Introduction to Data Science

22KDL

Lab02 - Regression

Deadline: **23h59 - 24/04/2024**
TA contact info: hduc.lee@gmail.com

Objective:
To understand and implement a linear regression model for real estate valuation using a dataset collected from Sindian District, New Taipei City, Taiwan.

Task description:
You are provided with a dataset (link) containing various features related to real estate properties in Sindian District. Your task is to build a linear regression model to predict the house price of a unit area based on the given features.

Requirements:

Data Exploration:
- Load the dataset into a suitable data structure (e.g., pandas DataFrame).
- Explore the dataset to understand its structure, feature types, and basic statistics.
- Check for missing values and handle them appropriately if necessary.

Data Preprocessing:
- Normalize the numerical features if required to ensure all features are on the same scale.
- Encode categorical features if any (in this dataset, there are no categorical features).
- Split the dataset into features (X) and the target variable (Y).

Feature Selection/Engineering:
- Analyze the correlation between features and the target variable.
- Select relevant features that are highly correlated with the target variable.
- Consider adding new features if they could improve model performance (e.g., feature combinations).

Model Training:
- Split the dataset into training and testing sets (e.g., 80% for training, 20% for testing).
- Initialize a linear regression model (e.g., using scikit-learn).
- Train the model using the training data.

Model Evaluation:
- Evaluate the trained model using appropriate metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared.

- Visualize the actual vs. predicted house prices to understand the model's performance visually.

Hyperparameter Tuning (Optional):

- Experiment with different hyperparameters of the linear regression model (if any) or other variants of linear model, regularization, etc to improve performance.
- Utilize techniques like cross-validation, grid search to find the optimal hyperparameters.

Conclusion and Further Analysis:

- Summarize the findings from the model evaluation.
- Discuss any limitations or assumptions made during the modeling process.
- Propose potential avenues for further analysis or model improvement.

Encouragement for Experimentation:

- Encourage students to experiment with different techniques for feature selection, model evaluation, and hyperparameter tuning.
- Encourage collaboration and discussion (BUT NOT **Plagiarism**) among students to share insights and learn from each other's approaches.
- Encourage students to leverage platforms like Google Colab/ Kaggle Notebook to experiment with the implementation. Google Colab provides free access to resources (CPU/GPU), facilitating faster experimentation.
- Encourage students to seek help from teaching assistants during lab sessions if they encounter difficulties.

Plagiarism Warning:

- Students are strictly prohibited from copying or reproducing the solution code from their peers. Each submission must be the individual work of the student. **Any instances of plagiarism or copying will result in a grade of 0 points for the assignment.**

Submission Guidelines:

- Jupyter Notebook containing:
  - Data loading and preprocessing steps.
  - Feature selection/engineering.
  - Model training and evaluation.
  - Conclusion and further analysis.
  - Visualizations (e.g., scatter plots, regression plots) demonstrating model performance and feature importance (if possible).
  - A brief report summarizing the methodology, results, and conclusions.
- Please send me your work before the due date.
- You can download the jupyter-notebook file (*.ipynb) by the following steps:
  - *File -> Download -> Download .ipynb*
- Name your notebook by the following pattern (same for Google Colab notebook's title): DS2024_Lab<LabID>_<StudentID>_<StudentName>.ipynb.
  - Example: DS2024_Lab01_21280075_NguyenVanA.ipynb
- The code results have to be printed out in the notebook.

- Include comments explaining key parts of the code if possible.
- Submit the notebook at: https://forms.gle/5qZyDbuRFxMfDdar5

There is **NO** acceptance for **cheating** or **copying**.

References:

1. https://utsavdesai26.medium.com/linear-regression-made-simple-a-step-by-step-tutorial-fb8e737ea2d9
2. https://www.kaggle.com/code/tanmayunhale/feature-selection-pearson-correlation#Breast-Cancer
3. https://www.kaggle.com/discussions/questions-and-answers/389354#2162654
4. https://scikit-learn.org/stable/modules/cross_validation.html