Kernel Density Estimation in Machine Learning

Team Members: Connor Hill, Gabriel Gray, Jay Woo, Julia Messegee

Introduction to KDE

What is Kernel Density Estimation (KDE)?

- A non-parametric method to estimate the probability density function (PDF) of any variable.

Advantages of Non-Parametric

- No assumption for underlying data distribution.
- Flexibility in modeling multimodal datasets.

Significance in Machine Learning

- Quick identification of key features in datasets (peaks, skewness, and spread).
- Improves probabilistic models and understanding data distributions.

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

$f(x)$: Estimated density at x

n: Number of data points

h: Bandwidth (smoothing parameter)

K: Kernel function

Kernel functions: Define the shape of the weighting function (Gaussian, Epanechnikov, Uniform, Triangular)

Bandwidth Selection h

Role of h

- Controls the smoothness of the density estimate.

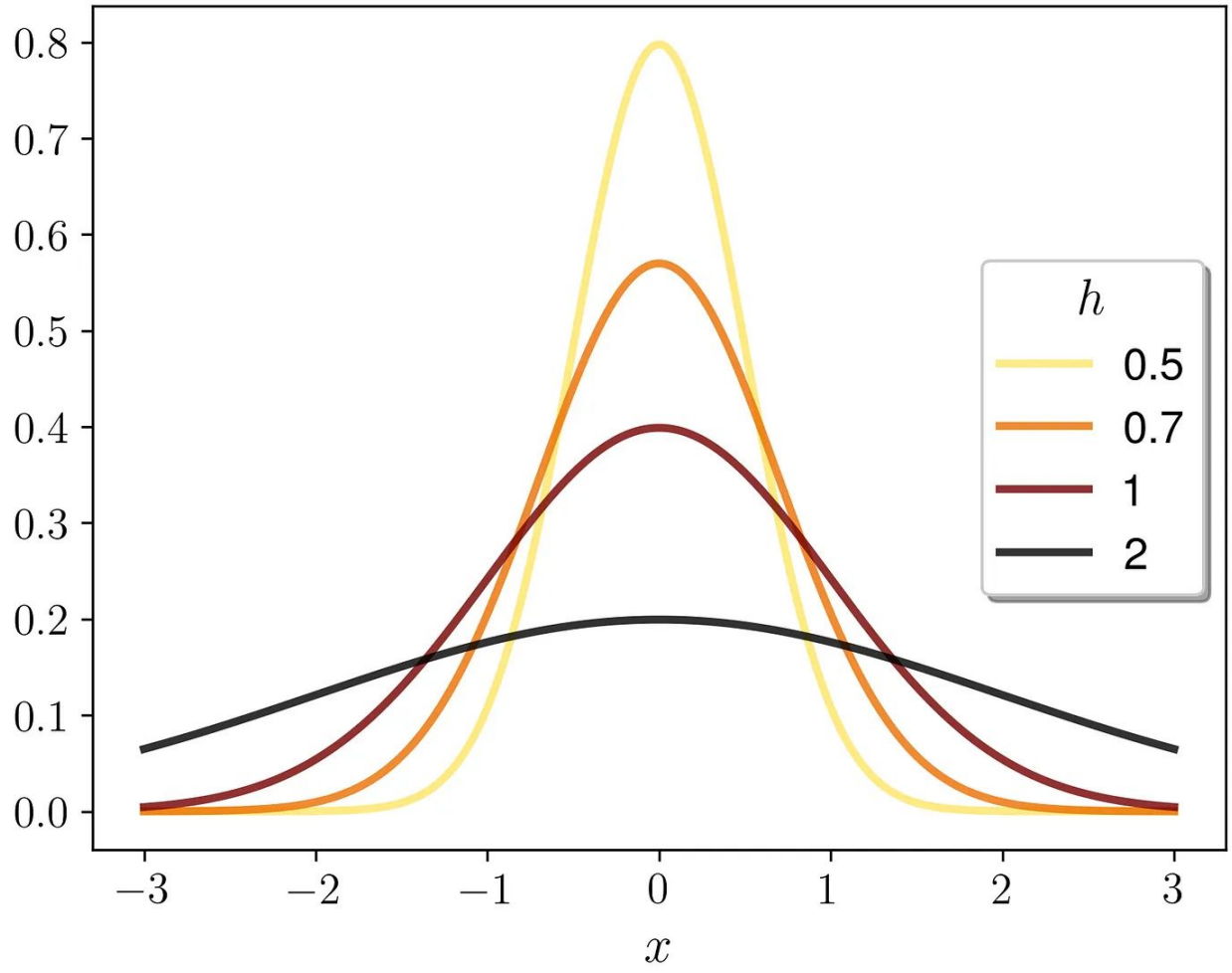
Small h vs Large h

- Captures data nuances but risks overfitting.
- Smoother density estimates but risks underfitting.

Bandwidth selection

- Cross-Validation: Least-Squares Cross-Validation (LSCV).
- Plug-in Methods: Solve for optimal h analytically.

All about finding balance.


0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−3
−2
−1
0
1
2
3
x
h
0.5
0.7
1
2

Kernel Functions

Gaussian Kernel:

- Smooth, bell-shaped curve.
- Sensitive to outliers.

Epanechnikov Kernel:

- Parabolic shape.
- Optimal in minimizing mean integrated squared error (MISE).

Choosing a Kernel:

- Impact on estimation is minor compared to bandwidth selection.
- Practical choice often defaults to Gaussian due to mathematical convenience.

KDE in the Real-World

Anomaly Detection:

- Identifying suspicious network activity/fraud detection by modeling normal behavior distributions.

Data Visualization:

- Implementing smooth histograms and contour plots in standard data analysis.

Finance:

- Estimating the distribution of asset returns for risk management.

Environmental Science:

- Modeling spatial distributions of pollutants or animal habitats.

Challenges and Considerations

Curse of Dimensionality:

- KDE performance degrades with high-dimensional data.
- Mitigated through dimensionality reduction techniques like PCA.

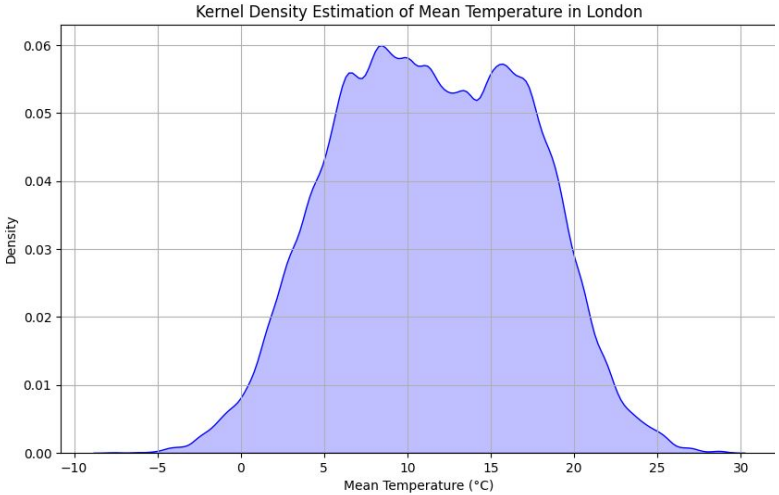
Bandwidth Selection Complexity:

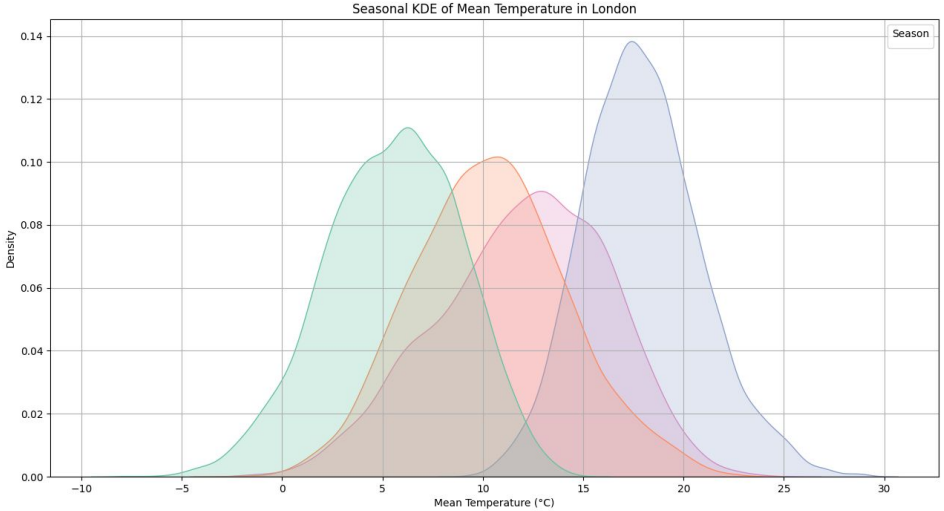
- Requires careful tuning as no one size works.

Computational Efficiency:

- KDE can be computationally intensive for larger datasets.
- Approximation methods, efficient algorithms can mitigate this.

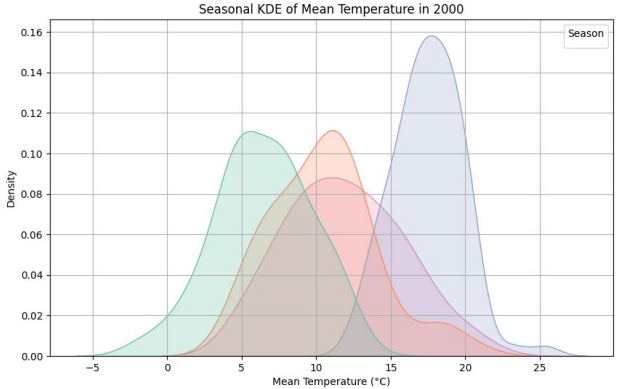
KDE on London Weather Data

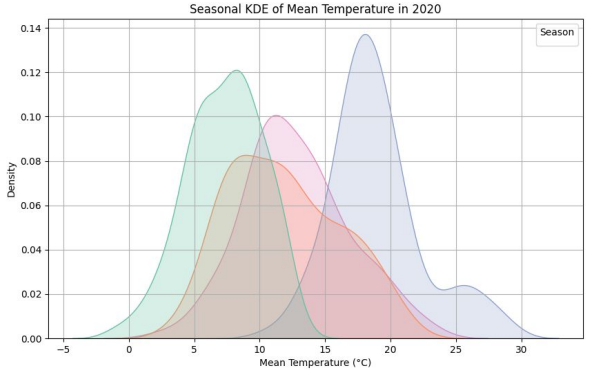


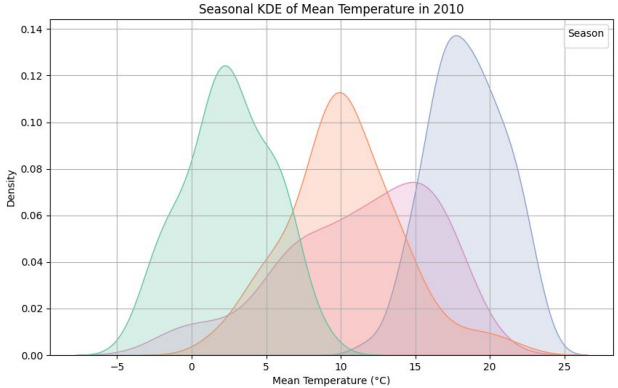


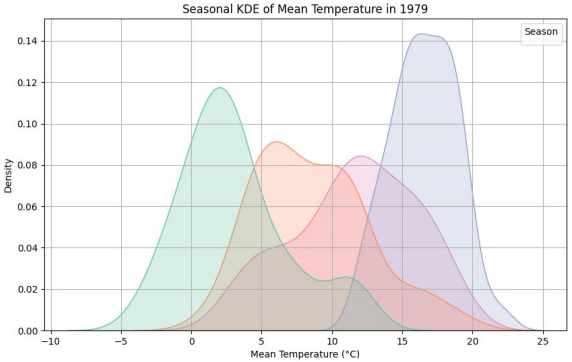
London Weather patterns from 1979-2021

1979, 2000, 2010, 2020 Comparison









Kernel Density Estimation In Machine Learning

Feature Engineering:

- Using density estimates as features for classification or regression models.

Probabilistic Modeling:

- Enhancing models like Naive Bayes or Hidden Markov Models with accurate PDFs.

Clustering:

- Mean Shift Clustering uses KDE to find clusters by locating maxima in the density function.