Johns Hopkins
Whiting School of Engineering

Algorithms for Data Science

Statistical Algorithms: Gaussian Mixture Models

Introducing Gaussian Mixture Models (GMMs)

GMMs provide the flexibility to model complex, multi-modal data distributions.

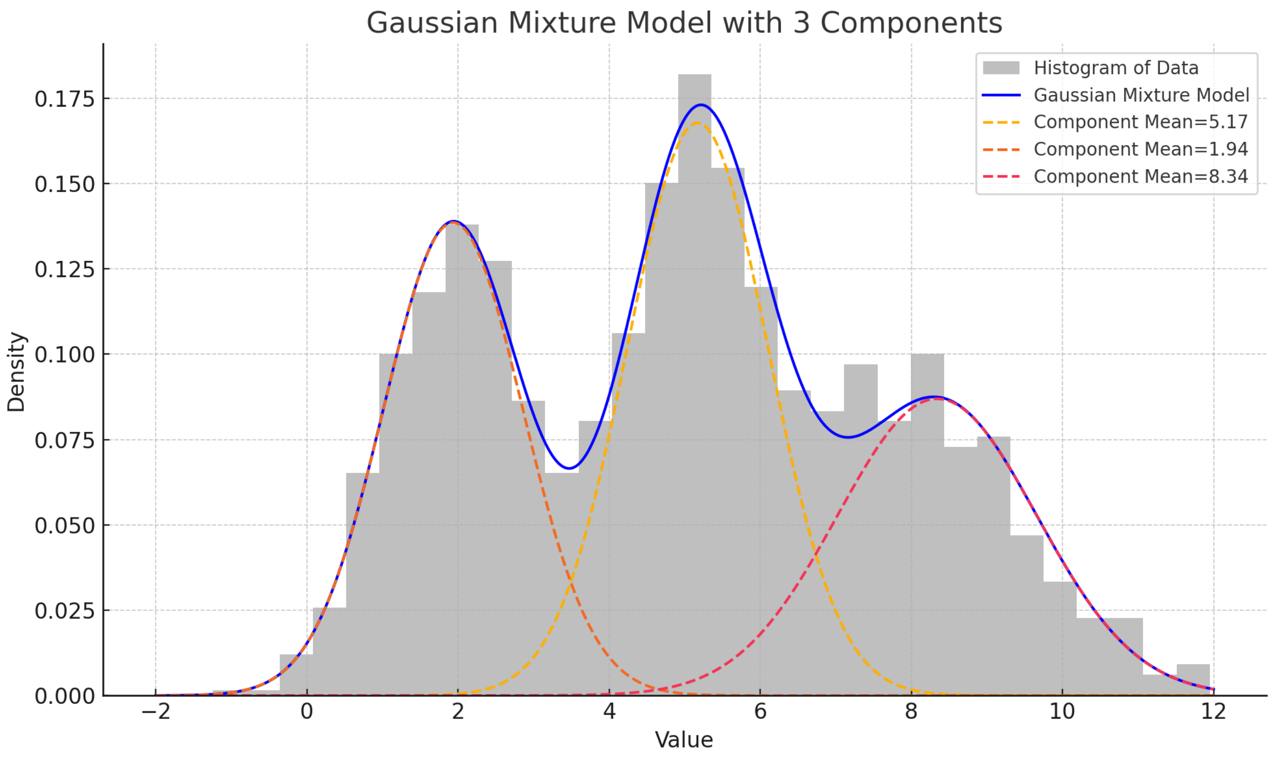
How do GMMs Work?

- Model data as a mixture of Gaussian distributions.
- Each data point is probabilistically assigned to one or more components.

Applications

- **Clustering:** Assign data points to overlapping groups.
- **Density Estimation:** Model the probability distribution of data.
- **Practical Use Cases:** Image segmentation, speech recognition.

Visualizing Gaussian Mixtures



JOHNS HOPKINS WHITING SCHOOL of ENGINEERING

GMMs: Mathematical Foundations

- GMM models data as a weighted sum of *K* Gaussian components:

$$P(x) = \sum_{k=1}^{K} \pi_k N\left(x \mid \mu_k, \Sigma_k\right)$$

Where:

π_k: Mixing coefficient (prior probability of component *k*, where $\sum_{k=1}^{K} \pi_k = 1$

μ_k : Mean vector of component *k*

Σ_k : Covariance matrix of component *k*

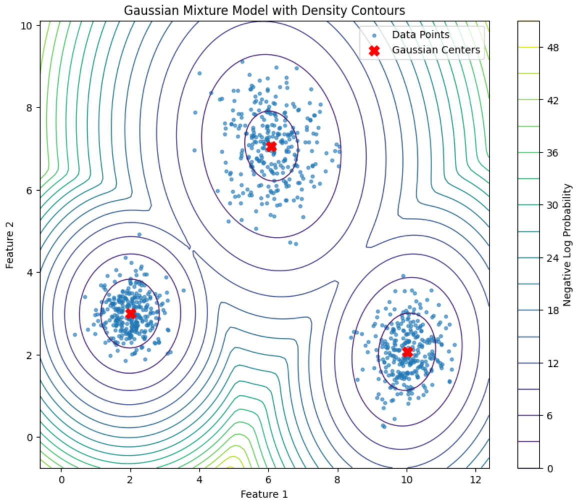
$N\left(x \mid \mu_k, \Sigma_k\right)$: Gaussian density function: $\dfrac{1}{(2\pi)^{D/2}\left|\Sigma_k\right|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T \sum_k^{-1}(x-\mu_k)}$

Parameters to Estimate (using EM algorithm): π_k, μ_k, Σ_k

GMMs: Mathematical Foundations

- Soft Clustering: Assigns probabilities γ_{nk} of data point x_n belonging to component *k*:

$$\gamma_{nk} = \frac{\pi_k N\left(x_n \mid \mu_k, \Sigma_k\right)}{\sum_{j=1}^{K} \pi_j N\left(x_n \mid \mu_j, \Sigma_j\right)}$$



GMM: Algorithm Analysis

1. **Initialization:**
   - Randomly initialize π_k, μ_k, Σ_k

   $\leftarrow O(1)$

2. **E-Step:**
   - Compute responsibilities γ_{nk} for each data point x_n and component *k*.

   $\leftarrow O\left(N \cdot K \cdot D^2\right)$

3. **M-Step:**
   - Mixing coefficients: $\pi_k = \frac{\sum_{n=1}^{N} \gamma_{nk}}{N}$
   - Means: $\mu_k = \frac{\sum_{n=1}^{N} \gamma_{nk} x_n}{\sum_{n=1}^{N} \gamma_{nk}}$
   - Covariance matrices: $\Sigma_k = \frac{\sum_{n=1}^{N} \gamma_{nk} \left(x_n - \mu_k\right)\left(x_k - \mu_k\right)^T}{\sum_{n=1}^{N} \gamma_{nk}}$

   $\leftarrow O\left(N \cdot K \cdot D^2\right)$

4. **Repeat Until Convergence:**
   - Compute responsibilities γ_{nk} for each data point x_n and component *k*.

Total Runtime Complexity:
$O\left(I \cdot N \cdot K \cdot D^2\right)$

GMM: Correctness

- **Theoretical Basis:** GMMs maximize the likelihood of the observed state:

$$\log P(X \mid \pi, \mu, \Sigma) = \sum_{n=1}^{N} \log \left(\sum_{k=1}^{K} \pi_k N\left(x_n \mid \mu_k, \Sigma_k\right) \right)$$

  - The EM algorithm guarantees monotonic increases in the log-likelihood at each iterations.
- **E-Step:** Responsibilities are computed to ensure probabilistic assignments.
- **M-Step:** Parameters are updated to maximize the likelihood given current responsibilities.
- **Convergence:** EM converges to a local maximum, the solution will depend on initialization and may reach suboptimal maxima.

GMM: Applications

Clustering

Grouping data into overlapping clusters such as in customer segmentation.

Density Estimation

Modeling probability distributions for anomaly detection such as in identifying fraudulent transactions.

Image Segmentation

Dividing an image into regions based on pixel intensity or color such as in medical imaging.

Speech Recognition

Modeling acoustic features for phoneme classification such as in identifying spoken words.

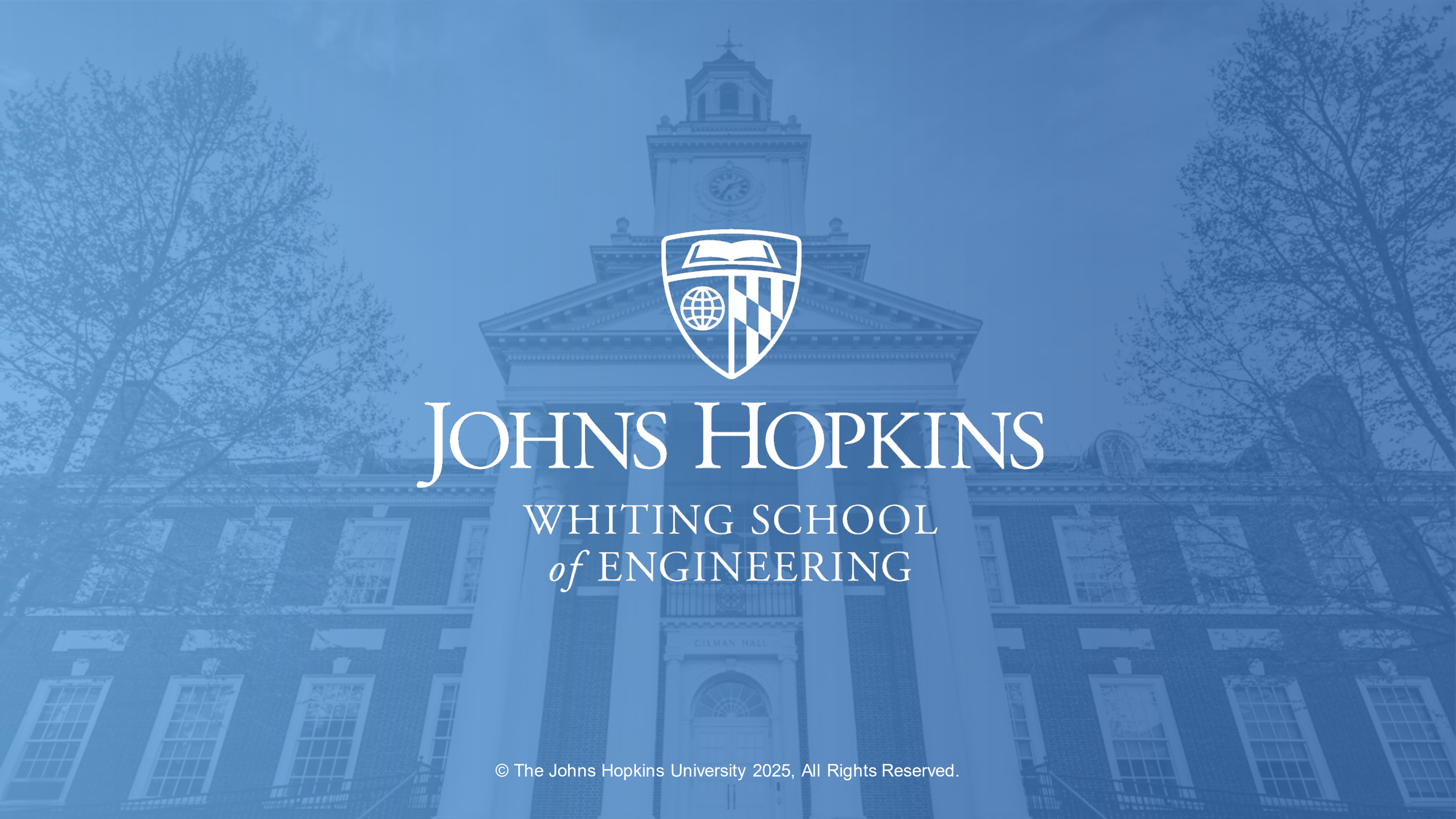
Wrapping Up Statistical Algorithms

Foundational Concepts

- Statistical algorithms leverage probability and statistics to model uncertainty and derive insights.
- Key characteristics include handling noise, modeling data distributions, and offering interpretability.

Key Algorithms

- **Expectation-Maximization (EM):** General framework for estimating parameters in models with latent variables.
- **Bayes Classifier:** Classifies under uncertainty by minimizing misclassification with known distributions.
- **Naive Bayes:** Simplifies Bayes by assuming feature independence, offering scalability with high-dimensional data.
- **Gaussian Mixture Models:** Models multi-modal distributions for clustering and density estimation.



JOHNS HOPKINS

WHITING SCHOOL of ENGINEERING