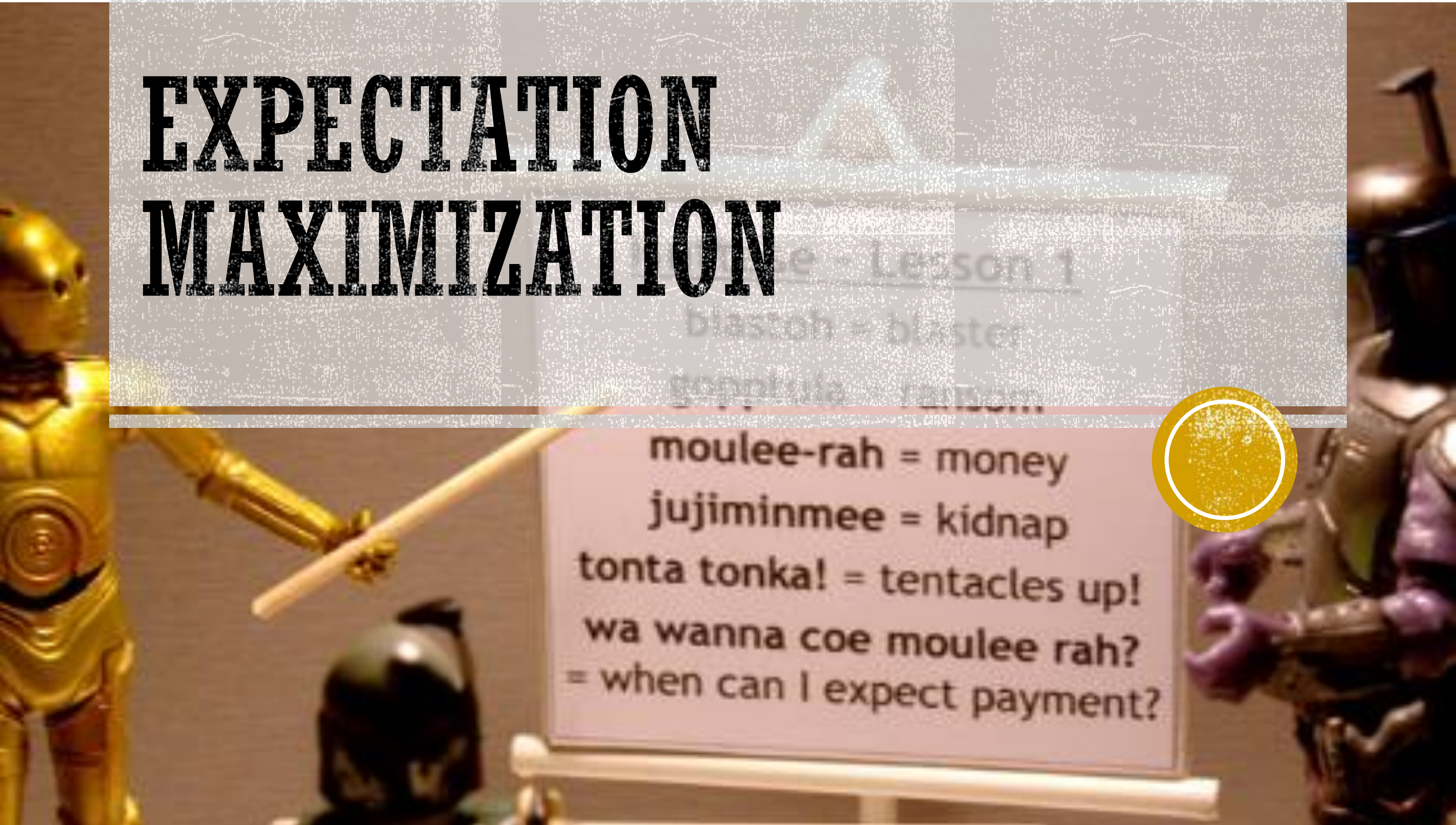
EXPECTATION MAXIMIZATION



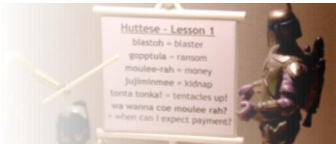
BACK TO CLUSTERING

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



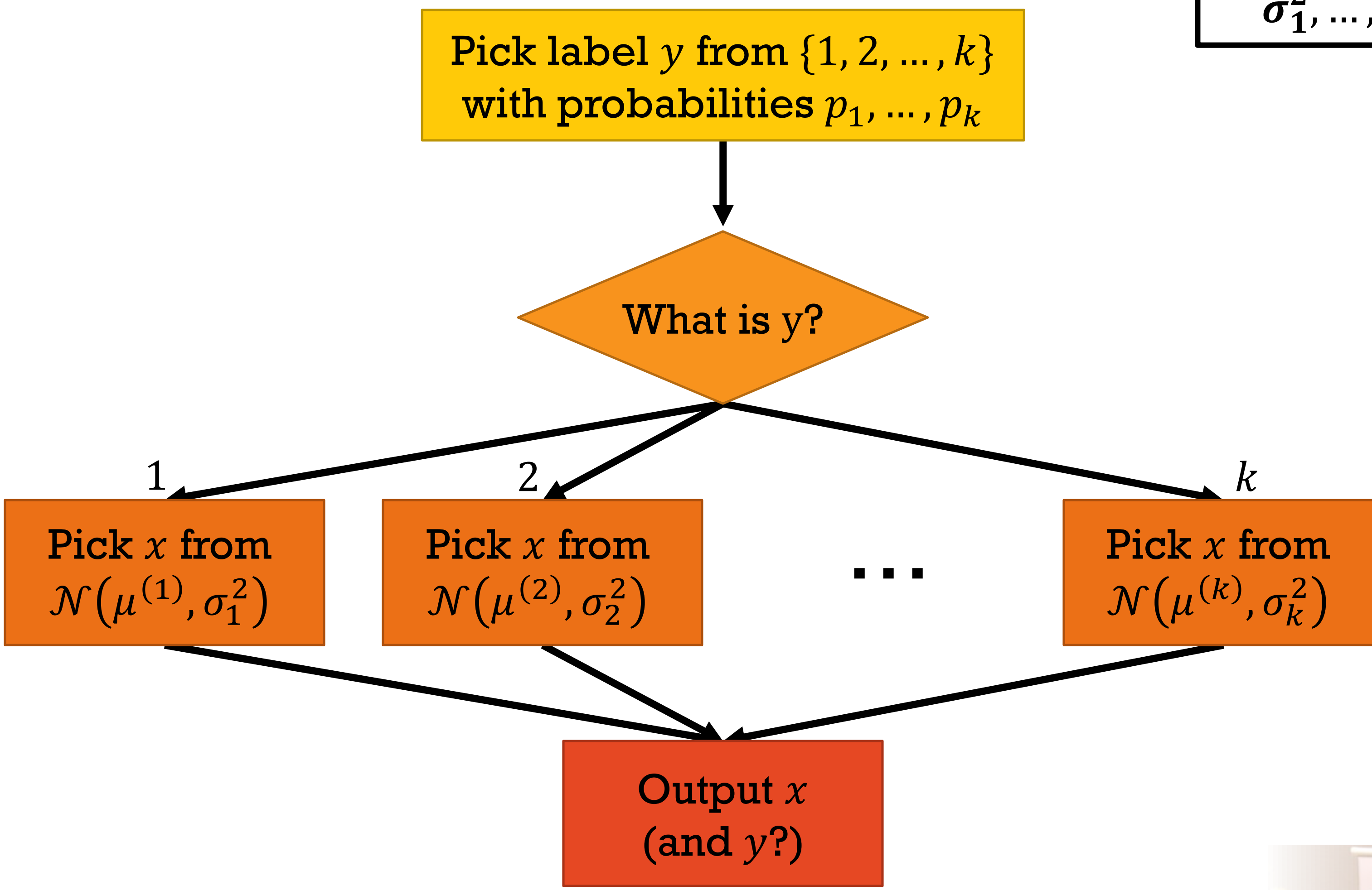
GENERATIVE MODEL

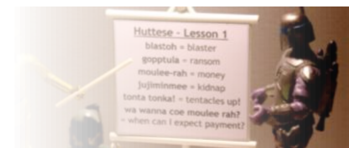
Model Parameters

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

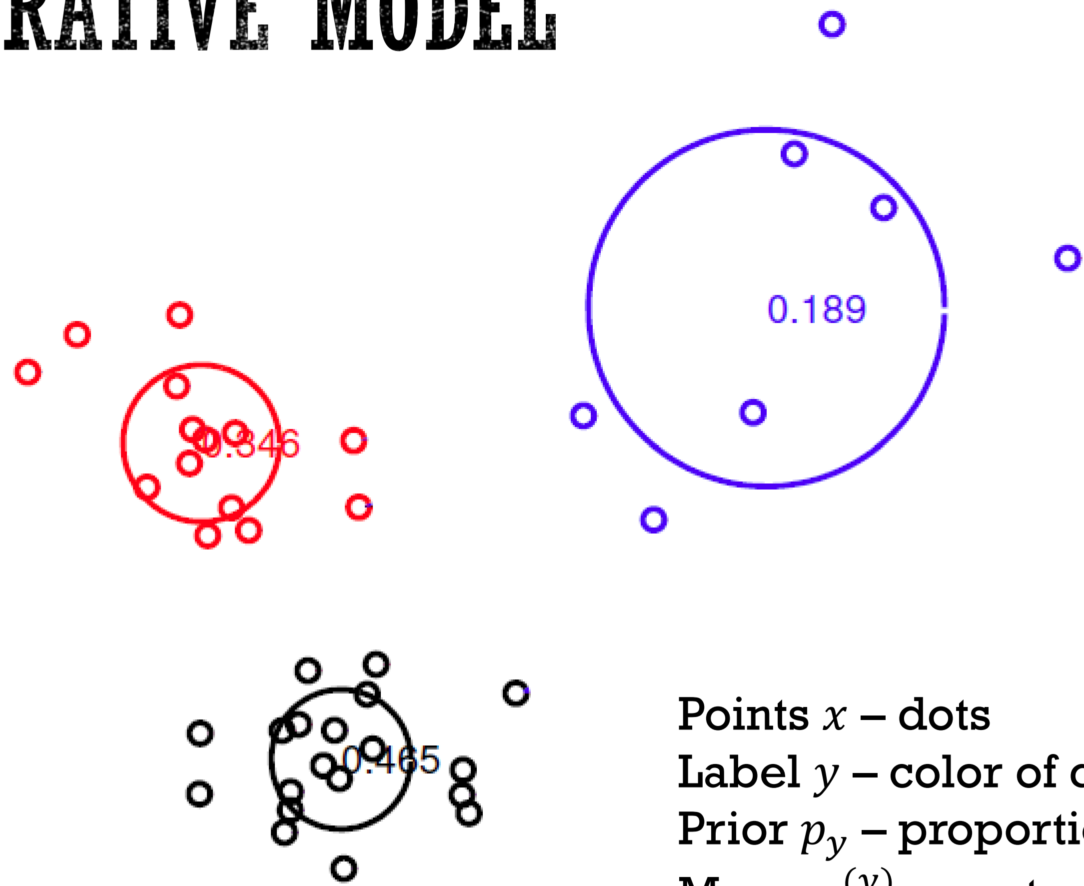
$\boldsymbol{\mu^{(1)}, \ldots, \mu^{(k)}}$

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

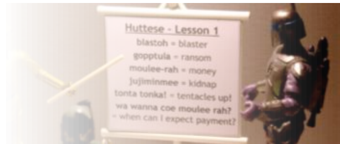




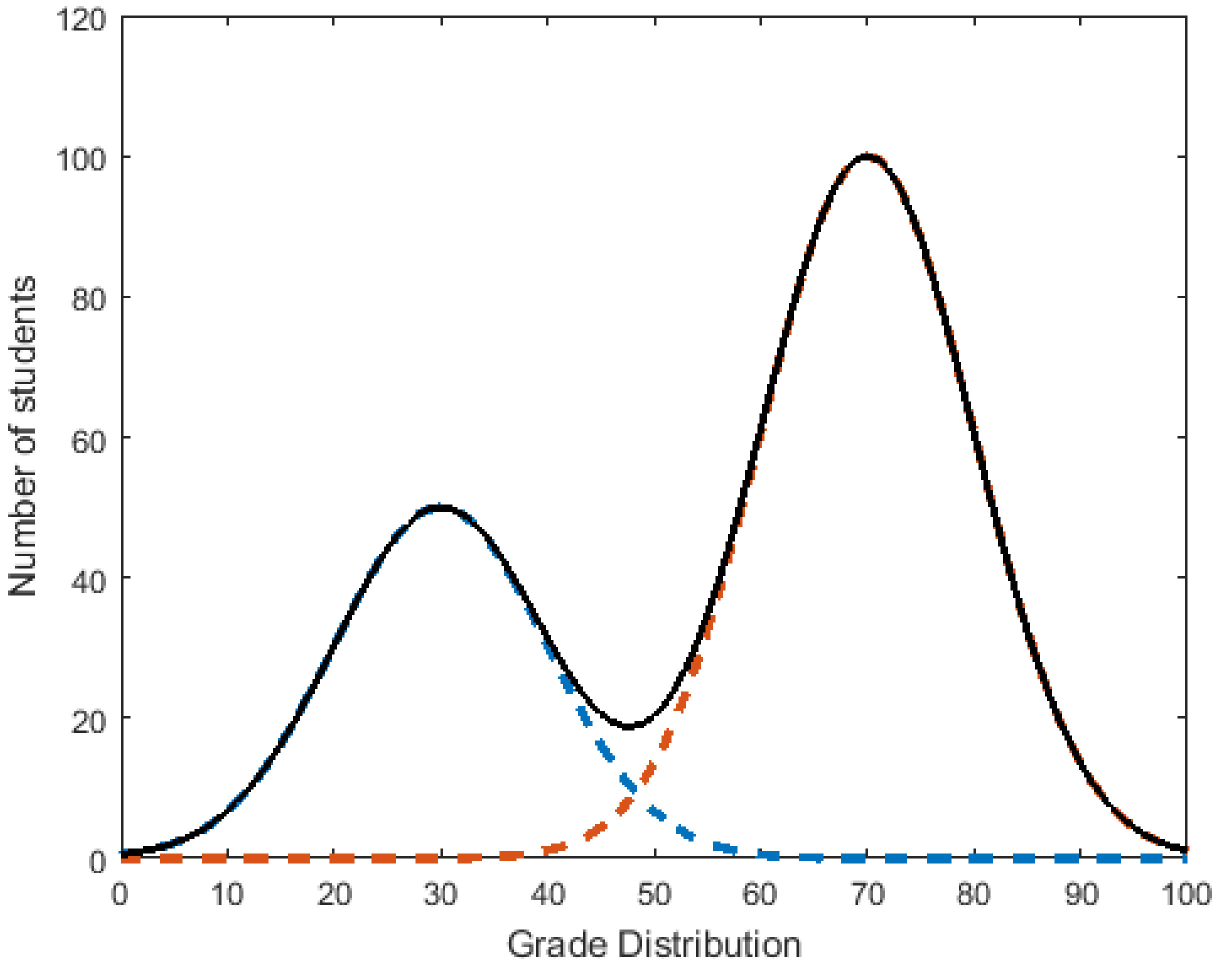
GENERATIVE MODEL

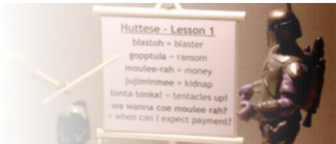


Points x – dots
Label y – color of dots
Prior p_y – proportion of dots
Mean $\mu^{(y)}$ – center of circle
Variance σ_y^2 – size of circle



GENERATIVE MODEL





OBSERVED LABELS

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

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

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

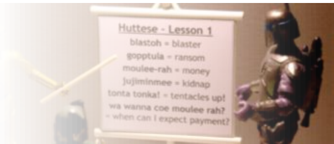
Data. $\mathcal{S}_n = \left\{\left(x^{(1)}, y^{(1)}\right), \left(x^{(2)}, y^{(2)}\right), \ldots, \left(x^{(n)}, y^{(n)}\right)\right\}$

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



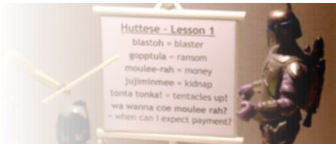
OBSERVED LABELS

Hard Labels (Given).

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

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).

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

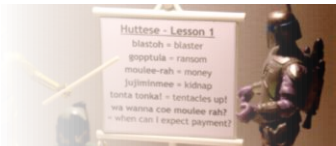
Maximum Likelihood Estimate.

$\hat{n}_y = \sum_{x \in \mathcal{S}_n} \delta(y|x)$ (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)

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

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

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

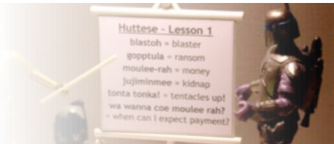
Data. $\mathcal{S}_n = \left\{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\right\}$

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



MIXTURE MODEL (HIDDEN LABELS)

PDF of Model

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

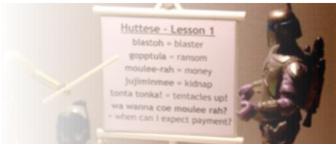
Hidden Labels $\quad P(x|\theta) \quad = \sum_{y=1}^{k} p_y P(x|y, \theta)$

Marginalizing over y

Log Likelihood

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

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



EXPECTATION-MAXIMIZATION

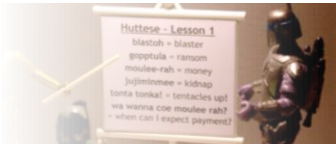
Log Likelihood.

No exact solution!

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

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

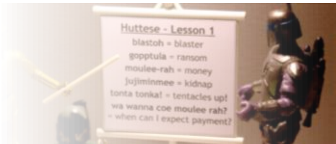
$\mu^{(y)}$ centroids from k-means algorithm

$\sigma_y^2 = \sigma^2$ 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

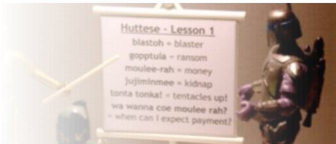
Maximization Step.

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

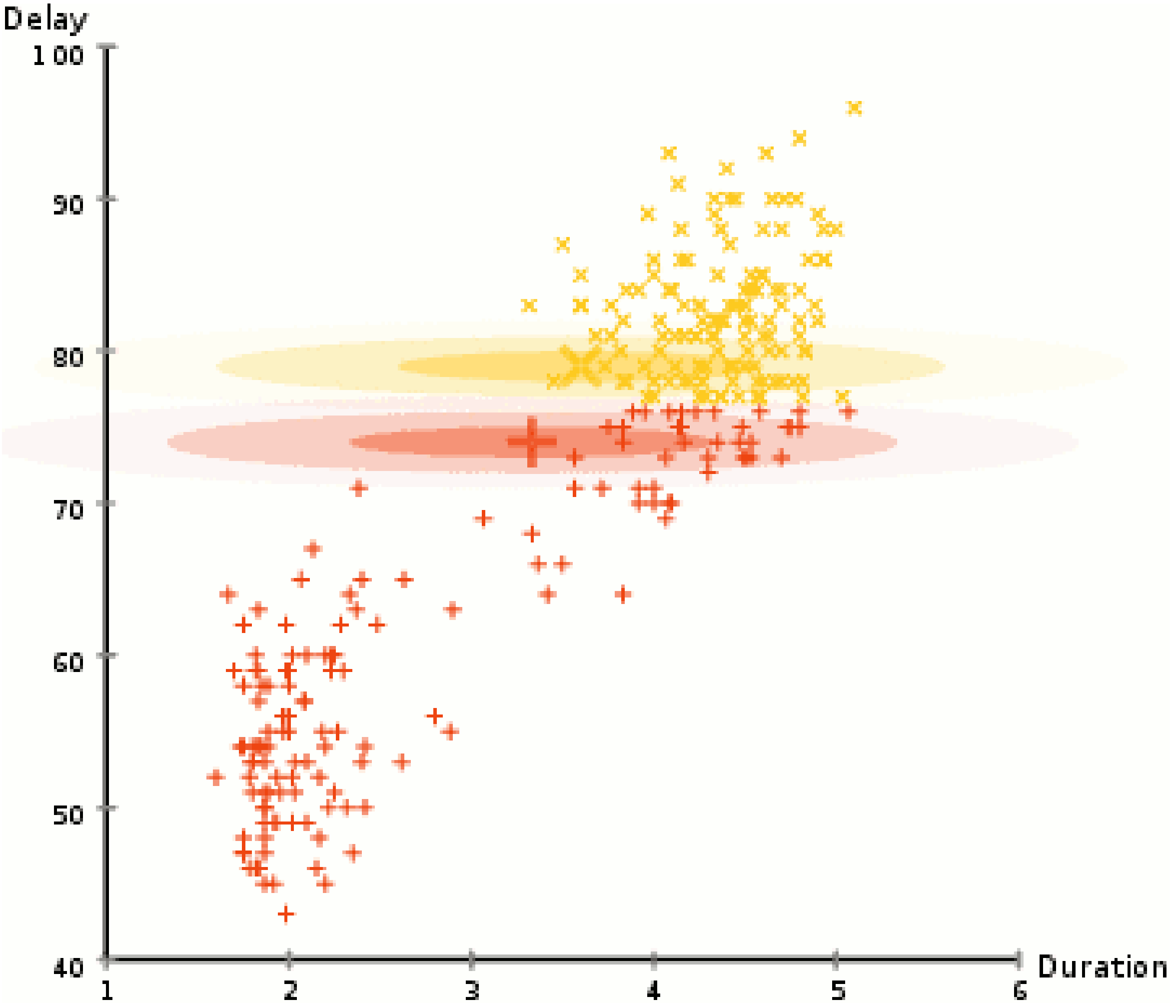
$\hat{p}_y = \hat{n}_y / n$ (effective fraction of points with label y)

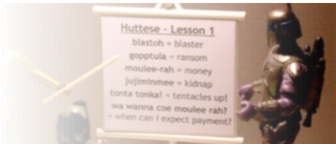
$\hat{\mu}^{(y)} = \frac{1}{\hat{n}_y} \sum_{x \in \mathcal{S}_n} p(y|x)x$ (weighted mean of points with label y)

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

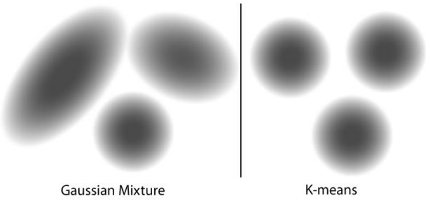


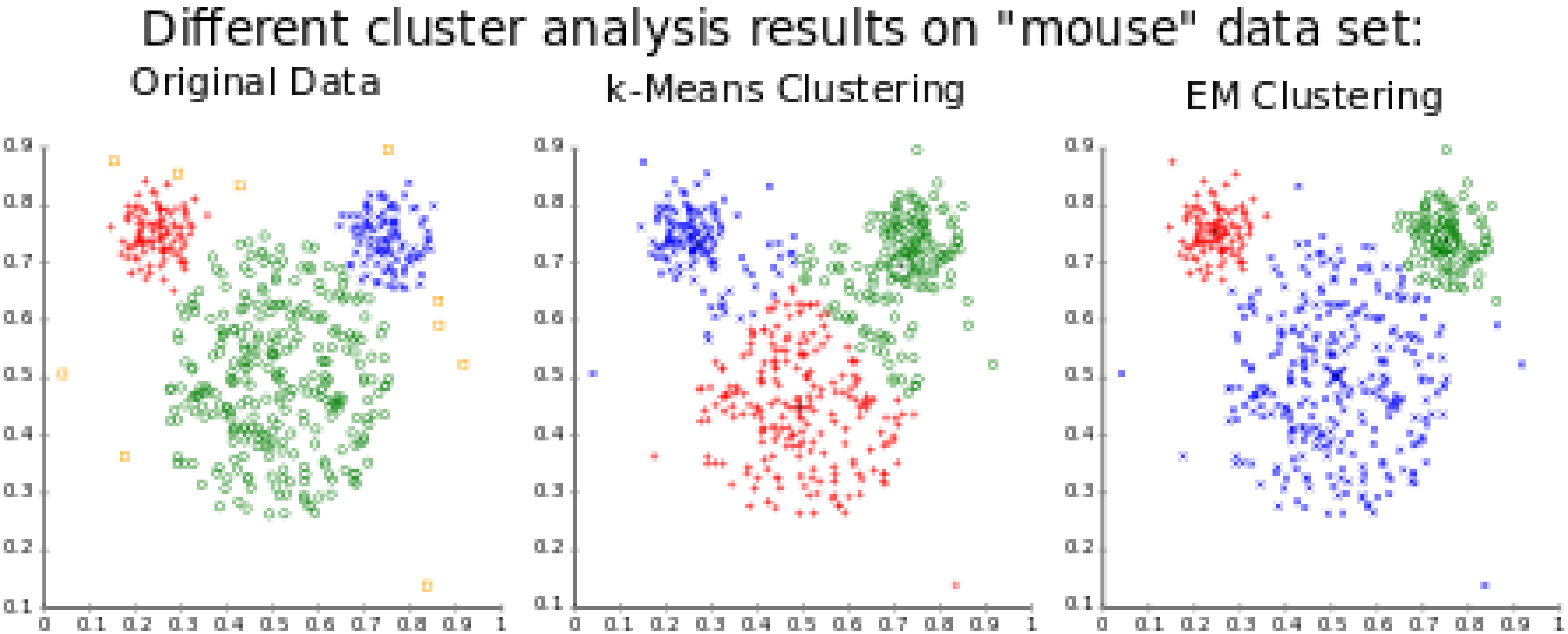
EXPECTATION-MAXIMIZATION

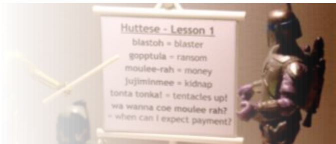




COMPARISON WITH K-MEANS

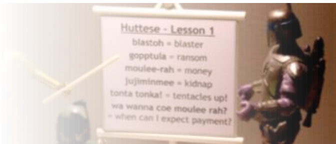






COMPARISON WITH K-MEANS

- 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.



SMOOTHING

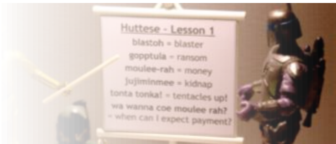
Problem.

- 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

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



SMOOTHING

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

Solution.

- Give prior probabilities to the σ_y.

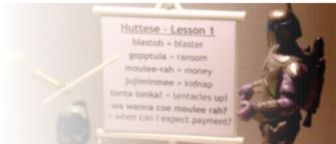
$$p\left(\sigma_y^2 \middle| \alpha_y, s_y^2\right) = C\,\left(2\pi\sigma_y^2\right)^{-\alpha_y d/2} \exp\left(-\frac{\alpha_y s_y^2}{2\sigma_y^2}\right)$$

- 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

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



SMOOTHING

Why do we choose prior probabilities of this form?

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

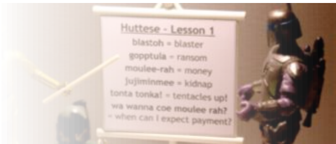
- Fix mean μ_y. Suppose we have α_y observations of $s_y + \mu_y$. The likelihood of these observations is

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

- The posterior probability of σ_y^2 will be

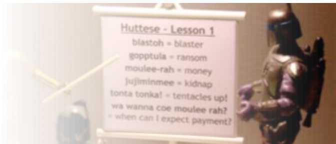
$$p\left(\sigma_y^2 \middle| \alpha_y, s_y^2\right) \propto p\left(\alpha_y, s_y^2 \middle| \sigma_y^2\right) p\left(\sigma_y^2\right).$$

- 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?



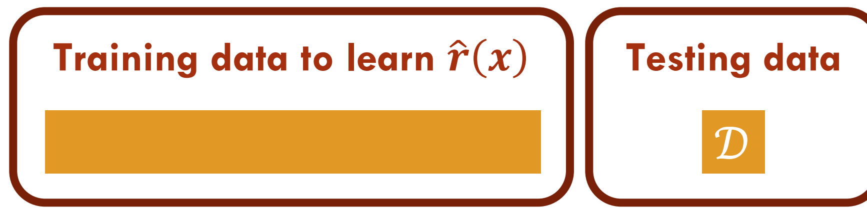
VALIDATION VS CROSS-VALIDATION

Method 1 (Simulation)

Estimate testing error using simple validation or cross-validation.

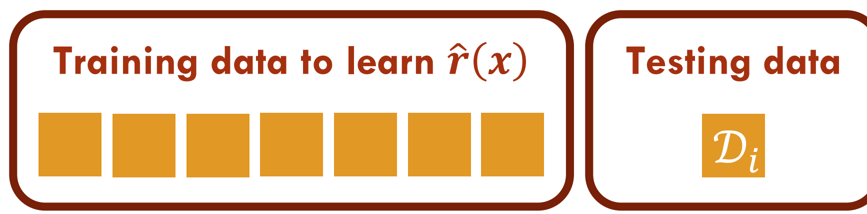
testing error

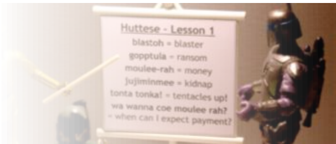
- $\hat{R}(\mathcal{D})$



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

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





BAYESIAN INFORMATION CRITERION

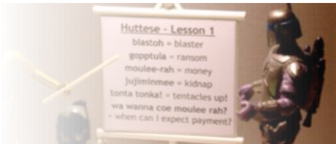
Method 2 (Marginal Likelihood)

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

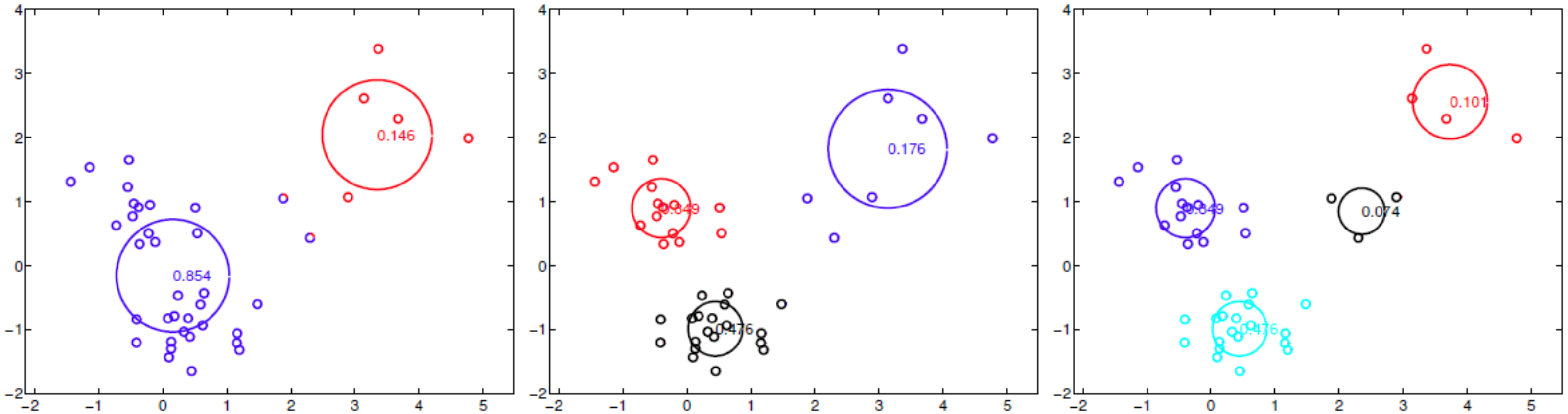
$$\mathrm{BIC}(\theta) = \mathcal{L}_n(\theta) - \frac{\#\text{ of free params}}{2} \log n$$

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

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

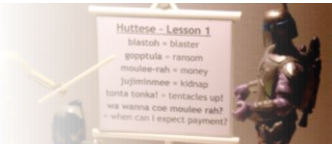


BAYESIAN INFORMATION CRITERION



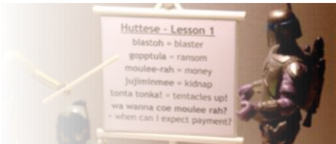
$$l(D;\hat{\theta}) = -118.25 \qquad l(D;\hat{\theta}) = -98.64 \qquad l(D;\hat{\theta}) = -94.11$$

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



SUMMARY

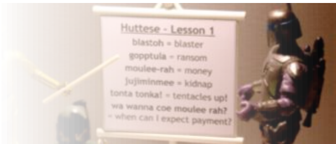
- Expectation-Maximization
  - Mixture Model
  - Clustering
  - Hidden Variables
  - Soft Labels
- Generalization
  - Priors and Smoothing
  - Model Selection
  - Validation and Cross-Validation
  - Bayesian Information Criterion



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.