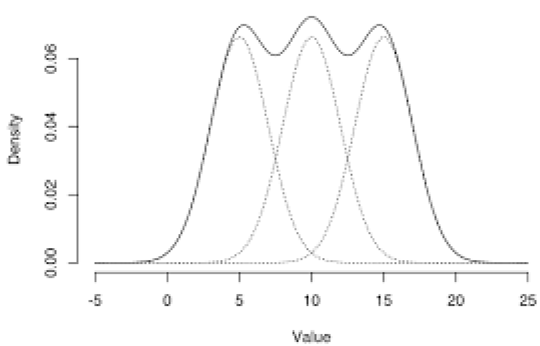
*Equations on slides

Gaussian Mixture Models(GMMs)

- More sophisticated and handles the problems of K-means
- Making the assumption of the underlying probability distribution of each data point. Ie calculates the probability a point belongs to a given clusters
    - Problems with K-means
        - Does not account for different cluster sizes, variances, and shapes
        - Does not allow points to belong to multiple clusters
        - Not Generative(It cannot create new data points)
            - Can put into clusters(discriminative)
        - We have to specify the K
    - Discriminative vs Generative Algorithms
        - Discriminative: Finds decision boundary
            - Ex: Logistic Regression, K-means
        - Generative: estimates probability distributions
            - Ex: Naive Bayes, GMMs

- Cluster sizes= How many data points belong to a cluster
    - How likely each of the clusters are
    - We want to maximize the likelihood of getting a data point
        - The goal is to find the model parameters that make the observed data most probable under the model
    - The GMM equation combines these Gaussian distributions, weighted by their respective mixing coefficients, to form the overall probability density function



Initialization Step:

- Probability: $\pi_k = 1/K$
- Mean: μ_k = choose random point
- Variance: (σk 2) = sample variance

- E Step: In the E-step, the step estimates the probability that each data point belongs to each cluster (Gaussian component)
- M Step: In the M-step, the step updates the model parameters to maximize the log-likelihood, using the calculations in the E-step.

- Generative Process:
    - Sample cluster k using ($\pi 1, \pi 2, \ldots, \pi k$)
    - Sample x from $N(\mu_k, \sigma k\ 2)$

Kernel Density Estimation(KDE)

- Kernel Density Estimation (KDE) is a non-parametric method used to estimate the probability density function of a random variable based on a finite set of data points. It smooths the distribution by placing a kernel on each data point and summing these kernels to produce a continuous estimate of the density.

  Components of Equation
    - Data Points (x_i):
        - The observed data points used to estimate the density.
    - Kernel Function(K):
        - A smooth, symmetric function that determines the shape of the influence of each data point.
    - Bandwidth (h):
        - A smoothing parameter that controls the width of the kernel.
        - Smaller h creates a more detailed estimate but can overfit (too spiky).
        - Larger h smooths the estimate more but can underfit (too flat).
- Common Kernel Functions
    - Uniform
    - Triangle
    - Gaussian
    - Cosine
- Width Selection:
    - As stated above: Smaller h creates a more detailed estimate but can overfit and larger h smooths the estimate more but can underfit

Missing Data

- Types of Missing Data
    - MCAR: Missing Completely At Random.
        - Not related to:
            - Specific values
            - Observed responses

    - MAR: Missing at Random. Not Related to Specific Values. Could be related to observed responses
    - MNAR: Missing Not At Random

- Techniques for handling missing data:
    - Try to prevent the problem in the first place. This means careful study design, follow-up with participants, etc
    - Omit rows with missing data(reduces n)
    - Omit only when value is needed
        - Naive bayes, per-feature estimates
    - Mean substitution(per-feature)
    - Imputation: Use similar examples to guess missing values
    - Last observation carried forward
        - Useful for time series data