CS 260: Foundations of Data Science

Prof. Thao Nguyen

Fall 2024

HAVERFORD
COLLEGE

Materials by Sara Mathieson

Admin

- **Midterm 2** returned today
- **Final project presentation** sign-up on Piazza
  - Email me pdfs of your slides the night before
  - Class attendance taken for Dec 09 & 11

Outline for today

- Gaussian Mixture Models (GMMs)
- Kernel Density Estimation (KDE)
- Missing data
- Go over Midterm 2

Outline for today

- Gaussian Mixture Models (GMMs)
- Kernel Density Estimation (KDE)
- Missing data
- Go over Midterm 2

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 (cannot create a new data point)

Discriminative vs. Generative Algorithms

- Discriminative: finds a decision boundary
  - Logistic regression, K-means
- Generative: estimates probability distributions
  - Naïve Bayes, Gaussian Mixture Models

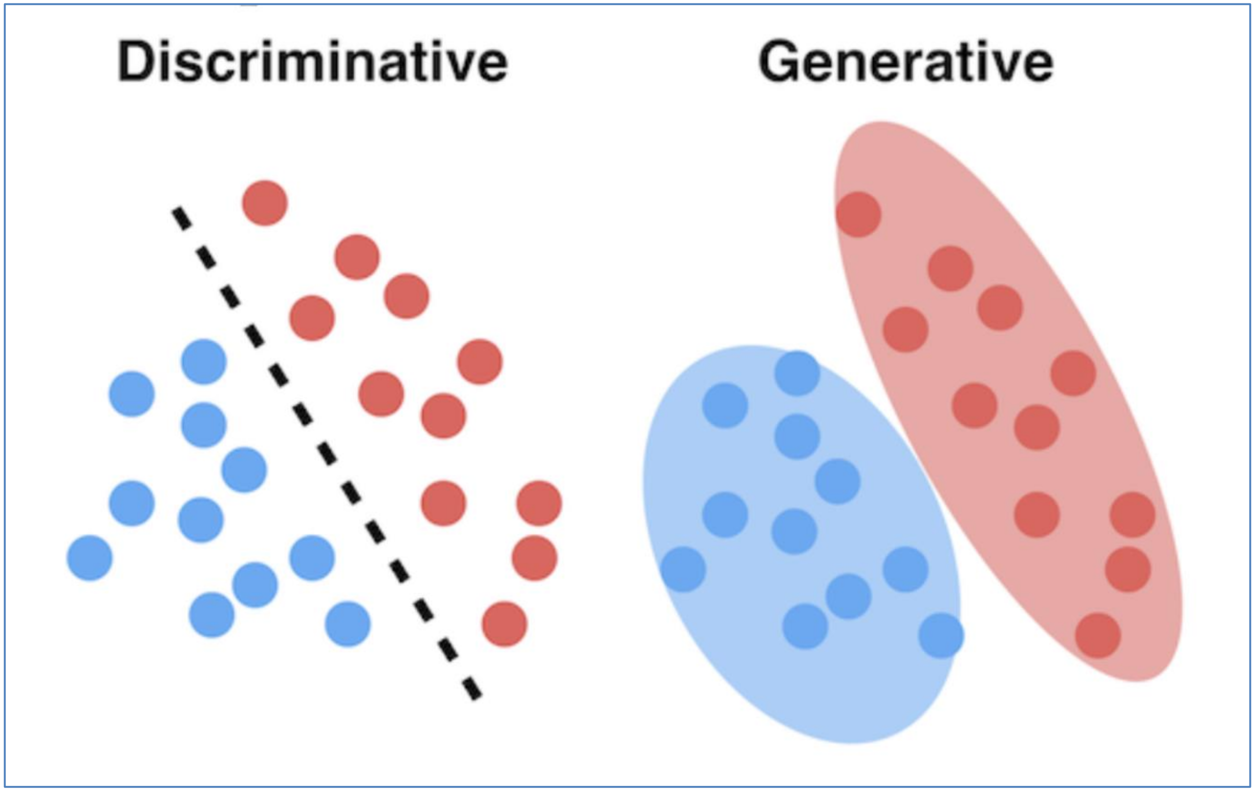


Figure: Ameet Soni

Gaussian Mixture Models (GMMs)

$$p(\vec{x}_i) = \sum_{k=1}^{K} p(\vec{x}_i, k) = \sum_{k=1}^{K} p(k)p(\vec{x}_i|k) = \sum_{k=1}^{K} \pi_k N(\vec{x}_i|\vec{\mu}_k, \sigma_k^2)$$

cluster membership

prior over cluster sizes

Gaussian distribution

- Maximize likelihood:

$$L(X) = \prod_{i=1}^{n} p(\vec{x}_i) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k N(\vec{x}_i|\vec{\mu}_k, \sigma_k^2)$$

Model parameters

Gaussian Mixture Models (GMMs)

- Initialization step: for each cluster
  - Probability $\pi_k = 1/K$ (uniform prior)
  - Mean $\vec{\mu}_k$ = choose random point
  - Variance σ_k^2 = sample variance

- E-step: "soft" assignment

$$w_{ik} = p(k|\vec{x}_i) = \frac{p(k)p(\vec{x}_i|k)}{p(\vec{x}_i)} = \frac{\pi_k N(\vec{x}_i|\vec{\mu}_k, \sigma_k^2)}{\sum_{j=1}^{K} \pi_j N(\vec{x}_i|\vec{\mu}_j, \sigma_j^2)}$$

probability that $\vec{x}_i$ came from cluster k

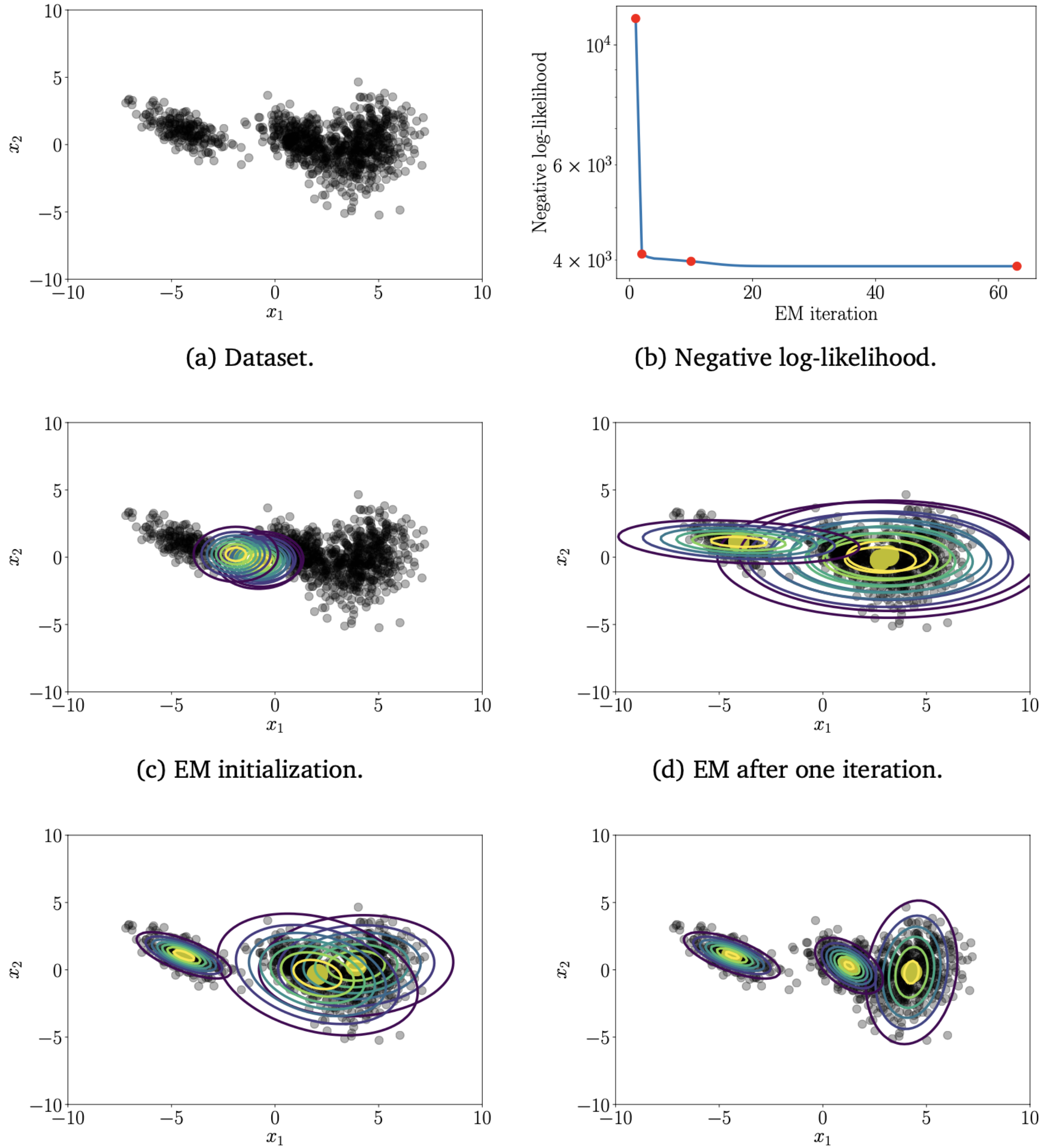
Gaussian Mixture Models (GMMs)

- <u>M-step:</u> parameter update

  $N_k = \sum_{i=1}^{n} w_{ik}$ (# of points assigned to cluster k)

  - $\pi_k = \frac{N_k}{n}$
  - $\vec{\mu}_k = \frac{1}{N_k} \sum_{i=1}^{n} w_{ik} \, \vec{x}_i$
  - $\sigma_k^2 = \frac{1}{N_k} \sum_{i=1}^{n} w_{ik} \left(\vec{x}_i - \vec{\mu}_k \right)^2$

    use updated mean



(a) Dataset.

(b) Negative log-likelihood.

(c) EM initialization.

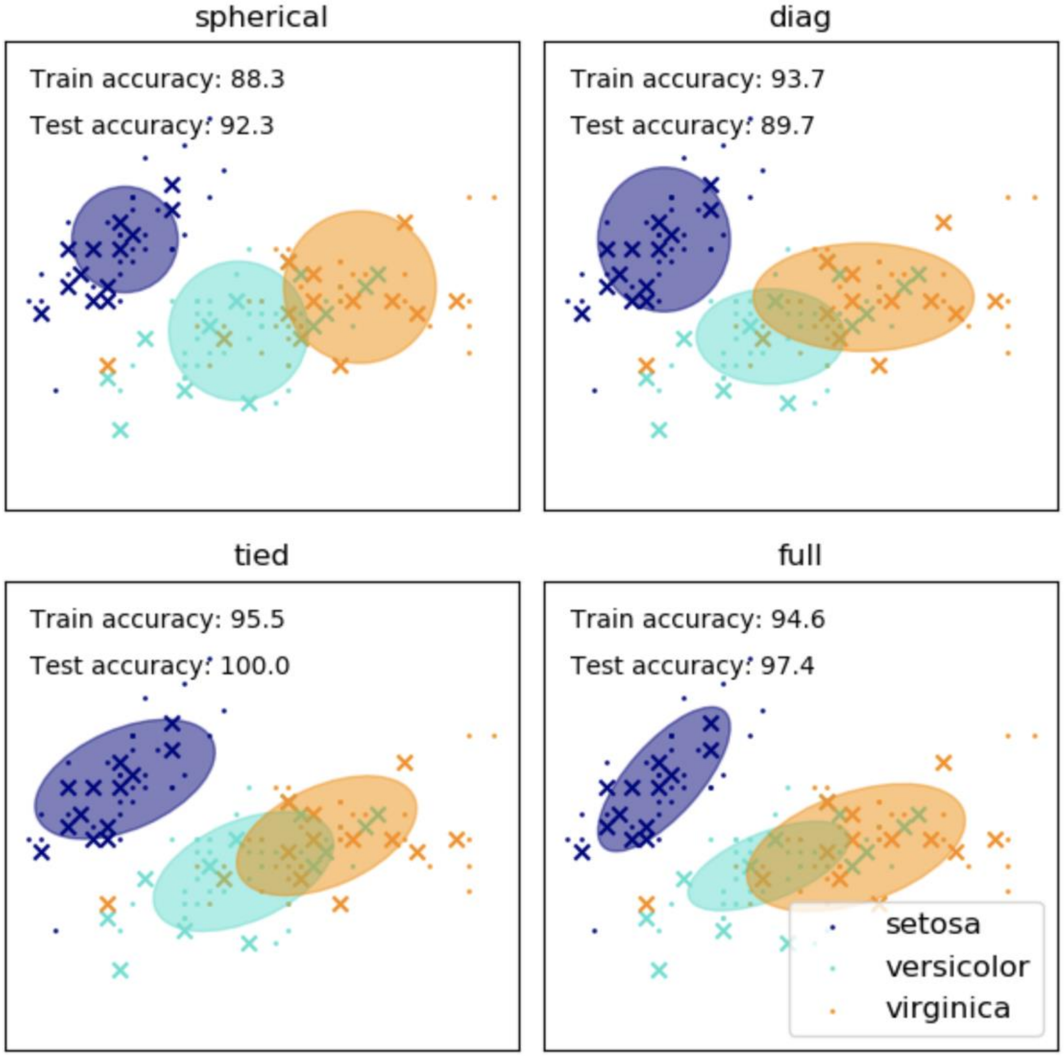
(d) EM after one iteration.

(e) EM after 10 iterations.

(f) EM after 62 iterations.

Figure 11.7 from MML textbook

Example of GMMs with different covariance constraints on the Iris flower data



https://scikit-learn.org/stable/auto_examples/mixture/plot_gmm_covariances.html#sphx-glr-auto-examples-mixture-plot-gmm-covariances-py

Generative Process

- Sample cluster k using $[\pi_1, \pi_2, \dots, \pi_k]$
- Sample x from $N(\vec{\mu}_k, \sigma_k^2)$

Outline for today

- Gaussian Mixture Models (GMMs)
- Kernel Density Estimation (KDE)
- Missing data
- Go over Midterm 2

KDE (Kernel Density Estimation)

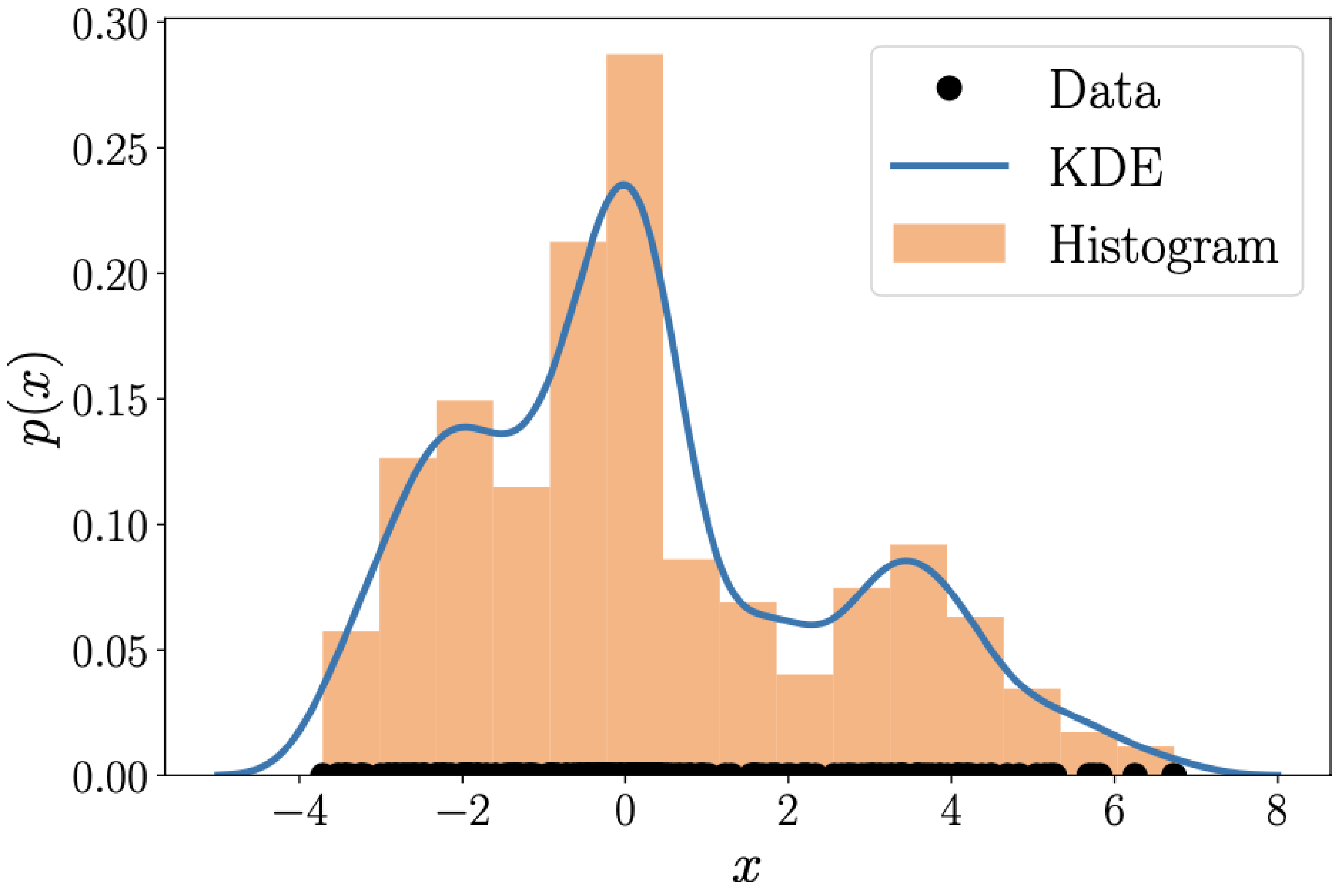


Figure 11.9 from MML textbook

KDE (Kernel Density Estimation)

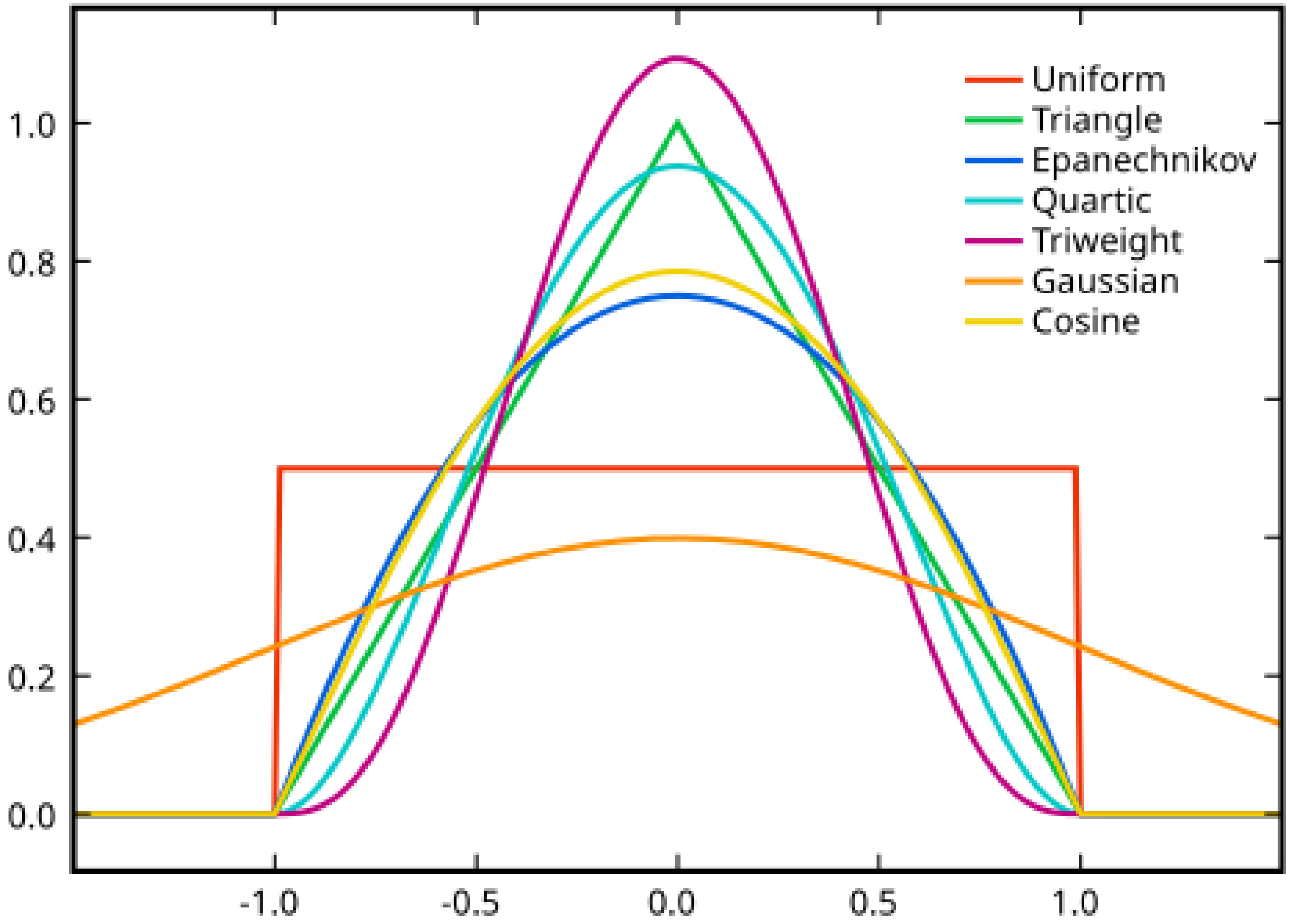
$$p(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x - x_i}{h}\right)$$

given point

kernel: weighting function

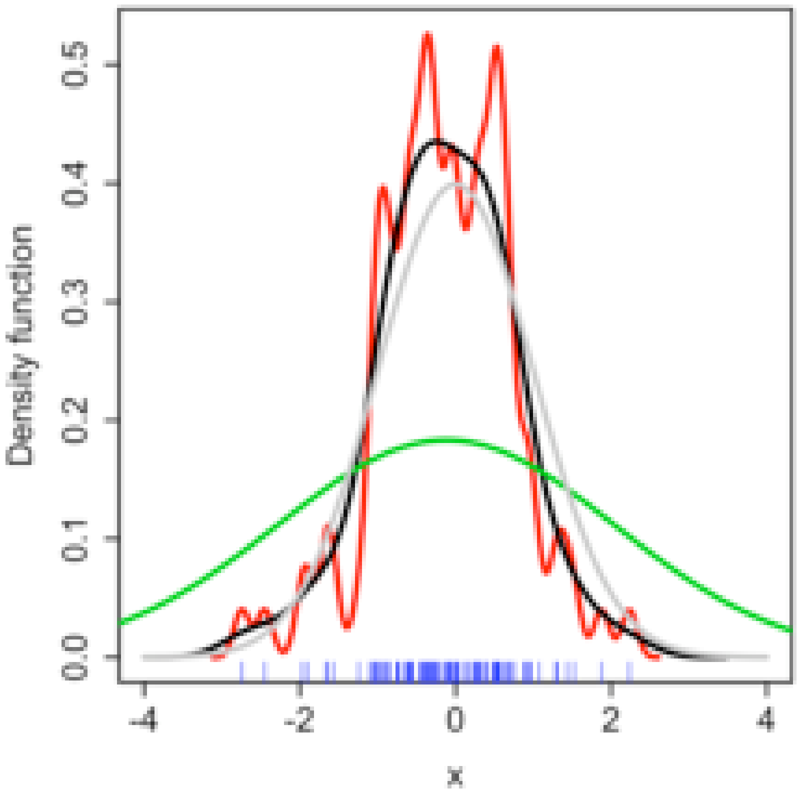
width: smoothing parameter

Commonly used kernel functions



Wikipedia

Width selection



Wikipedia

Kernel density estimate (KDE) with different bandwidths of a random sample of 100 points from a standard normal distribution. Grey: true density (standard normal). Red: KDE with h=0.05. Black: KDE with h=0.337. Green: KDE with h=2.

Outline for today

- Gaussian Mixture Models (GMMs)
- Kernel Density Estimation (KDE)
- Missing data
- Go over Midterm 2

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

- MNAR: Missing Not At Random

Reference: "The prevention and handling of the missing data" Kang (2013)

Techniques for handling missing data

- Try to prevent the problem in the first place
  - Careful study design, follow-up with participants, etc
- Omit rows with missing data (reduces *n*)
- Omit only when value is needed
  - i.e. Naïve Bayes, per-feature estimates
- Mean substitution (per feature)

Reference: "The prevention and handling of the missing data" Kang (2013)

Techniques for handling missing data

- Imputation
  - Use similar examples to guess the missing values
  - Can be done locally or globally
- Last observation carried forward
  - Useful for time-series data

Reference: "The prevention and handling of the missing data" Kang (2013)

Outline for today

- Gaussian Mixture Models (GMMs)
- Kernel Density Estimation (KDE)
- Missing data
- Go over Midterm 2

Midterm solutions
not posted online