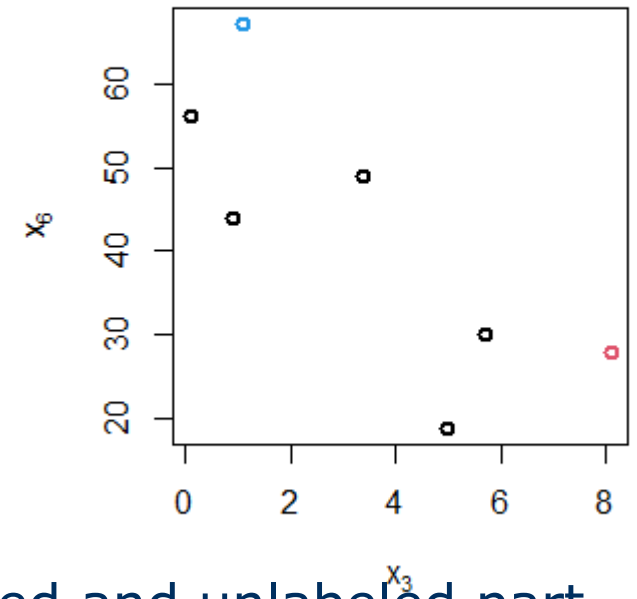
- We will discuss two approaches to deal with situations where labeling is difficult/expensive:

- **Semi-supervised learning**
  We have a dataset which is partly labeled
- **Active machine learning**
  We have an unlabeled (or partly labeled) dataset but have the opportunity to choose some data points to be labeled by an expert

Semi-supervised learning

- The data available for training consists in such situations of many data points where a few have labels and many are unlabeled, e.g.:

x_1	x_2	x_3	x_4	x_5	x_6	x_7	y
1	0	1.1	0	1	67	0	1
0	0	0.9	0	4	44	1	NA
0	0	3.4	1	3	49	1	NA
1	1	8.1	1	3	28	1	2
0	1	0.1	1	2	56	1	NA
1	0	5.0	0	1	19	1	NA
1	0	5.7	1	1	30	1	NA



- Information is in both labeled and unlabeled part

Stockholm University

Semi-supervised learning

- We know well how to deal with labeled data (supervised learning)
- In a previous lecture, you have discussed how to obtain information from unlabeled data (unsupervised learning)
- We want to use now both, the information in the labels and the huge amount of unlabeled data

Blood pressure example

- Effect of a drug to be measured and n patients (randomly chosen out of a population of patients) treated with the drug
- X_i, $i=1,\ldots,n$, observed for each patient after drug-treatment (reduction in blood pressure in mmHg)
- Known that population consists of two groups:
  - Group 1 ($Y_i=1$) responds only barely to drug (smaller X_i)
  - Group 2 ($Y_i=2$) responds well (larger X_i)
- Based on genetics, we could determine if someone belongs to Group 1 or 2
- But most patients are not genotyped (then Y_i =NA)

Model for blood pressure example

- Data generative model assumed:
  Gaussian mixture model for X_i with two components (non-responder (Y=1) and responder (Y=2) population)
- Model:

$$p(y = 1) = \pi_1 = \pi, \qquad p(y = 2) = \pi_2 = 1 - \pi$$
$$p(x|y = m) = \mathcal{N}(x|\mu_m; \sigma_m), \qquad m = 1,2$$

- 5 parameters: $\pi;\ \mu_1; \sigma_1; \mu_2; \sigma_2$
- Density for unlabeled data points:

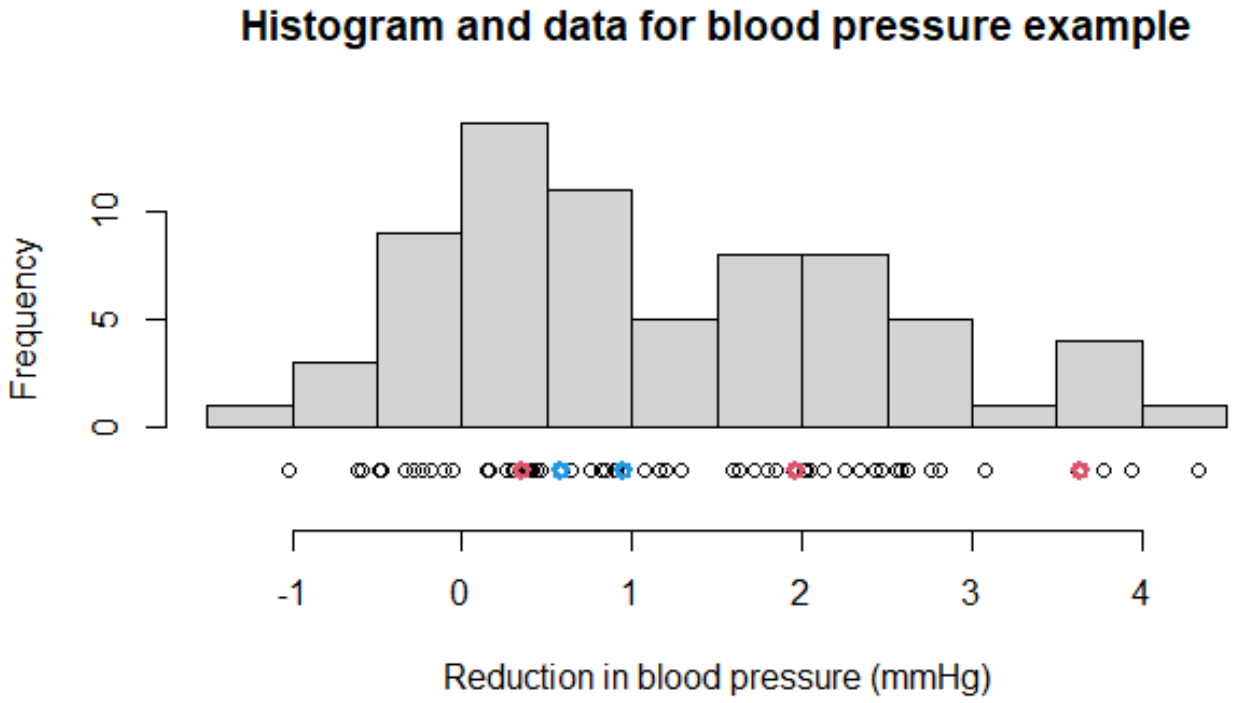
$$p(x) = \pi\mathcal{N}(x|\ \mu_1; \sigma_1) + (1 - \pi)\mathcal{N}(x|\mu_2; \sigma_2)$$

  with π=probability to be a non-responder

- EM algorithm can compute parameters

Blood pressure example

- Blood pressure observed after drug treatment for $n=70$ patients; five patients genotyped (three of them belong to Responder-group 2, red; two to Group 1, blue)



Stockholm University

EM algorithm for semi-supervised case

- In the E-step, the probabilities to belong to group m, $w_i(m)$, are predicted
- Unsupervised case:

$$w_i(m) = p\left(y = m \middle| \boldsymbol{x}_i, \widehat{\boldsymbol{\theta}}\right) = \frac{\hat{\pi}_m \mathcal{N}(\boldsymbol{x}_i | \hat{\mu}_m; \hat{\sigma}_m)}{\sum_{j=1}^{M} \hat{\pi}_j \mathcal{N}(\boldsymbol{x}_i | \hat{\mu}_j; \hat{\sigma}_j)}$$

- Semi-supervised case:

$$w_i(m) = \begin{cases} p\left(y = m \middle| \boldsymbol{x}_i, \widehat{\boldsymbol{\theta}}\right), & \text{if } y_i \text{ is missing} \\ 1, & \text{if } y_i = m \\ 0, & \text{otherwise} \end{cases}$$

EM algorithm for semi-supervised case

- E-step:

$$w_i(m) = \begin{cases} p(y = m | \boldsymbol{x}_i, \widehat{\boldsymbol{\theta}}), & \text{if } y_i \text{ is missing} \\ 1, & \text{if } y_i = m \\ 0, & \text{otherwise} \end{cases}$$

- M-step (no difference to unsupervised case):

$$\hat{\pi}_m = \frac{1}{n} \sum_{i=1}^{n} w_i(m) \,,$$

$$\widehat{\boldsymbol{\mu}}_m = \frac{1}{\sum_{i=1}^{n} w_i(m)} \sum_{i=1}^{n} w_i(m) \boldsymbol{x}_i \,,$$

$$\widehat{\boldsymbol{\Sigma}}_m = \frac{1}{\sum_{i=1}^{n} w_i(m)} \sum_{i=1}^{n} w_i(m) (\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_m)(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_m)^T \,.$$

EM algorithm for semi-supervised case

- Expected log-likelihood (which is maximized):

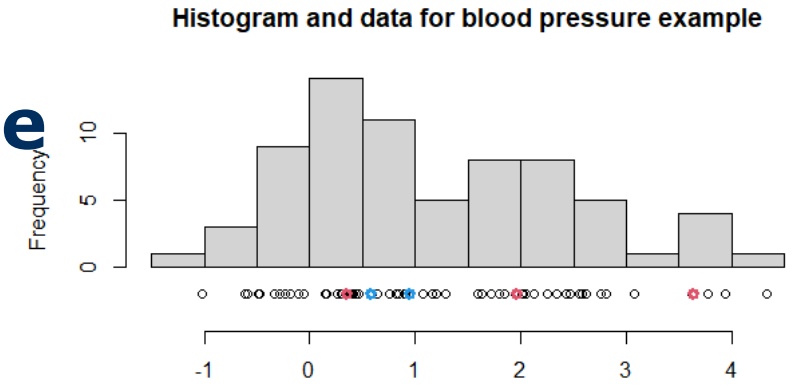
$$Q(\theta) = \sum_{i=1}^{n} \sum_{m=1}^{M} w_i(m) \{ \ln \mathcal{N}(\boldsymbol{x}_i | \hat{\mu}_m; \hat{\sigma}_m) + \ln \pi_{y_i} \}$$

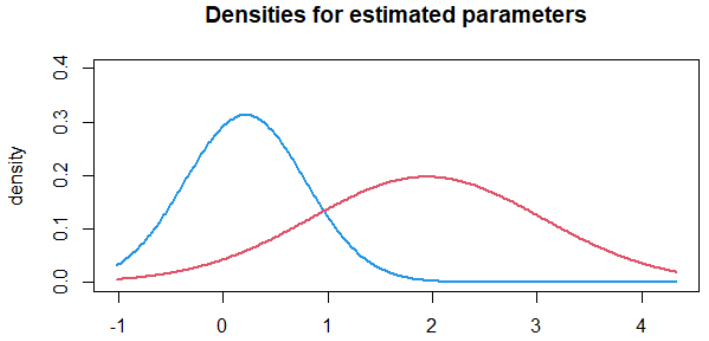
- This is the same as in the unsupervised case, but $w_i(m)$ is now set to 0 or 1 for the labeled data points
- Expectation over log-likelihood done for unlabeled points
- Stopping rule for algorithm can be based on $Q(\theta)$:
  If change in $Q(\theta)$ between two iterations small, stop

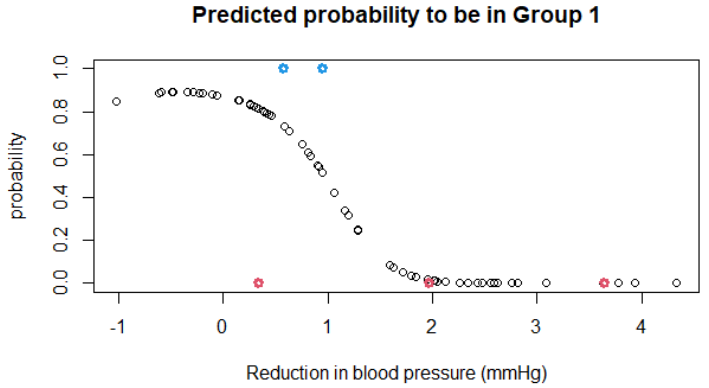
- **`GMM_EM.R`** on homepage (see L9) can be modified from unsupervised to semi-supervised learning

Blood pressure example

- EM algorithm estimates model parameters (corresponding densities in figure)
- EM algorithm provides also probabilities to belong to Group 1 (non-responders) for each unlabeled data point



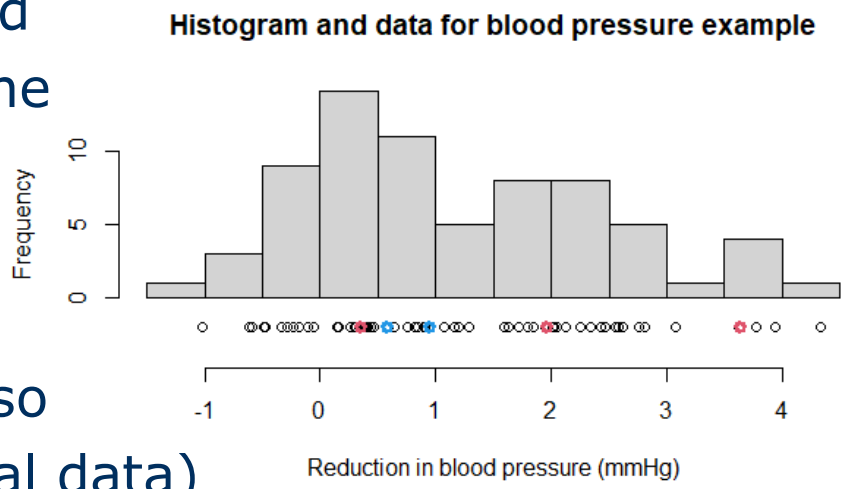




Stockholm University

Choosing starting values for the EM algorithm in Gaussian mixture models

- We can look at the data and guess the components in the mixture, their mean and variance (in a simulation study, looking at data and deciding is not possible; also difficult for high dimensional data)

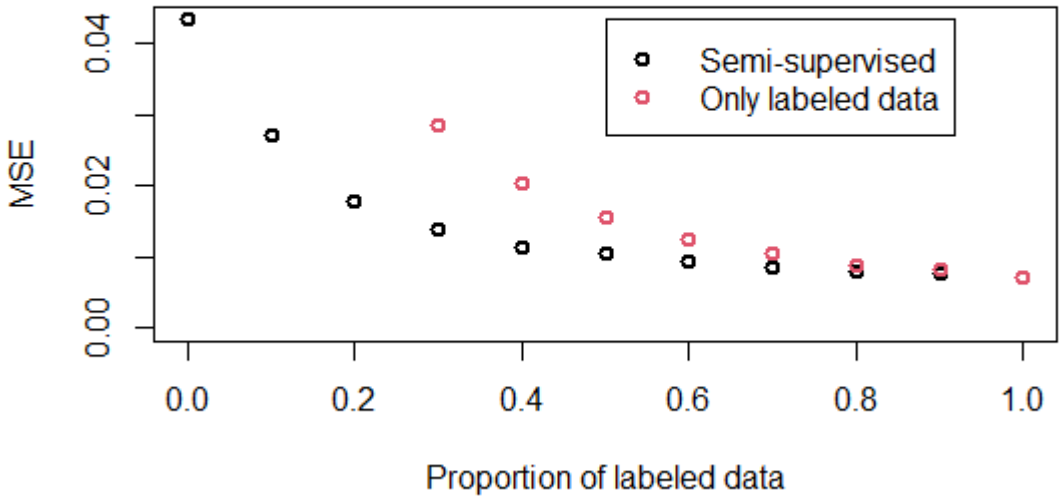


- We can use a heuristic rule to determine starting values
- We can try a grid of starting parameter values and choose then the best result
- We can first run some other classification algorithm

Stockholm University

Gain from semi-supervised learning (compared to supervised)

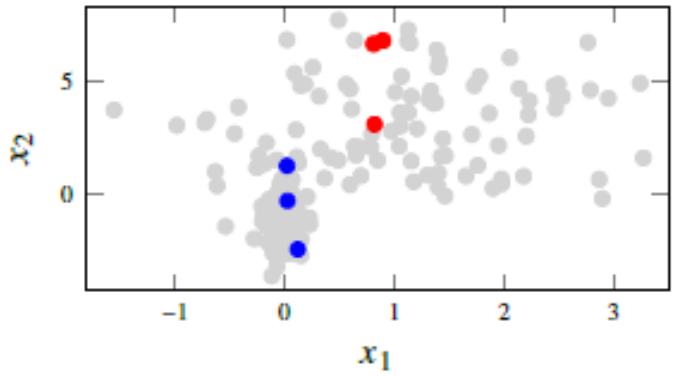
- A naïve way to analyze partially labeled data is to ignore the unlabeled part
- In context of previous example, we simulated data repeated times and calculated the mean squared error (MSE) of the parameter estimates
- Done for n=70 and ratio of labeled data of r=0, 0.1, …, 1



Stockholm University

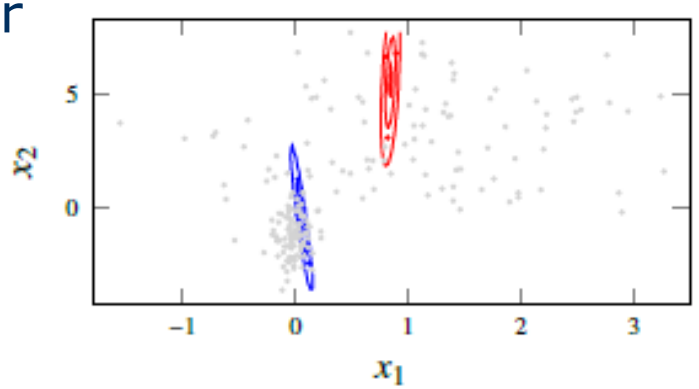
Gain from semi-supervised learning (compared to supervised)

- Unlabeled data (grey) and 6 data points which are labeled (red/blue)



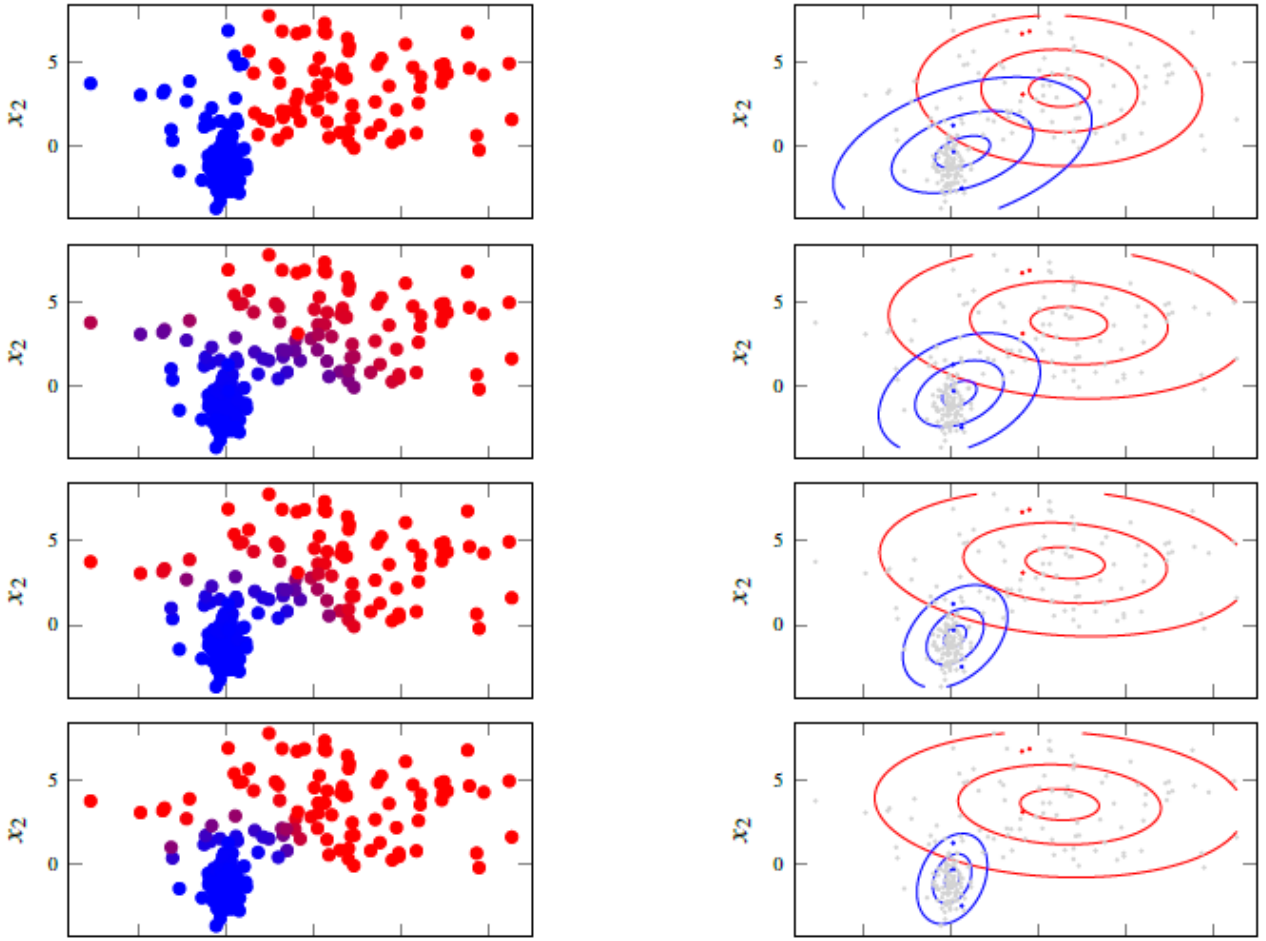
Pictures from: Lindholm et al. (2021). http://smlbook.org

- If model is trained using labeled data only (supervised); result is poor



Stockholm University

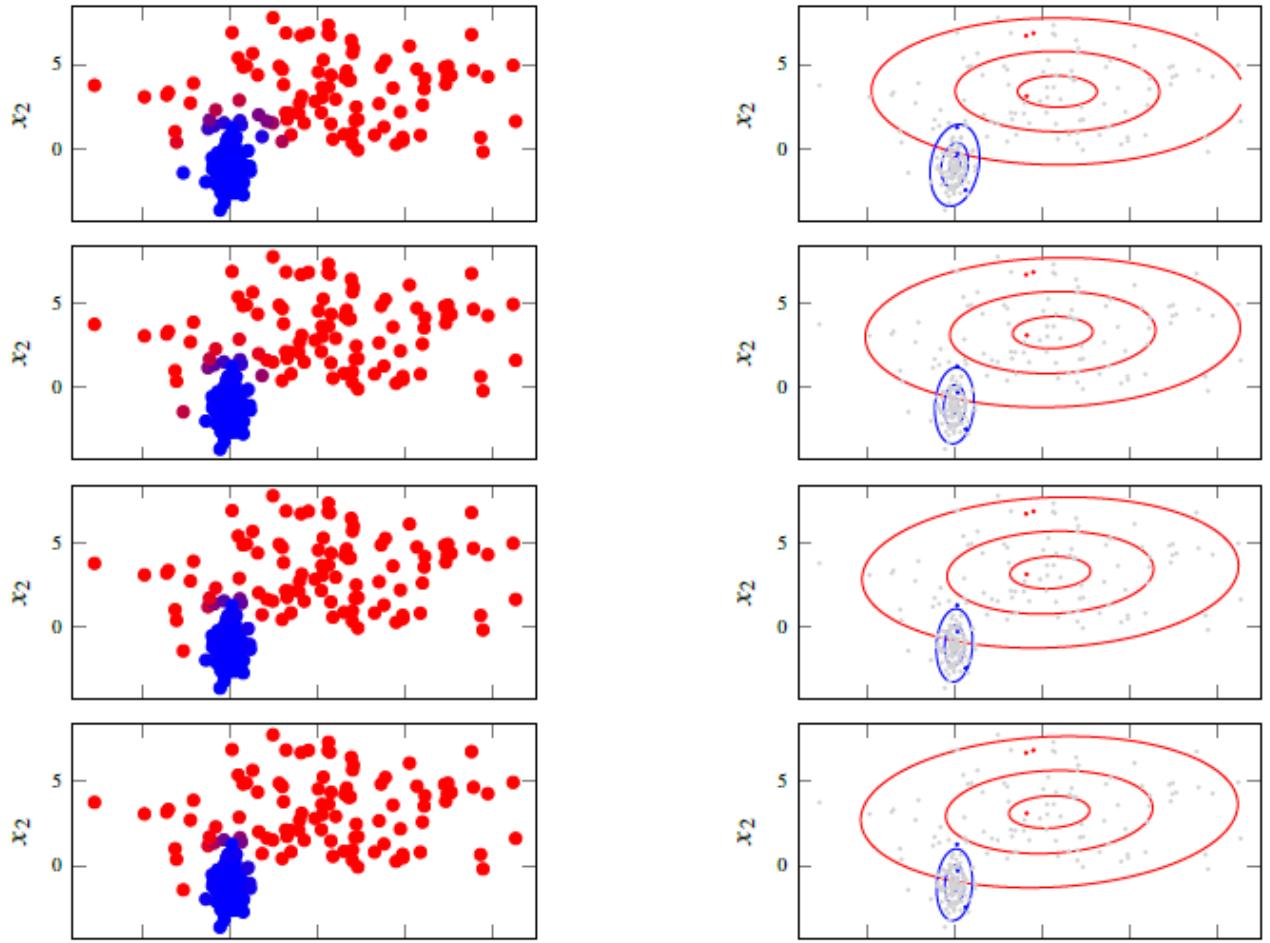
EM algorithm for semi-supervised learning



Picture from: Lindholm et al. (2021). http://smlbook.org

Stockholm University

EM algorithm for semi-supervised learning

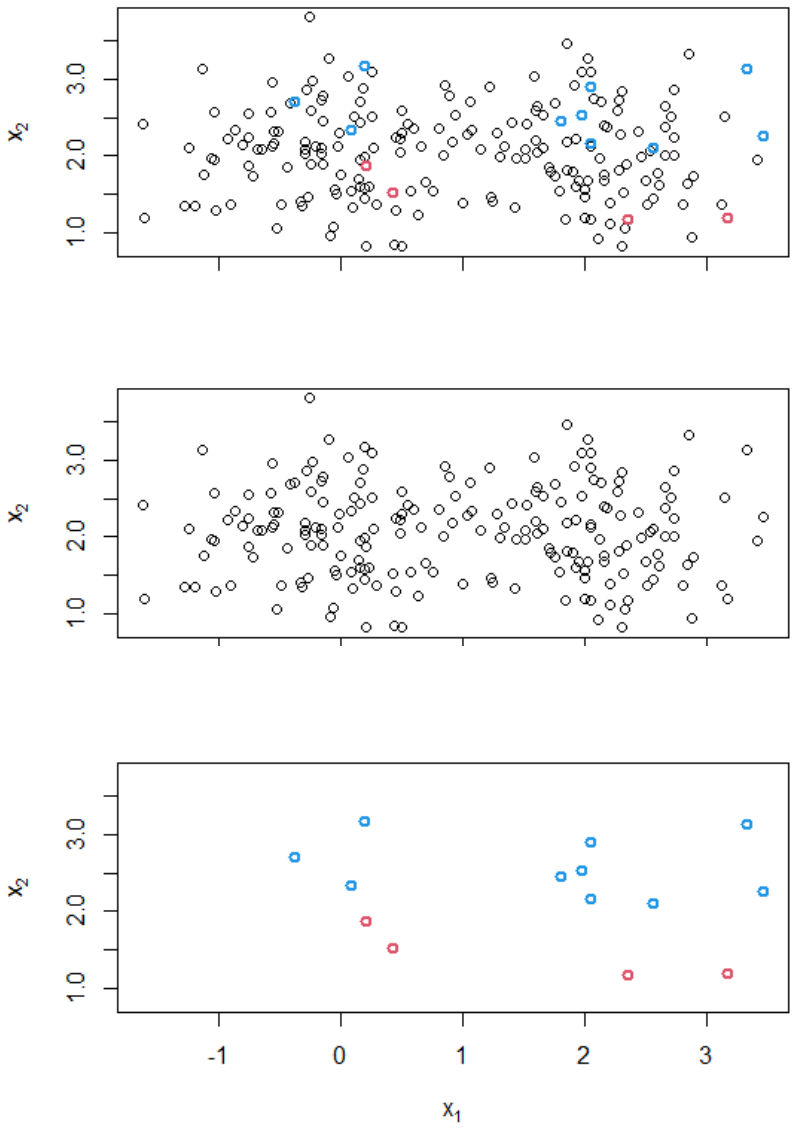


Picture from: Lindholm et al. (2021). http://smlbook.org

Stockholm University

Supervised, semi-supervised, unsupervised

- How would you analyze this semi-supervised data?
- Would you trust a data generative model like GMM for x-data without labels?
- Or would you focus on the much smaller dataset with labeled data?



Supervised, semi-supervised, unsupervised

- Say that we trust our generative model (GMM for semi-supervised data)
- Which of these two results is the best classification (highest log-likelihood)?