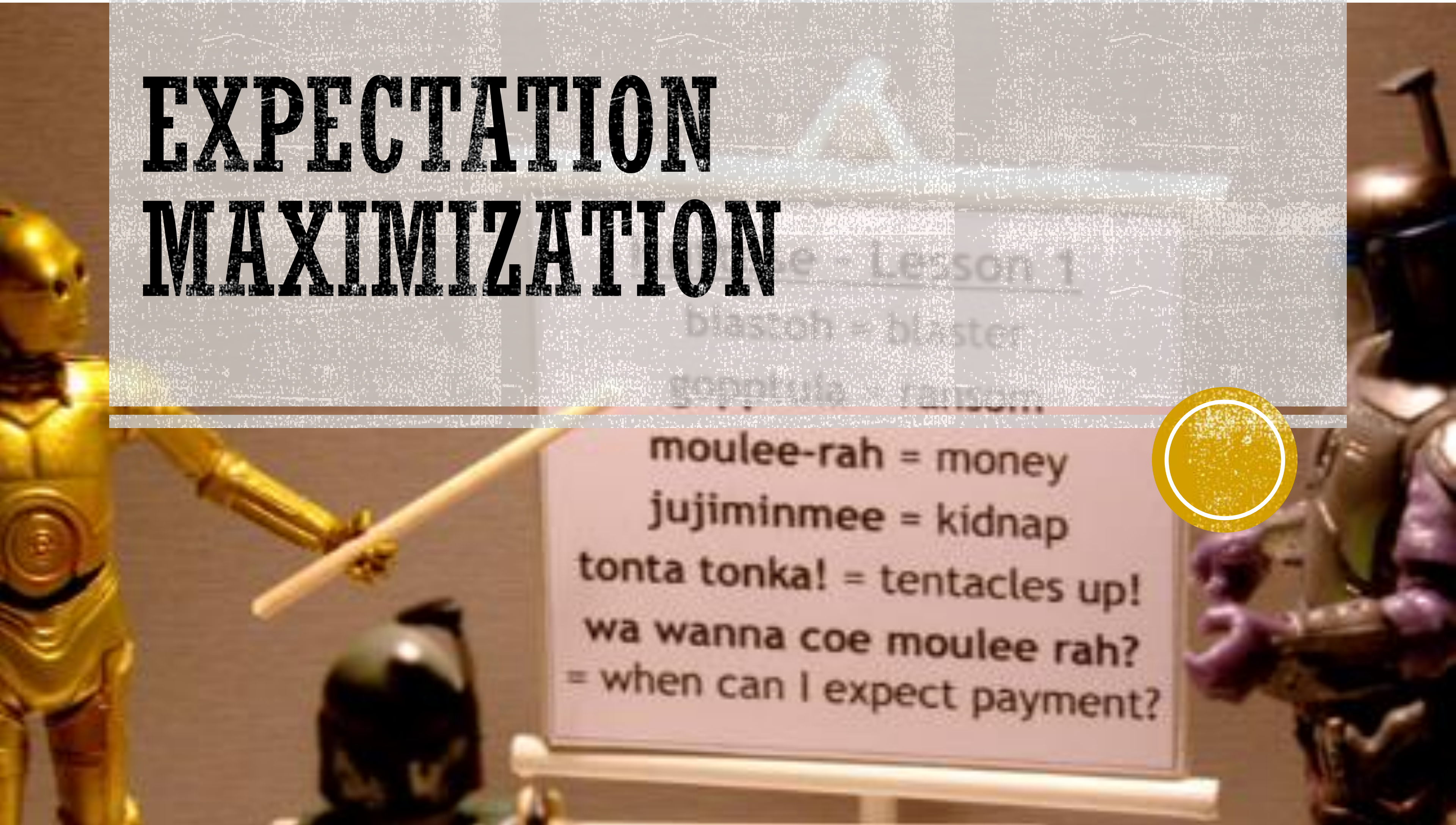
EXPECTATION MAXIMIZATION



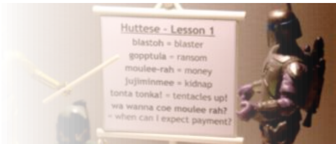
BACK TO CLUSTERING

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



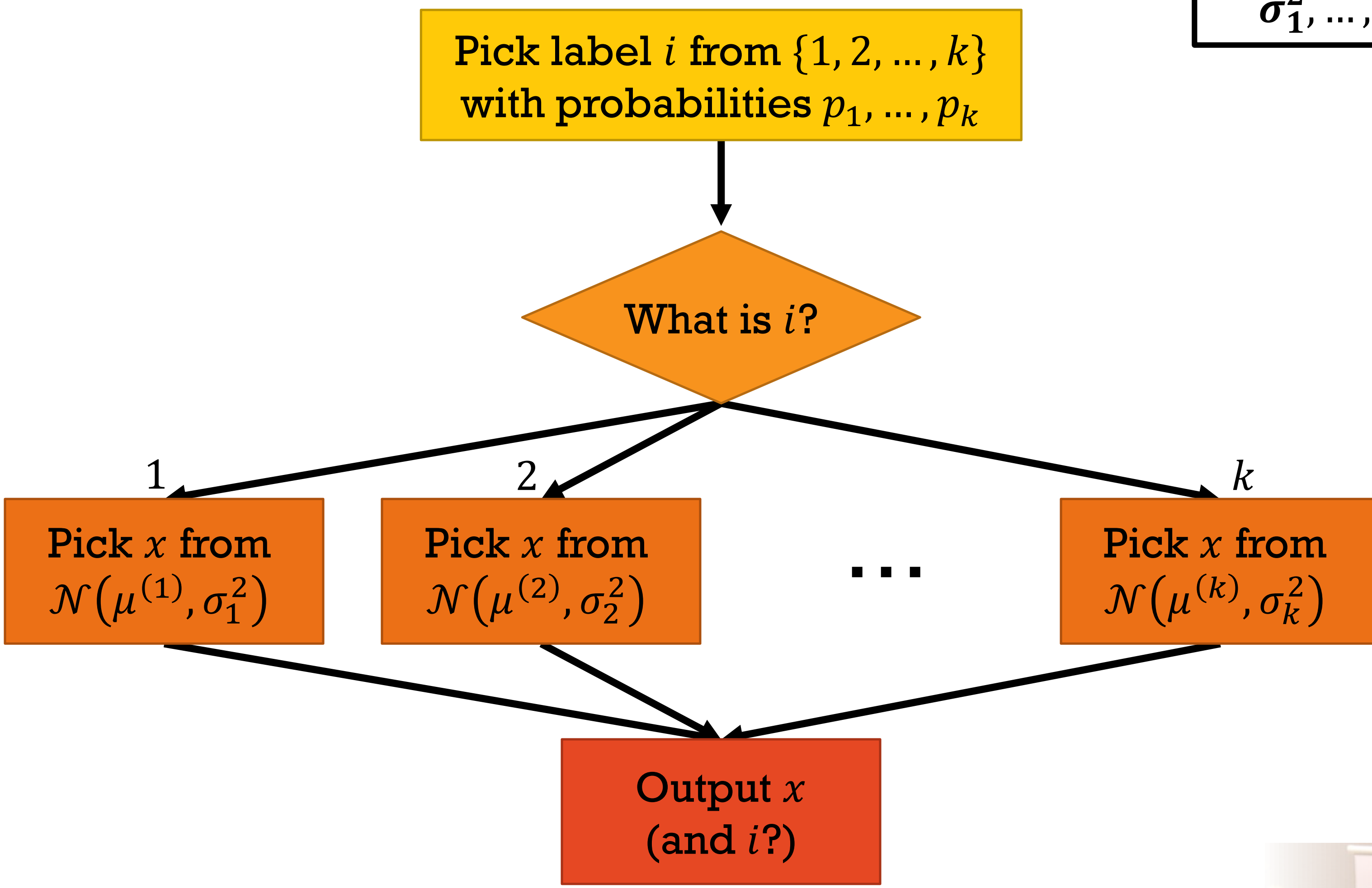
MIXTURE MODELS

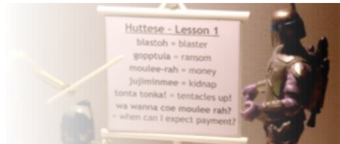
Model Parameters

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

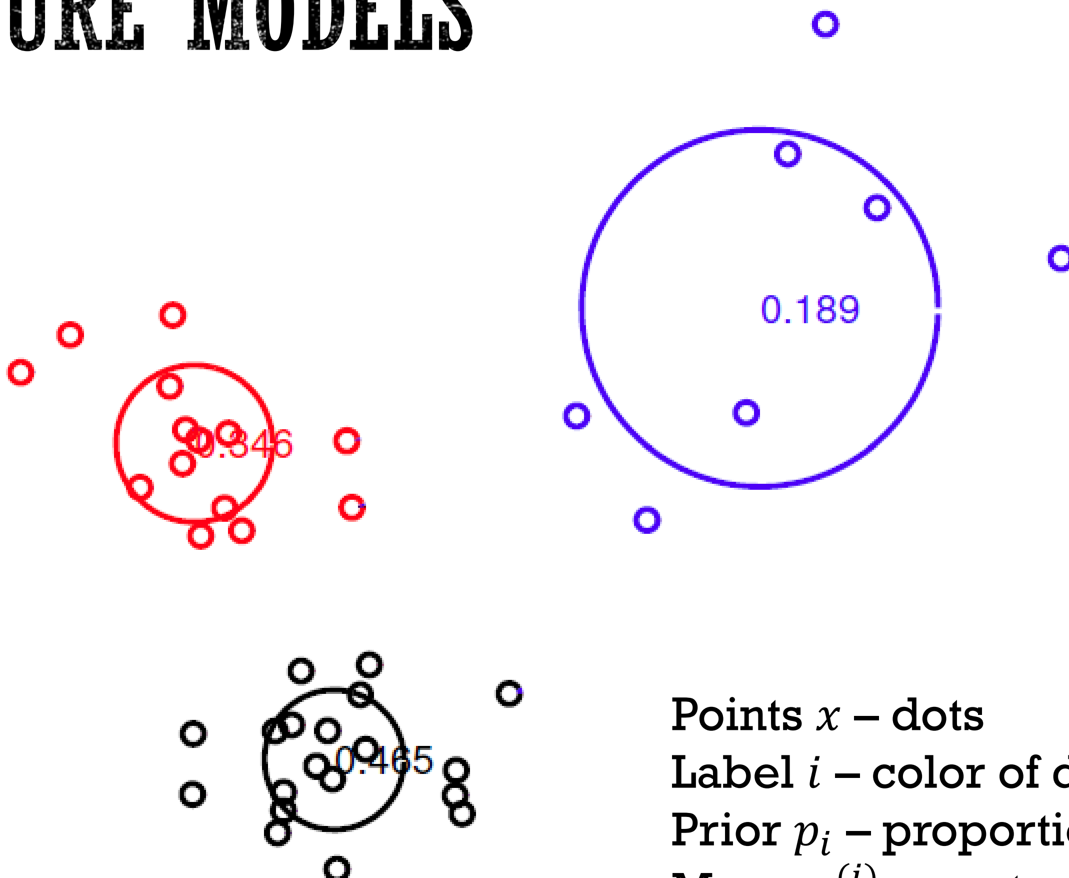
$\boldsymbol{\mu^{(1)}, \dots, \mu^{(k)}}$

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

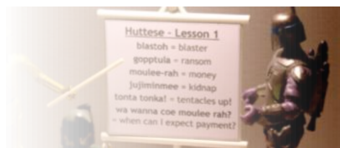




MIXTURE MODELS



Points x – dots
Label i – color of dots
Prior p_i – proportion of dots
Mean $\mu^{(i)}$ – center of circle
Variance σ_i – size of circle



MIXTURE MODELS

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

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

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

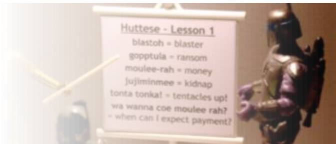
Data. $\mathcal{D} = \left\{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\right\}$

PDF of spherical Gaussian

$$P\left(x \middle| \mu^{(i)}, \sigma_i^2\right) = \left(2\pi\sigma_i^2\right)^{-d/2} \exp\left\{-\frac{1}{2\sigma_i^2}\left\|x - \mu^{(i)}\right\|^2\right\}$$

PDF of mixture model

$$P(x|\theta) = \sum_{i=1}^{k} p_i\, P\left(x \middle| \mu^{(i)}, \sigma_i\right)$$



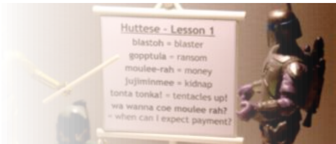
OBSERVED LABELS

Hard Labels (Given).

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

Log Likelihood.

$$\begin{aligned} \ell(\theta) &= \sum_{x \in \mathcal{D}} \sum_{i=1}^{k} \delta(i|x) \log\left\{p_i P\left(x \middle| \mu^{(i)}, \sigma_i^2\right)\right\} \\ &= \sum_{i=1}^{k} \sum_{x \in \mathcal{D}} \delta(i|x) \log\left\{p_i P\left(x \middle| \mu^{(i)}, \sigma_i^2\right)\right\} \\ &= \sum_{i=1}^{k} \sum_{x \in \mathcal{D}} \delta(i|x) \log\left\{P\left(x \middle| \mu^{(i)}, \sigma_i^2\right)\right\} + \sum_{i=1}^{k} \sum_{x \in \mathcal{D}} \delta(i|x) \log(p_i) \end{aligned}$$



OBSERVED LABELS

Hard Labels (Given).

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

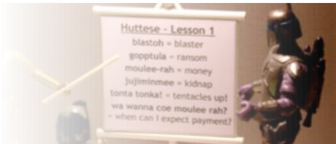
Maximum Likelihood Estimate.

$\hat{n}_i = \sum_{x \in \mathcal{D}} \delta(i|x)$ (number of points with label i)

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

$\hat{\mu}^{(i)} = \frac{1}{\hat{n}_i} \sum_{x \in \mathcal{D}} \delta(i|x) x$ (mean of points with label i)

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



HIDDEN LABELS

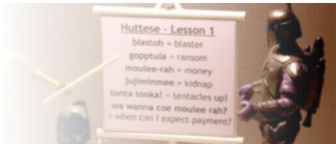
No exact solution!

Log Likelihood.

$$\ell(\theta) = \sum_{x \in \mathcal{D}} \log\left\{ \sum_{i=1}^{k} p_i P\left(x \middle| \mu^{(i)}, \sigma_i^2\right) \right\}$$

Numerical Algorithm.

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



EXPECTATION-MAXIMIZATION

Initialize Parameters.

$p_i = 1/k$ for all i

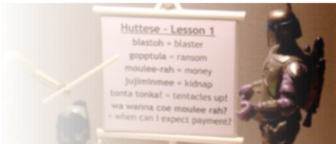
$\mu^{(i)}$ centroids from k-means algorithm

$\sigma_i^2 = \sigma^2$ the sample variance, for all i

Expectation Step.

Compute soft labels

$$p(i|x) = \frac{p(i,x)}{p(x)} = \frac{p_i P\left(x \middle| \mu^{(i)}, \sigma_i^2\right)}{\sum_{j=1}^{k} p_j P\left(x \middle| \mu^{(j)}, \sigma_j^2\right)}$$



EXPECTATION-MAXIMIZATION

Maximization Step.

$\hat{n}_i = \sum_{x \in \mathcal{D}} p(i|x)$ (effective number of points with label i)

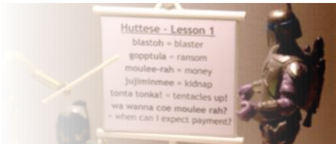
$\hat{p}_i = \hat{n}_i / n$ (effective fraction of points with label i)

$\hat{\mu}^{(i)} = \frac{1}{\hat{n}_i} \sum_{x \in \mathcal{D}} p(i|x)x$ (weighted mean of points with label i)

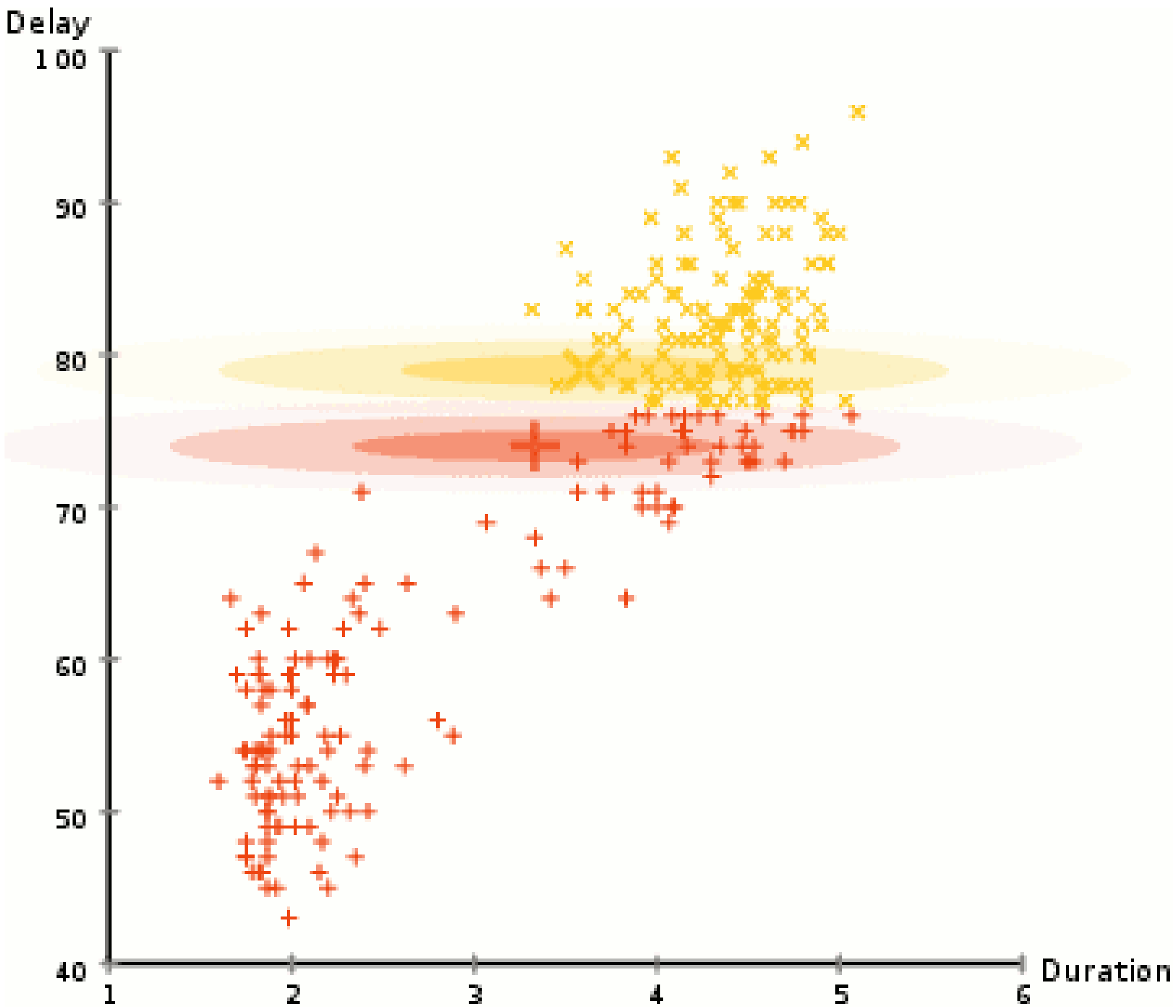
$\hat{\sigma}_i^2 = \frac{1}{d\hat{n}_i} \sum_{x \in \mathcal{D}} p(i|x) \left\| x - \hat{\mu}^{(i)} \right\|^2$ (weighted variance of points with label i)

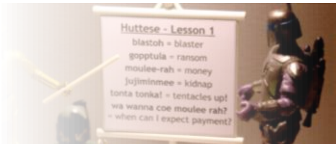
Caveat.

- Like k-means, it 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.

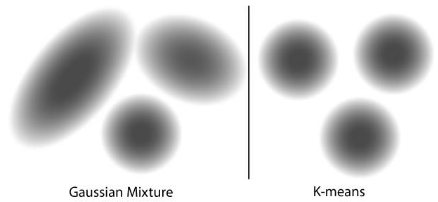


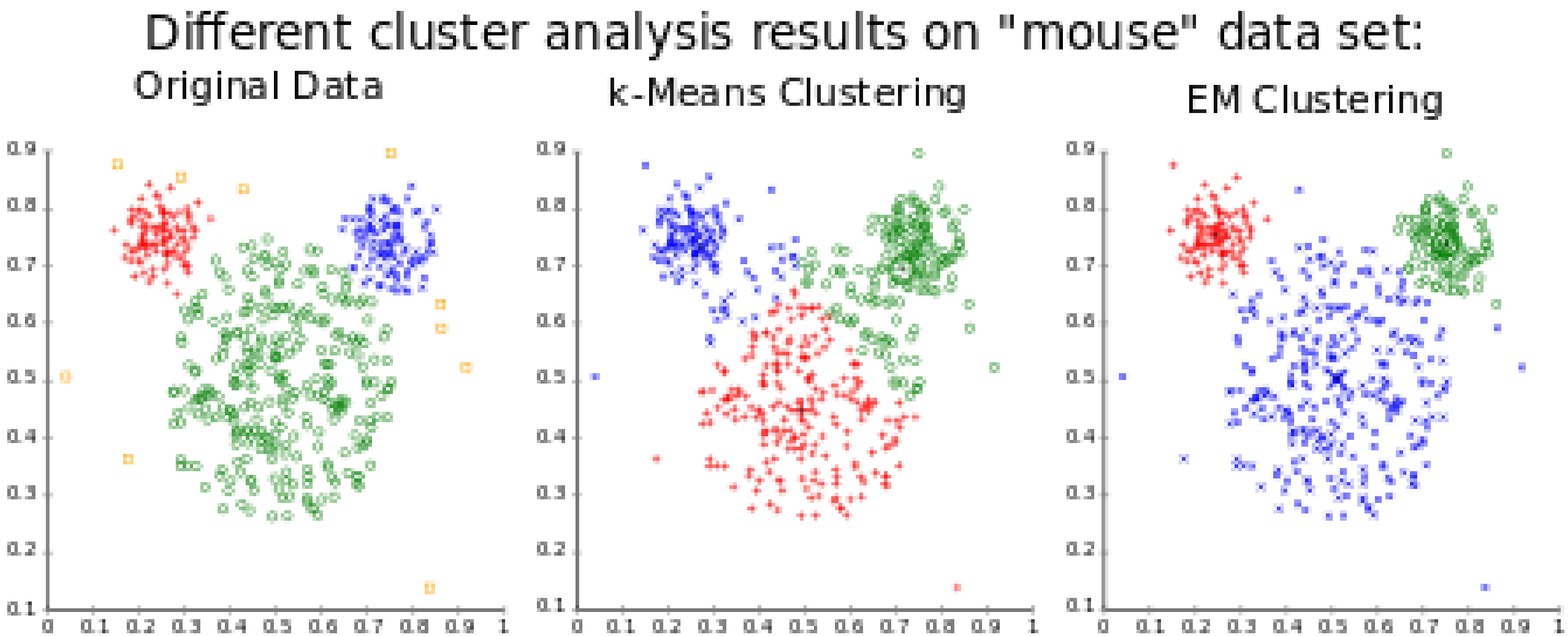
EXPECTATION-MAXIMIZATION

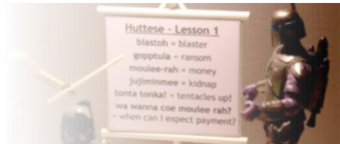




COMPARISON WITH K-MEANS







SMOOTHING

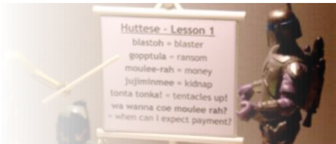
Problem.

- We want to maximize

$$\ell(\theta) = \sum_{x \in \mathcal{D}} \log \left\{ \sum_{i=1}^{k} p_i \left(2\pi\sigma_i^2\right)^{-d/2} \exp\left(-\frac{1}{2\sigma_i^2} \left\| x - \mu^{(i)} \right\|^2\right) \right\}$$

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

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



SMOOTHING

Solution.

- Give prior probabilities to the σ_i.

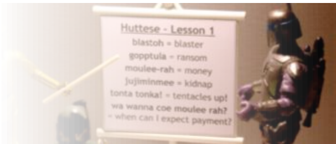
$$p\left(\sigma_i^2 \middle| \alpha_i, s_i^2\right) = C\left(2\pi\sigma_i^2\right)^{-\alpha_i d/2} \exp\left(-\frac{\alpha_i s_i^2}{2\sigma_i^2}\right)$$

- New objective is to maximize the log posterior probability.

$$\ell(\theta) = \sum_{x\in\mathcal{D}} \log\left\{ \sum_{i=1}^{k} p_i P\left(x \middle| \mu^{(i)}, \sigma_i^2\right) p\left(\sigma_i^2 \middle| \alpha_i, s_i^2\right) \right\}$$

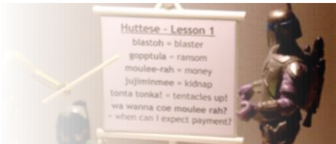
- New maximization step for $\hat{\sigma}_i^2$ is given by

$$\hat{\sigma}_i^2 = \frac{1}{d(\alpha_i + \hat{n}_i)}\left(\alpha_i s_i^2 + \sum_{x\in\mathcal{D}} p(i|x)\left\|x - \hat{\mu}^{(i)}\right\|^2\right).$$



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)

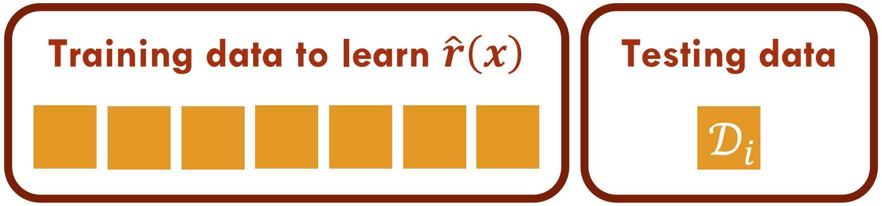
Estimate testing error using validation or cross-validation.

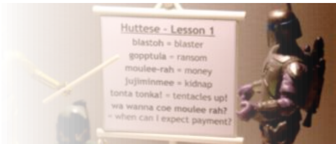
testing error

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

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

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





BAYESIAN INFORMATION CRITERION

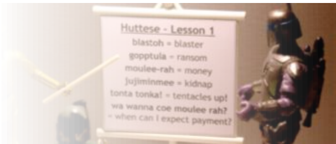
Method 2 (Marginal Likelihood)

Use marginal likelihood integral to select model. But computing this integral is tedious, so we approximate it using the BIC.

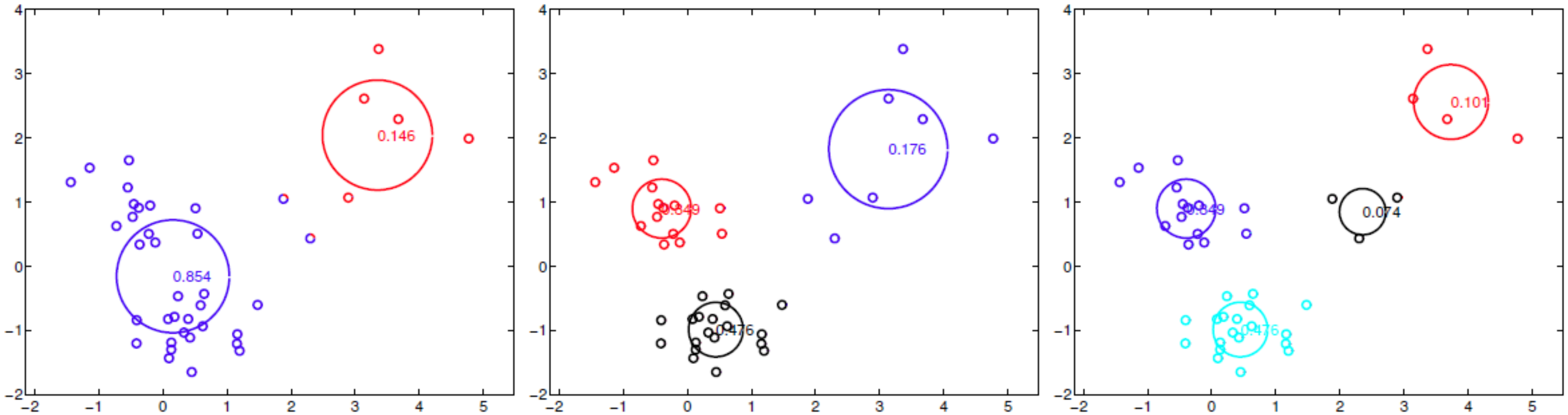
$$\mathrm{BIC}(\theta) = \ell(\theta) - \frac{\#\text{ of free params}}{2} \log n$$

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

$$\mathrm{BIC}(\theta) = \ell(\theta) - \frac{k(d+2)-1}{2} \log n$$

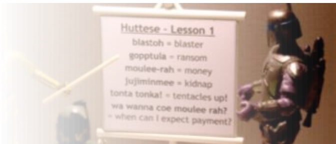


BAYESIAN INFORMATION CRITERION



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

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



SUMMARY

- Generative Models
  - Multinomial and Bag-of-Words
  - Multivariate and Spherical Gaussian
  - Independent and Identically Distributed
  - Maximum Likelihood Estimate
  - Stochastic Gradient Descent
  - Log Likelihood Ratio
- Expectation-Maximization
  - Mixture Model
  - Clustering
  - Hidden Variables
  - Soft Labels
- Generalization
  - Smoothing and Pseudo-Counts
  - Model Selection
  - Validation and Cross-Validation
  - Bayesian Information Criterion