Laboratory assignment

Component 4

Authors: Ichim Stefan, Mirt Leonard
Group: 246/1

January 8, 2025

Related Works

The California Housing dataset, while widely used in machine learning research, has seen relatively few applications in unsupervised learning. The dataset was originally introduced by Pace and Barry (1997) for spatial autoregression analysis, with features specifically selected and engineered for housing price prediction. It contains only 8 features plus the target variable (median house value), with limited redundancy and clear numerical relationships between variables. These characteristics have made it particularly suitable for supervised learning tasks, especially regression problems. The dataset's popularity was further cemented through its inclusion in scikit-learn as a benchmark dataset for predictive modeling, establishing strong precedent for supervised approaches in the literature.

1 Unsupervised Learning Task

Several studies have explored unsupervised learning techniques on the California Housing dataset. Two notable approaches have employed K-means clustering and Principal Component Analysis (PCA), respectively.

1.1 K-means Clustering Analysis

Xiao (2024) proposed a Comprehensive K-means clustering approach that evaluates the stability and consistency of clustering results. Their method assesses both the common patterns that emerge across different clustering trials and the robustness of these patterns. When applied to the California Housing dataset, their analysis revealed that as housing prices gradually converged on specific latitude values and decreased in longitude, several key metrics increased together: median house value, median income, number of households, population, and number of rooms. This suggested that areas in the mid-west portion of California tend to be more densely populated with larger houses.

1.2 Privacy-Preserving PCA

Kwatra et al. (2024) investigated PCA from a privacy-preserving perspective, analyzing both utility and privacy aspects of eigenvector computation. Their work is particularly relevant for scenarios where housing data needs to be analyzed while preserving individual privacy. They evaluated different approaches including k-anonymity and synthetic data generation using CTGAN.

1.3 Performance Analysis

1.3.1 Technical Terminology

Before examining the results, we define key technical concepts used in this analysis:

- **K-anonymous Data:** A privacy protection technique where each record is modified to be indistinguishable from at least k-1 other records. For example, with k=20, each house's characteristics are adjusted until they match at least 19 other houses, protecting individual privacy while maintaining overall patterns.
- **Synthetic Data:** Artificially generated data points that preserve the statistical properties of the original dataset without containing any real records. This provides complete privacy protection while maintaining the dataset's analytical utility.
- **R^2 Score:** Also known as the coefficient of determination, this metric ranges from 0.0 to 1.0 and indicates how well the reduced-dimensional representation preserves the variance in the original data. An R^2 of 0.781 means 78.1% of the original data's variability is captured.

1.3.2 Clustering Performance

The hierarchical clustering analysis revealed distinct patterns across different data versions:

Data Version	Cluster Merging Height	Interpretation
Original	250	Highest separation between clusters, indicating clear natural groupings in the raw data
Synthetic	150	Lower merging height suggests more uniform distribution of synthetic data points
k-anonymous	150-250	Intermediate separation, showing partial preservation of cluster structure while maintaining privacy

Table 1: Hierarchical Clustering Results Analysis

The cluster merging height indicates the dissimilarity level at which clusters are combined. Higher values suggest more distinct, well-separated clusters, while lower values indicate more closely related groupings.

1.3.3 PCA Performance Analysis

The Principal Component Analysis (PCA) results demonstrate the trade-off between dimensionality reduction and information preservation: Table 2.

1.3.4 Privacy-Utility Trade-off

The results demonstrate how different privacy preservation techniques affect data utility:

- **K-anonymous PCA:** This approach modifies the original data to ensure privacy (k=20) before applying PCA. The minimal reduction in R^2 scores (roughly 1-2% lower than original PCA) suggests that privacy can be preserved while maintaining most of the data's analytical value.
- **Synthetic Data Impact:** The lower clustering height (150 vs 250) in synthetic data indicates some loss of natural grouping structure, though the patterns remain sufficiently preserved for meaningful analysis.

Method	PCs	R^2 Score	Analysis
Baseline	All	0.781 ± 0.019	Reference performance using all original features
Original PCA	3	0.148 ± 0.034	Significant information loss with aggressive reduction
Original PCA	4	0.455 ± 0.038	Notable improvement with fourth component
Original PCA	5	0.631 ± 0.034	Major recovery of explanatory power
Original PCA	6	0.697 ± 0.029	Close to baseline performance
k-anonymous PCA	3	0.134 ± 0.030	Minimal privacy impact on low-dimensional projection
k-anonymous PCA	4	0.445 ± 0.035	Preserved relationship structure
k-anonymous PCA	5	0.623 ± 0.029	Strong performance despite anonymization
k-anonymous PCA	6	0.689 ± 0.032	Near-original performance with privacy guarantees

Table 2: Detailed PCA Performance Analysis

- **Performance vs. Privacy:** The progression of R^2 scores shows that with 5-6 principal components, both original and k-anonymous versions retain approximately 80% of the baseline performance, suggesting this is an optimal balance between dimensionality reduction and privacy preservation.

1.3.5 Statistical Significance

The performance differences between original and k-anonymous PCA are statistically insignificant (p greater than 0.05) for all component counts, demonstrating that privacy preservation does not meaningfully impact the utility of the dimensionality reduction.

2 Supervised Regression Task

2.1 Chen (2024)

The research by Chen (2024) presents a comprehensive analysis of supervised regression methods applied to the California Housing dataset from 1990. The study implements multiple regression approaches to predict house prices, beginning with traditional linear regression as a baseline model. The methodology extends to more sophisticated techniques, particularly Support Vector Regression (SVR) with various kernel functions including linear, polynomial, and Radial Basis Function (RBF). Additionally, the research explores deep learning through a neural network architecture comprising four hidden layers with 128, 64, 32, and 16 neurons respectively, utilizing ReLU activation functions. The performance evaluation primarily relies on Root Mean Square Error (RMSE), calculated as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

where y_i represents actual values and $\hat{y}_i$ represents predicted values. The empirical results demonstrate varying levels of predictive accuracy across the different models. The following table summarizes the performance metrics:

Model	Training RMSE	Test RMSE
Linear Regression	67,796	66,638
SVR (Linear)	69,470	68,070
SVR (Polynomial)	63,527	66,299
SVR (RBF)	58,182	57,405
DNN	59,912	59,263

Table 3: Model Performance Comparison

The analysis reveals that the SVR model with RBF kernel achieves superior performance, demonstrating the lowest RMSE values in both training and test sets. This suggests that the RBF kernel's ability to capture non-linear relationships in the housing data surpasses other approaches. The deep neural network emerges as the second most effective method, while linear regression and SVR with linear kernel show higher error rates, indicating their limitations in modeling complex housing price relationships. The findings highlight the particular suitability of RBF kernel-based SVR for real estate price prediction applications, offering valuable insights for both academic research and practical implementation in the real estate industry. As noted by Chen (2024), these results provide a foundation for future work in model optimization and feature engineering within the domain of housing price prediction.

2.2 Chen et al. (2022)

The research by Chen et al. (2022) presents a comprehensive analysis of supervised regression methods applied to the California Housing dataset from 1990. The study implements multiple regression approaches to predict house prices, beginning with traditional linear regression as a baseline model. The methodology extends to more sophisticated techniques, particularly Support Vector Regression (SVR) with various kernel functions including linear, polynomial, and Radial Basis Function (RBF). Additionally, the research explores deep learning through a neural network architecture comprising four hidden layers with 128, 64, 32, and 16 neurons respectively, utilizing ReLU activation functions. The performance evaluation primarily relies on Root Mean Square Error (RMSE). The empirical results demonstrate varying levels of predictive accuracy across the different models. The following table summarizes the performance metrics:

Table 4: Key Results from Chen et al. (2022)

Model	Training MSE	Test MSE
Random Forest	0.208	0.290
Gradient Boosting	0.215	0.295

The analysis reveals that the SVR model with RBF kernel achieves superior performance, demonstrating the lowest RMSE values in both training and test sets. This suggests that the RBF kernel's ability to capture non-linear relationships in the housing data surpasses other approaches. The deep neural network emerges as the second most effective method, while linear regression and SVR with linear kernel show higher error rates, indicating their limitations in modeling complex housing price relationships.

The Chen et al. (2022) paper further explores the data preprocessing and feature engineering aspects of the problem. The authors examined the importance of geospatial features, such as ocean proximity, and created additional categorical variables to capture this information. They also investigated the non-normal distribution of the house prices and applied log-transformation to make the data more Gaussian-like.

The researchers used cross-validation and K-Fold methods to tune the hyperparameters of each regression model, including Ridge Linear Regression, Random Forest, and Gradient Boosting. The Random Forest model achieved the best performance with an MSE of 0.290 on the test set, followed by Gradient Boosting with an MSE of 0.295.

The findings from the Chen et al. (2022) study highlight the particular suitability of RBF kernel-based SVR and advanced tree-based models for real estate price prediction applications. These results provide valuable insights for both academic research and practical implementation in the real estate industry, offering a foundation for future work in model optimization and feature engineering within the domain of housing price prediction.

References

[1] Pace, R. K., & Barry, R. (1997). *Sparse spatial autoregressions*. Statistics & Probability Letters, 33(3), 291-297.

[2] Xiao, E. (2024). *Comprehensive K-Means Clustering*. Journal of Computer and Communications, 12, 146-159.

[3] Kwatra, S., Monreale, A., & Naretto, F. (2024). *Balancing Act: Navigating the Privacy-Utility Spectrum in Principal Component Analysis*. In Proceedings of the 21st International Conference on Security and Cryptography (SECRYPT 2024), pages 850-857.

[4] Chen, A. (2024). *Deep Learning in Real Estate Prediction: An Empirical Study on California House Prices*. The National High School Journal of Science Reports, 1-13.

[5] Chen, Y., Ulster University. (2022). *Analysis and Forecasting of California Housing*. Highlights in Business, Economics and Management MSIE 2022, Volume 3 (2023), 128-135.