Project Proposal - Project 4

Connor Hill, Jay Woo, Gabriel Gray

Project Statement

We have decided to explore Kernel Density Estimation (KDE) and its applications through machine learning in statistics. Our motivation for doing so lies in the fact that KDE is a useful non-parametric technique for estimating the probability density function (PDF) without reliance on predefined distribution forms. This flexibility makes it ideal for modeling complex, multimodal distributions seen in real-world data, revealing its importance in practical applications.

Kernel Density Estimation functions as our chosen algorithm by first placing a kernel of our choosing at individual data points n. They operate as functions that determine the influence of the data point on the estimate for the location. The bandwidth h controls the width of the kernel and affects the density estimate's smoothness. Smaller bandwidth is better for details, larger for more general estimates.

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

And this KDE formula provides the general outline for summing the kernel contributions for each data point. This generates the estimate of the data distribution, good for breaking down multimodal datasets.

Figuring the underlying probability distribution can improve ML models that are based on probability. As an example, bayes' net is based on probability, and KDE can improve bayes' net by giving better probabilities from a raw data set. Therefore, understanding KDE would be very helpful for designing ML models.

Datasets

- London Weather Data
  - https://www.kaggle.com/datasets/emmanuelfwerr/london-weather-data
  - The dataset featured below was created by reconciling measurements from requests of individual weather attributes provided by the European Climate Assessment (ECA). The measurements of this particular dataset were recorded by a weather station near Heathrow airport in London, UK.
- World City air and water quality
  - https://www.kaggle.com/datasets/cityapiio/world-cities-air-quality-and-water-polution
  - Air quality varies from 0 (bad quality) to 100 (top good quality)
  - Water pollution varies from 0 (no pollution) to 100 (extreme pollution)

Summary of Discussion

The team (Gabriel, Jay, Connor) decided to focus on KDE because it was a non parametric way to estimate PDFs without a reliance on any predefined distributions. This approach allows complex analysis with multimodal datasets which as a team we decide would be complex enough and have implications in our future.

Another consideration we had was using MineRL which is a Minecraft based simulator that allows reinforcement learning algorithms.Using MineRL we would attempt to teach an efficient mining algorithm. In the end we decided that going down the KDE path was best for our team.

Team Roles

- Lead/Liaison - Will be overseeing the project as it develops, paying attention to deadlines and goals to keep the project on track. Primary tasks will be taking the lead on team meetings, managing plans, and serving as a point of contact between the team and the instructor/TA's. In addition to these tasks, delegating other tasks and monitoring progress is important.
    - Connor Hill
- Programming Lead - Primarily working with the chosen algorithm(s) and overseeing its integration and development for the project. Primary tasks include implementing and testing models utilized in the project, and ensuring these models score high in accuracy and efficiency.
    - Connor Hill, Jay Woo, Gabriel Gray (Tri-Leads)
- Research Lead - As we are taking a more educational approach to our project by taking an unseen algorithm and expanding on it, we need to ensure that every aspect of our research is accurate. This role includes taking the lead on research-heavy portions of the project and ensuring other members are properly implementing the findings pertaining to our topic.
    - Jay Woo
- Data Handler - This role works closely with the programming lead to provide/clean up data utilized by models during the project. Ensuring we have the proper data sets for these algorithms will streamline their implementation and provide consistent, accurate results.
    - Gabriel Gray

Timeline

Interval	Goal
Week 10	Creating tasks based on project topic and scope. Assigning tasks to group members based on role and work distribution.
Week 11	Studying KDE academic papers/examples to solidify understanding. Deriving important equations to ensure understanding for the implementations.
Week 12	Cleaning and Parsing data to prepare to train our algorithm. Writing and testing the first iteration of the KDE algorithm.
Week 13	Refining the KDE algorithm and continued testing. Applying visualization tools to plot curves and density estimates.
Week 14	Finalizing testing and compiling the results. Performing analysis and kernel comparisons.

If we are unable to meet the goals outlined above, we may choose to do one or more of the following: Finding and utilizing alternative datasets, simplifying the scope of the project, or leveraging existing libraries for assistance with our modeling. In dire cases, we may go with our back up algorithm of Reinforcement Learning with an accompanying proposal change.