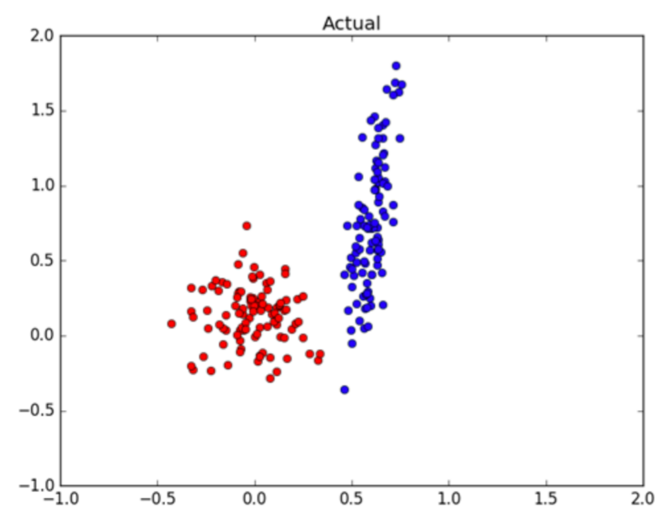
Expectation Maximization and Gaussian Mixture Model

https://www.cs.utah.edu/~piyush/teaching/EM_algorithm.pdf
https://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm
https://www.kdnuggets.com/2016/08/tutorial-expectation-maximization-algorithm.html

Introduction

- We are presented with some unlabelled data and we are told that it comes from a multivariate Gaussian distribution.
- Our task is to come up with the hypothesis for the means and the variances of each distribution
- For example, we have data drawn from two Gaussians. We need to estimate the means and variances of the x's and the y's of the blue and red distribution. How are we going to do this?



Back to Clustering

Classification. Training two Gaussians given data labeled $+, -$

Clustering. Training two Gaussians given unlabeled data

Algorithms.

1. k-Means
   a. Given hard labels, compute centroids
   b. Given centroids, compute hard labels
2. Expectation-Maximization
   a. Given soft labels, compute Gaussians
   b. Given Gaussians, compute soft labels

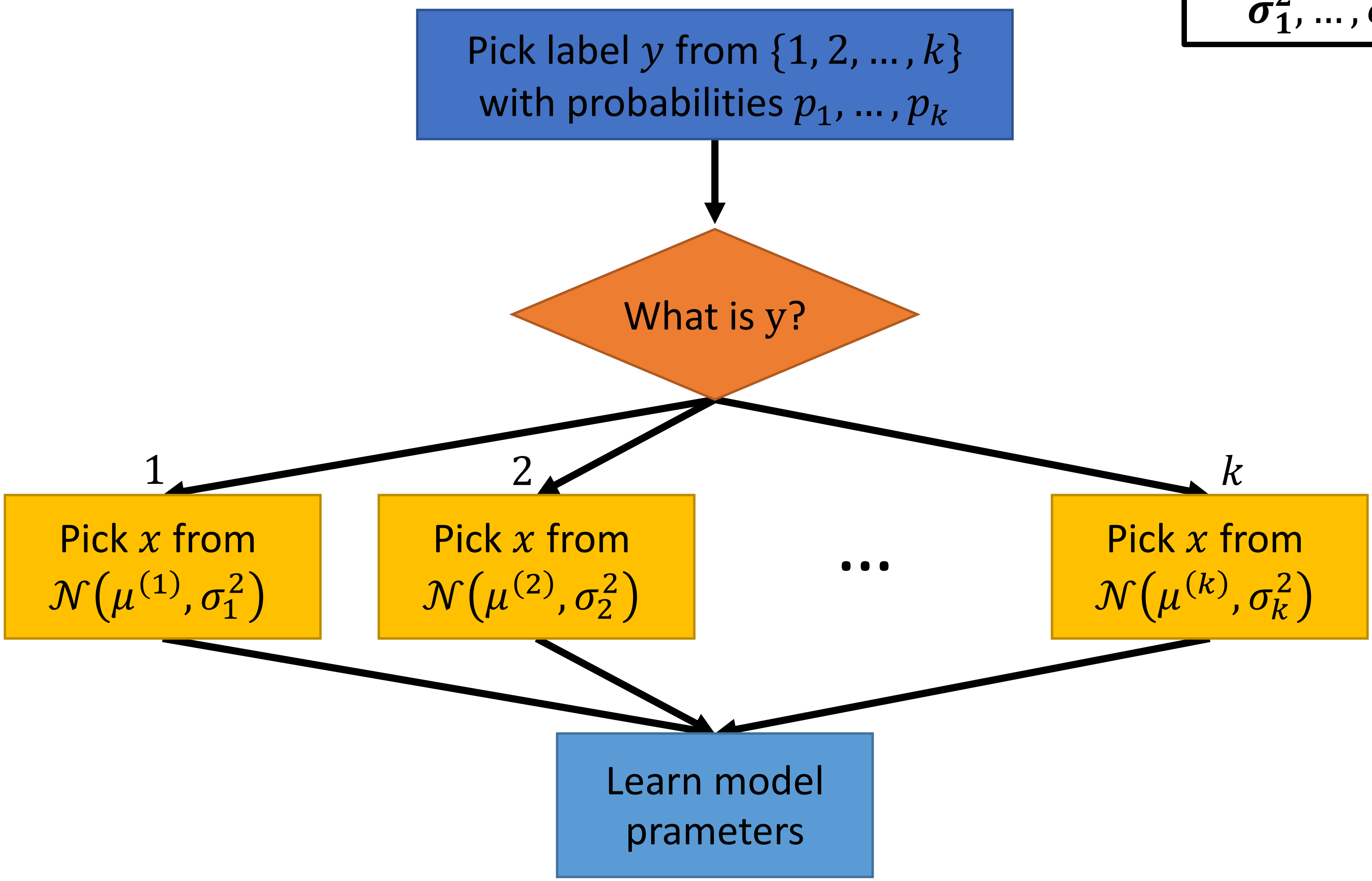
Generative Model

Model Parameters

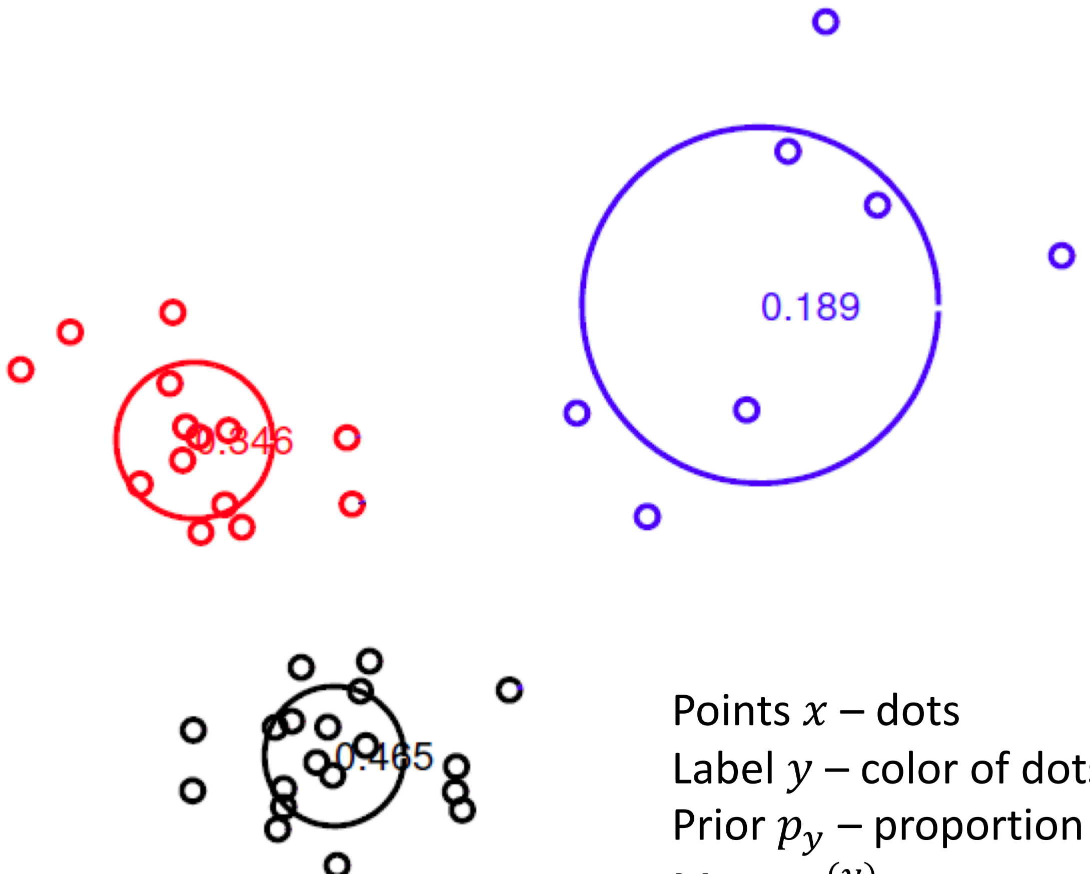
$\boldsymbol{p_1, \dots, p_k}$

$\boldsymbol{\mu^{(1)}, \dots, \mu^{(k)}}$

$\boldsymbol{\sigma_1^2, \dots, \sigma_k^2}$

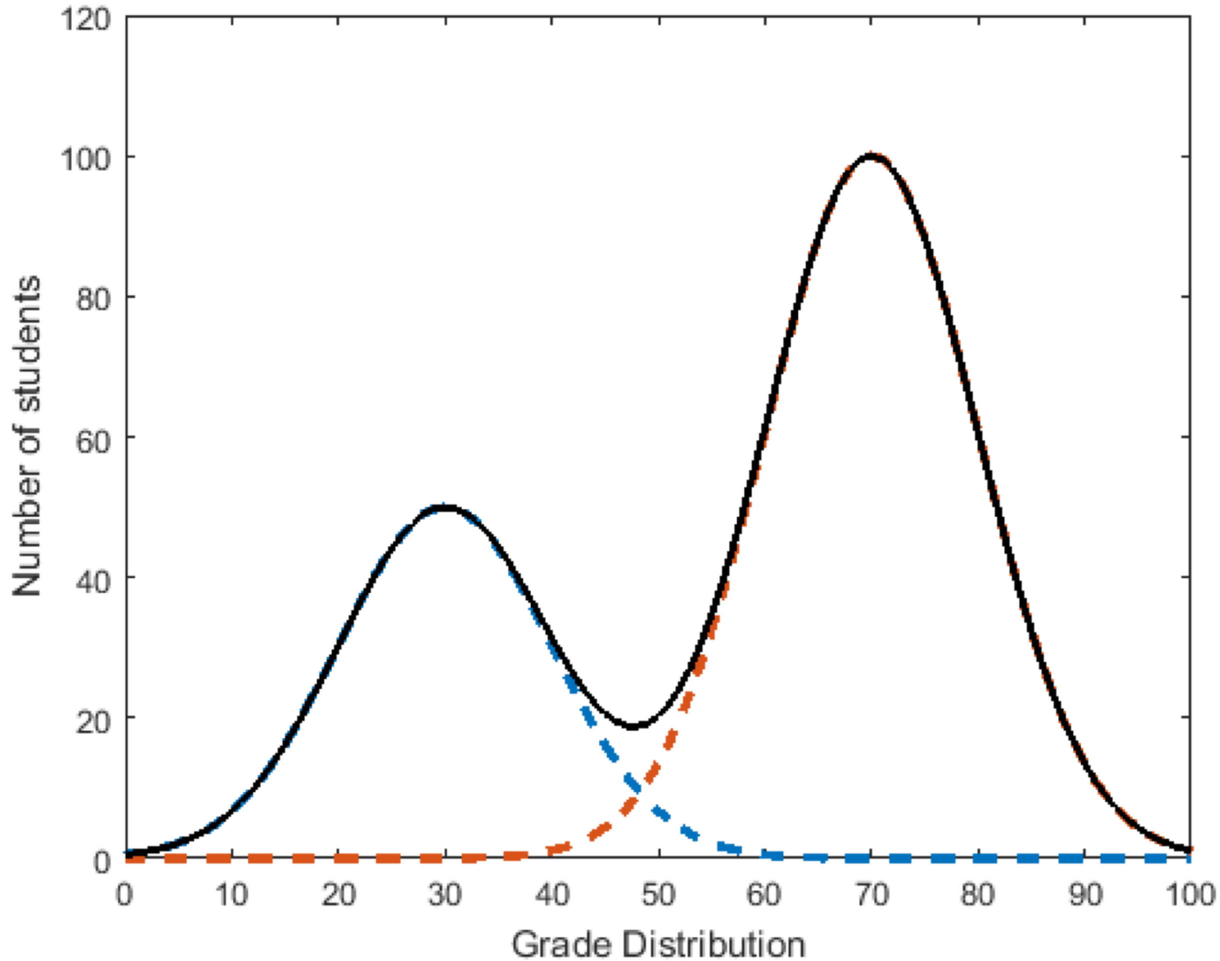


Generative Model

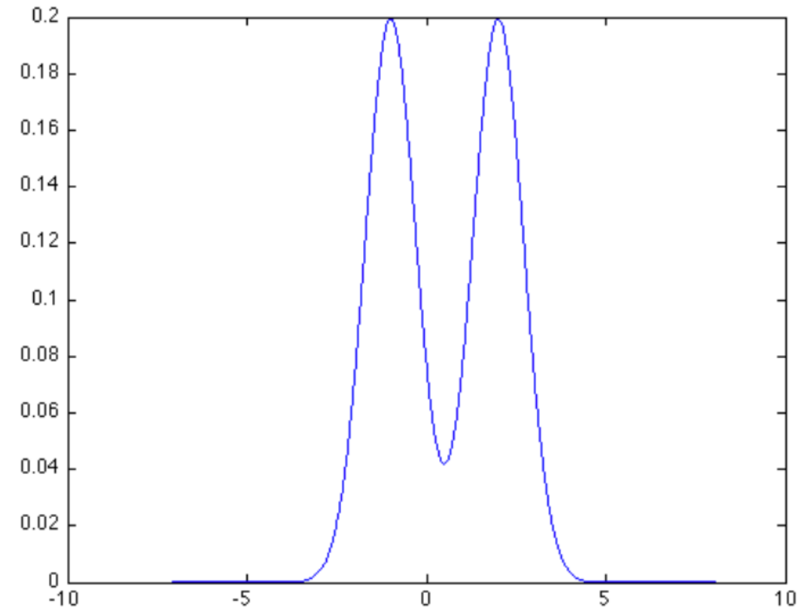


Points x – dots
Label y – color of dots
Prior p_y – proportion of dots
Mean $\mu^{(y)}$ – center of circle
Variance σ_y^2 – size of circle

Generative Model



Observed Labels



Label. $y \sim \text{Multinomial}(p_1, \dots, p_k)$

Point. $x \sim \mathcal{N}\left(\mu^{(y)}, \sigma_y^2\right)$

Parameters. $\theta = \left\{p_1, \dots, p_k, \mu^{(1)}, \dots, \mu^{(k)}, \sigma_1^2, \dots, \sigma_k^2\right\}$

Example: $P(y = 1) = 0.5, P(y = 2) = 0.5$

$$x|y = 1 \sim N(-1, 1)$$

$$x|y = 2 \sim N(2, 1)$$

We have $x \sim \frac{1}{2} N(-1, 1) + \frac{1}{2} N(2, 1)$

Observed Labels

Label. $y \sim \text{Multinomial}(p_1, \ldots, p_k)$

Point. $x \sim \mathcal{N}\left(\mu^{(y)}, \sigma_y^2\right)$

Parameters. $\theta = \left\{p_1, \ldots, p_k, \mu^{(1)}, \ldots, \mu^{(k)}, \sigma_1^2, \ldots, \sigma_k^2\right\}$

Data. $\mathcal{S}_n = \left\{\left(x^{(1)}, y^{(1)}\right), \left(x^{(2)}, y^{(2)}\right), \ldots, \left(x^{(n)}, y^{(n)}\right)\right\}$

PDF of Spherical Gaussian

$$P(x|y, \theta) = \left(2\pi\sigma_y^2\right)^{-d/2} \exp\left\{-\frac{1}{2\sigma_y^2}\left\|x - \mu^{(y)}\right\|^2\right\}$$

PDF of Model $\quad P(x, y|\theta) = p_y P(x|y, \theta)$

Log Likelihood $\quad \mathcal{L}_n(\theta) = \sum_{(x,y)\in\mathcal{S}_n} \log p_y P(x|y, \theta)$

Observed Labels

Hard Labels (Given).

$$\delta\left(y \middle| x^{(t)}\right) = \begin{cases} 1 & \text{if label } y^{(t)} \text{ equals y,} \\ 0 & \text{otherwise.} \end{cases}$$

Log Likelihood.

$$\begin{aligned} \mathcal{L}_n(\theta) &= \textstyle\sum_{(x,y)\in\mathcal{S}_n} \log p_y P(x|y,\theta) \\ &= \textstyle\sum_{x\in\mathcal{S}_n} \sum_{y=1}^{k} \delta(y|x) \log\left\{p_y P(x|y,\theta)\right\} \\ &= \textstyle\sum_{y=1}^{k} \sum_{x\in\mathcal{S}_n} \delta(y|x) \log\left\{p_y P(x|y,\theta)\right\} \\ &= \textstyle\sum_{y=1}^{k} \sum_{x\in\mathcal{S}_n} \delta(y|x) \log\{P(x|y,\theta)\} + \sum_{y=1}^{k} \sum_{x\in\mathcal{S}_n} \delta(y|x) \log\left(p_y\right) \end{aligned}$$

Observed Labels

Hard Labels (Given).

$$\delta\left(y \middle| x^{(t)}\right) = \begin{cases} 1 & \text{if label } y^{(t)} \text{ equals y,} \\ 0 & \text{otherwise.} \end{cases}$$

Maximum Likelihood Estimate.

$\hat{n}_y = \sum_{x \in \mathcal{S}_n} \delta(y|x)$ (number of points with label y)

$\hat{p}_y = \hat{n}_y / n$ (fraction of points with label y)

$\hat{\mu}^{(y)} = \frac{1}{\hat{n}_y} \sum_{x \in \mathcal{S}_n} \delta(y|x) x$ (mean of points with label y)

$\hat{\sigma}_y^2 = \frac{1}{d\hat{n}_y} \sum_{x \in \mathcal{S}_n} \delta(y|x) \left\| x - \hat{\mu}^{(y)} \right\|^2$ (variance of points with label y)

Mixture Model (Hidden Labels)

Example:

Model: $P(y = 1) = 0.5, P(y = 2) = 0.5$

$$x|y = 1 \sim N(\mu_1, 1)$$

$$x|y = 2 \sim N(\mu_2, 1)$$

Goal: Given data $x^{(1)}, x^{(2)}, \dots, x^{(n)}$ (but no $y^{(n)}$ observed)

Find maximum likelihood estimates of μ_1 and μ_2

EM basic idea: if $y^{(n)}$ were known ➡ two easy-to-solve separate ML problems

EM iterates over

E-step: For $i = 1, \dots, n$, fill in missing data $y^{(n)}$ according to what is most likely given the current model μ

M-step: run ML for completed data, which gives new model μ

Mixture Model (Hidden Labels) (EM for Mixture of Gaussians)

Label. $y \sim \mathrm{Multinomial}(p_1, \dots, p_k)$

Point. $x \sim \mathcal{N}\left(\mu^{(y)}, \sigma_y^2\right)$

Parameters. $\theta = \left\{p_1, \dots, p_k, \mu^{(1)}, \dots, \mu^{(k)}, \sigma_1^2, \dots, \sigma_k^2\right\}$

Data. $\mathcal{S}_n = \left\{x^{(1)}, x^{(2)}, \dots, x^{(n)}\right\}$

PDF of Spherical Gaussian

$$P(x|y,\theta) = \left(2\pi\sigma_y^2\right)^{-d/2} \exp\left\{-\frac{1}{2\sigma_y^2}\left\|x - \mu^{(y)}\right\|^2\right\}$$

PDF of Model $\quad P(x|\theta) = \sum_{y=1}^{k} p_y P(x|y,\theta)$

Log Likelihood $\quad \mathcal{L}_n(\theta) = \sum_{x \in \mathcal{S}_n} \log \sum_{y=1}^{k} p_y P(x|y,\theta)$

Comparison between hidden labels and observed labels

PDF of Model

Observed Labels $P(x, y|\theta) = p_y P(x|y, \theta)$

Hidden Labels $P(x|\theta) = \sum_{y=1}^{k} p_y P(x|y, \theta)$

Marginalizing over y

Log Likelihood

Observed Labels $\mathcal{L}_n(\theta) = \sum_{(x,y)\in\mathcal{S}_n} \log p_y P(x|y, \theta)$

Hidden Labels $\mathcal{L}_n(\theta) = \sum_{x\in\mathcal{S}_n} \log \sum_{y=1}^{k} p_y P(x|y, \theta)$

Expectation-Maximization

No exact solution!

Log Likelihood.

$$\mathcal{L}_n(\theta) = \sum_{x \in \mathcal{S}_n} \log \sum_{y=1}^{k} p_y P(x|y,\theta)$$

Numerical Algorithm.

1. Initialize parameters $\theta = \{p_1, \ldots, p_k, \mu^{(1)}, \ldots, \mu^{(k)}, \sigma_1^2, \ldots, \sigma_k^2\}$
2. Repeat until convergence:
   a. **E-Step.** Given parameters θ, compute soft labels $p(y|x)$.
   b. **M-Step.** Given soft labels $p(y|x)$, compute parameters θ.

Expectation-Maximization

Initialize Parameters.

$p_y = 1/k$ for all y

$\mu^{(y)}$ centroids from k-means algorithm

$\sigma_y^2 = \sigma^2$ the sample variance, for all y

Expectation Step.

Compute soft labels

$$p(y|x) = \frac{p(y,x)}{p(x)} = \frac{p_y P\left(x \middle| \mu^{(y)}, \sigma_y^2\right)}{\sum_{z=1}^{k} p_z P\left(x \middle| \mu^{(z)}, \sigma_z^2\right)}$$

Expectation-Maximization

Maximization Step.

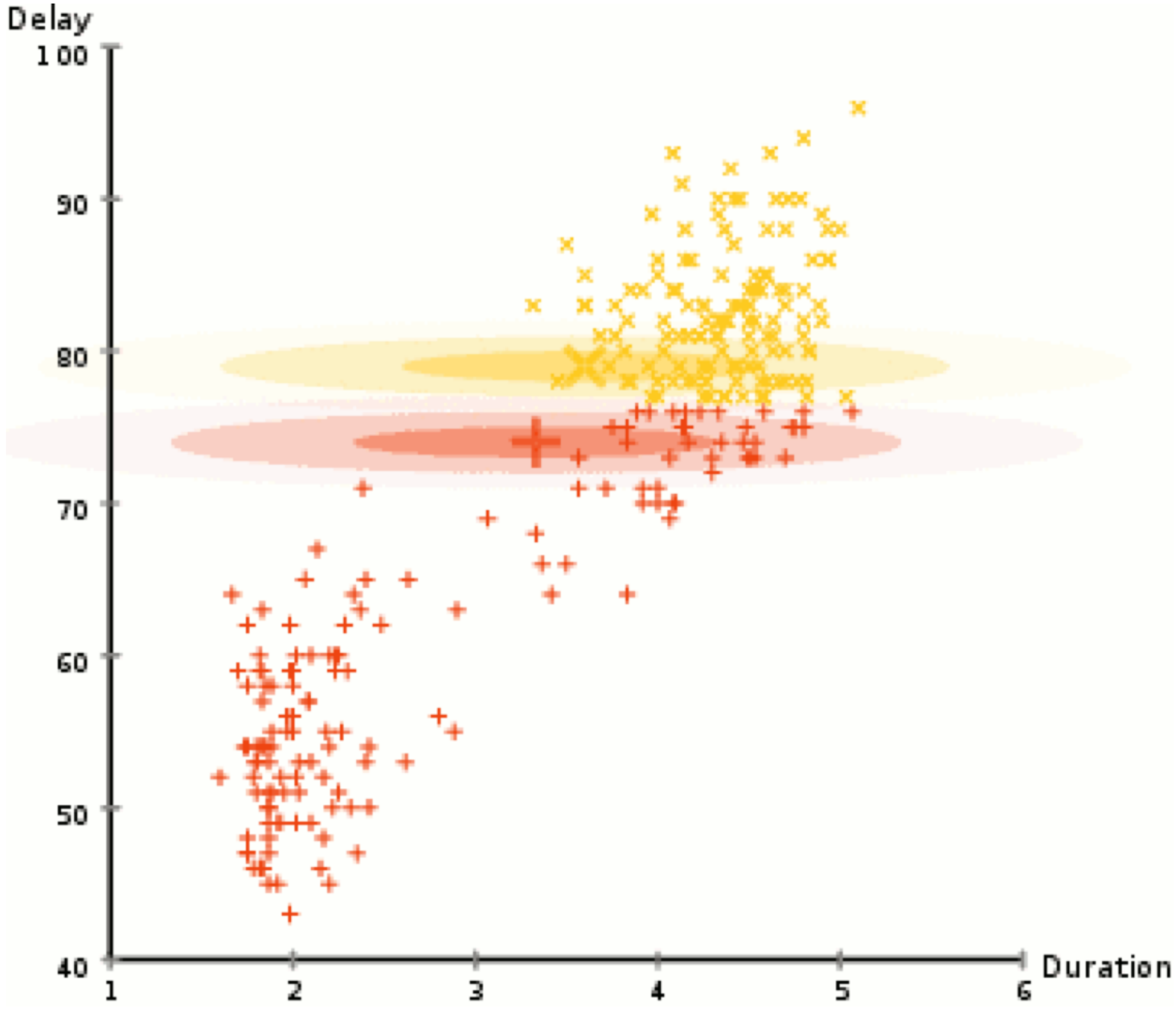
$\hat{n}_y = \sum_{x \in \mathcal{S}_n} p(y|x)$ (effective number of points with label y)

$\hat{p}_y = \hat{n}_y / n$ (effective fraction of points with label y)

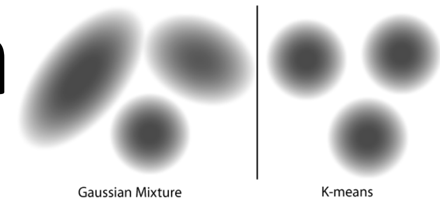
$\hat{\mu}^{(y)} = \frac{1}{\hat{n}_y} \sum_{x \in \mathcal{S}_n} p(y|x) x$ (weighted mean of points with label y)

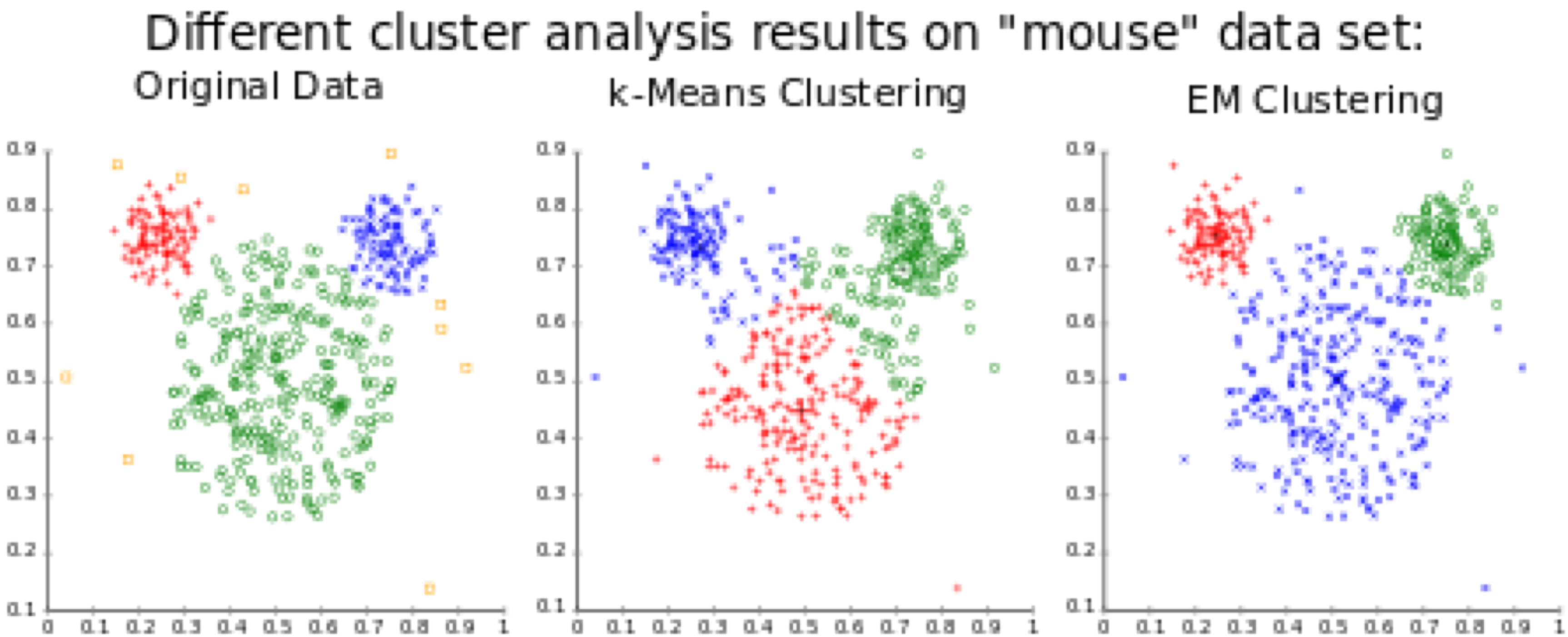
$\hat{\sigma}_y^2 = \frac{1}{d\hat{n}_y} \sum_{x \in \mathcal{S}_n} p(y|x) \left\| x - \hat{\mu}^{(y)} \right\|^2$ (weighted variance of points with label y)

Expectation-Maximization



Comparison with k-Mea





Comparison with K-Means

- Like k-means, EM clustering may get stuck in local minima.
- Unlike k-means, the local minima are more favorable because soft labels allow points to move between clusters slowly.

Not in syllabus

Smoothing

Problem. (when the number of training data is small)

- We want to maximize

$$\mathcal{L}_n(\theta) = \sum_{x \in \mathcal{S}_n} \log\left\{ \sum_{y=1}^{k} p_y \left(2\pi\sigma_y^2\right)^{-d/2} \exp\left(-\frac{1}{2\sigma_y^2} \left\|x - \mu^{(y)}\right\|^2\right) \right\}$$

- Let $\mu^{(1)} = x^{(1)}$ be equal to a data point.
- Term in inner sum becomes $\left(2\pi\sigma_y^2\right)^{-d/2} \exp(0)$.
- As σ_y tends to zero, $\mathcal{L}_n(\theta)$ will tend to infinity!
- In fact, if $x^{(1)}$ is the only point with soft label $p(1|x) \neq 0$, then

$$\hat{\sigma}_1^2 = \frac{1}{d\hat{n}_1} \sum_{x \in \mathcal{S}_n} p(1|x) \left\|x - \hat{\mu}^{(1)}\right\|^2 = 0.$$

Not in syllabus

Smoothing

These are called *conjugate priors*, designed to ensure that prior and posterior have the same form.

Solution.

- Give prior probabilities to σ_y.

$$p\left(\sigma_y^2 \middle| \alpha_y, s_y^2\right) = C\left(2\pi\sigma_y^2\right)^{-\alpha_y d/2} \exp\left(-\frac{\alpha_y s_y^2}{2\sigma_y^2}\right)$$

- New objective is to maximize the log posterior probability.

$$\mathcal{L}_n(\theta) = \sum_{x\in\mathcal{S}_n} \log\left\{ \sum_{y=1}^{k} p_y P\left(x \middle| \mu^{(y)}, \sigma_y^2\right) p\left(\sigma_y^2 \middle| \alpha_y, s_y^2\right) \right\}$$

- New maximization step for $\hat{\sigma}_y^2$ is given by

$$\hat{\sigma}_y^2 = \frac{1}{d(\alpha_y + \hat{n}_y)}\left(\alpha_y s_y^2 + \sum_{x\in\mathcal{S}_n} p(y|x)\left\|x - \hat{\mu}^{(y)}\right\|^2\right).$$

Not in syllabus

Smoothing

Why do we choose prior probabilities of this form?

$$p\left(\sigma_y^2 \middle| \alpha_y, s_y^2\right) = C\ \left(2\pi\sigma_y^2\right)^{-\alpha_y d/2} \exp\left(-\frac{\alpha_y s_y^2}{2\sigma_y^2}\right)$$

- Fix mean μ_y. Suppose we have α_y observations of $s_y + \mu_y$. The likelihood of these observations is

$$p\left(\alpha_y, s_y^2 \middle| \sigma_y^2\right) = \left(2\pi\sigma_y^2\right)^{-\alpha_y d/2} \exp\left(-\frac{\alpha_y s_y^2}{2\sigma_y^2}\right).$$

- The posterior probability of σ_y^2 will be

$$p\left(\sigma_y^2 \middle| \alpha_y, s_y^2\right) \propto p\left(\alpha_y, s_y^2 \middle| \sigma_y^2\right) p\left(\sigma_y^2\right).$$

- Use this posterior as a *prior* for maximum likelihood estimation.

Model Selection

- By setting $p_{k+1} = 0$, we see that (mixture model with k clusters) contained in (mixture model with $k+1$ clusters).
- Therefore, likelihood for (mixture model with $k+1$ clusters) is greater or equal to that of (mixture model with k clusters).
- How to choose the right k and prevent over-/under-fitting?

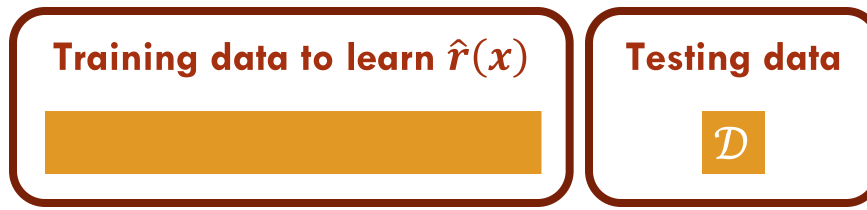
Validation vs Cross-Validation

Method 1 (Simulation)

Estimate testing error using simple validation or cross-validation.

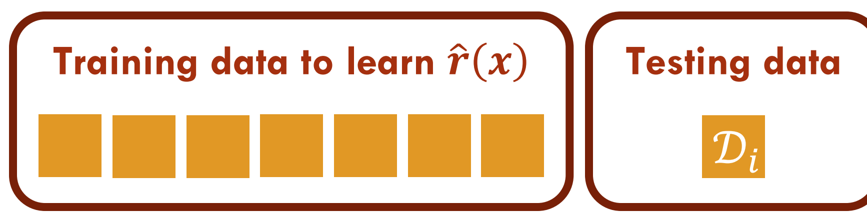
testing error

- $\hat{R}(\mathcal{D})$



***k*-fold cross-validation**.

- $\hat{R}_{\text{CV}} = \frac{1}{k}\sum_{i=1}^{k} \hat{R}(\mathcal{D}_i)$



Bayesian Information Criterion

Method 2 (Marginal Likelihood)

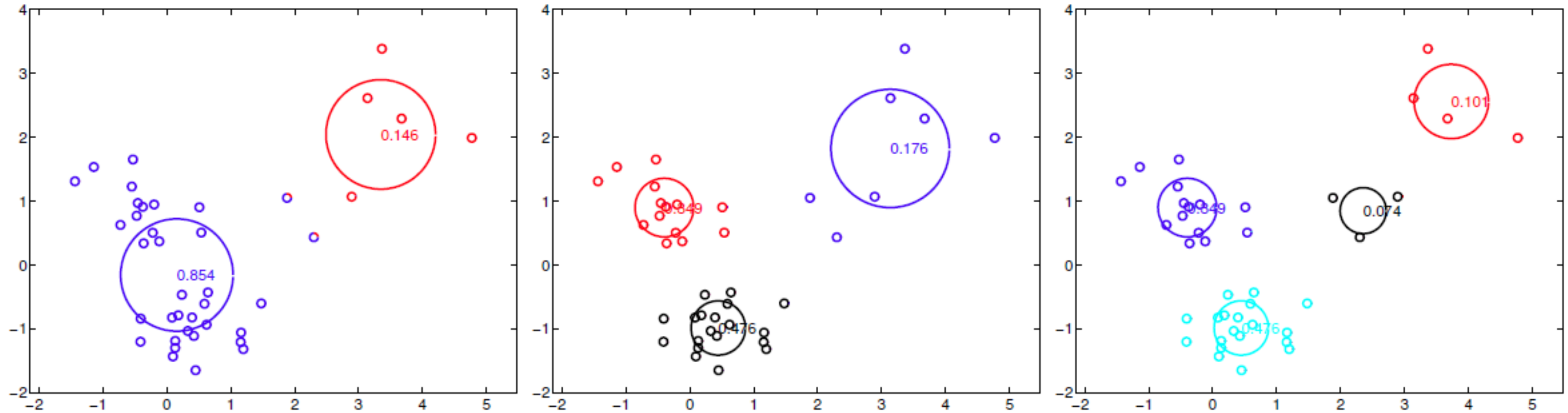
Maximize the marginal likelihood integral. But computing this integral is tedious, so we approximate it using the BIC.

$$\mathrm{BIC}(\theta) = \mathcal{L}_n(\theta) - \frac{\#\text{ of free params}}{2} \log n$$

For Gaussian mixtures, we have $k(d+2)-1$ free parameters.

$$\mathrm{BIC}(\theta) \qquad = \mathcal{L}_n(\theta) - \frac{k(d+2)-1}{2} \log n$$

Bayesian Information Criterion



$$l(D;\hat{\theta}) = -118.25 \qquad l(D;\hat{\theta}) = -98.64 \qquad l(D;\hat{\theta}) = -94.11$$

$$BIC(D;\hat{\theta}) = -131.16 \qquad BIC(D;\hat{\theta}) = -118.93 \qquad BIC(D;\hat{\theta}) = -121.78$$

Summary

- Expectation-Maximization
  - Mixture Model
  - Clustering
  - Hidden Variables
  - Soft Labels
- Generalization
  - Priors and Smoothing
  - Model Selection
  - Validation and Cross-Validation
  - Bayesian Information Criterion

Intended Learning Outcomes

Expectation-Maximization

- Write down the distribution of a Gaussian mixture model. Write down the log likelihood of a given data set.
- Describe the expectation-maximization algorithm. In particular, describe how the parameters may be initialized effectively, and describe how the soft labels are computed in the E-step, and describe how the parameters are updated in the M-step.
- Explain how the EM algorithm may be used in clustering, and describe the differences between k-means and EM clustering.
- Explain how prior probabilities on the variances σ_y^2 may be used to obtain smoothed estimates for the parameters.

Intended Learning Outcomes

Model Selection

- List some strategies for selecting the number of clusters.
- Describe the differences between validation and cross-validation.
- Write down the Bayesian Information Criterion, and explain how it may be used for model selection.