Naïve Bayes Documentation

Necessary libraries:

- Os
- Numpy
- Matplotlib
- Pandas

Dataset obtained from https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/overview

User input is used to determine five categories pertaining to certain data fields in our dataset: SalePrice, SaleCondition, SaleType, YrSold, and MoSold. SalePrice in particular is considered the cutoff estimate for the program. This allows the user to choose different combinations of data field options to determine how the other four categories impact the sale price prediction.

The algorithm/function utilizes data from the training set (train.csv) provided by our dataset for its probability calculations. It follows the naïve bayes formula for determining probability by assuming features (categories/data fields) are conditionally independent.

$$\underbrace{P(c \mid x)}_{\text{Posterior Probability}} = \frac{\overbrace{P(x \mid c)}^{\text{Likelihood}}\,\overbrace{P(c)}^{\text{Class Prior Probability}}}{\underbrace{P(x)}_{\text{Predictor Prior Probability}}}$$

$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

The goal is to predict the probability of a sale price to be >= or < the cutoff, given four data fields. This is done by:

- Counting number of items in “SalePrice” in the training set that are >= and < the cutoff, then calculating their probabilities: prob(>=) and prob(<).
- Go through a category and its respective “SalePrice”, count the number of items that match the user input AND are >= the cutoff, same with < the cutoff.

- Calculate the probability of the category given >= and the probability of the category given <, then multiply to prob(>=) and prob(<).
- Repeat for every category to get two probability ratios: prob(>=|X) and prob(<|X).

The function then returns those two probabilities and then compares them. If prob(>=|X) > prob(<|X) then the sale price is predicted to be >= the cutoff, and vice versa.

Lastly, the program produces a bar chart of all the distributions of every category based on the cutoff to better visualize the data.