National University of Singapore
School of Computing
CS3244: Machine Learning
Tutorial 06

Regression Metrics and Data Processing

1. **Regression Evaluation Metrics**

   In this problem, we discuss the regression metric, Mean Absolute Percentage Error (MAPE). MAPE is a measure of prediction accuracy of a forecasting method. Mathematically, it is expressed by a ratio defined by:

   $$MAPE = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

   where y_i and $\hat{y}_i$ denote the actual values and forecast/prediction values at data point i. Here, n denotes the number of data points in total.

   (a) Suppose that $y = [1, 2, 3, 4, 5]$ and $\hat{y} = [1, 2.5, 3, 4.1, 4.9]$. Calculate the MAPE of this prediction.

   (b) Assume that you are a supply-chain manager. You are using MAPE to judge your regression forecasts about **product demands** for the next month. We list them here. Discuss whether the listed MAPE shortcomings below will or will not affect you.

      i. Data with zeroes or close to zeroes.
      ii. Heavier penalty when predictions are higher than actual data.
      iii. Assumption that zero in the unit of the data has a special meaning.

   (c) Finally, we define an alternative to MAPE, which is SMAPE (Symmetric-MAPE). We define SMAPE as follows.

   $$SMAPE = \frac{100\%}{n} \sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{(\hat{y}_i + y_i)/2}$$

   Analyze how SMAPE fixes MAPE's problems.

2. **Curse of Dimensionality**

   (a) **Feature Selection (Wrapper):** Observe figure 1 regarding Recursive Feature Elimination (RFE) closely and answer the questions below.

      i. Describe the general trend in the graph.
      ii. Is this graph theoretically possible if we implement the high-level RFE algorithm described in the lecture? Explain your answer.

   (b) **Feature Selection (Filter):** Consider the following correlation matrix shown in figure 2 about cell nucleus data for breast cancer patients. There are six features and their correlations within each other.

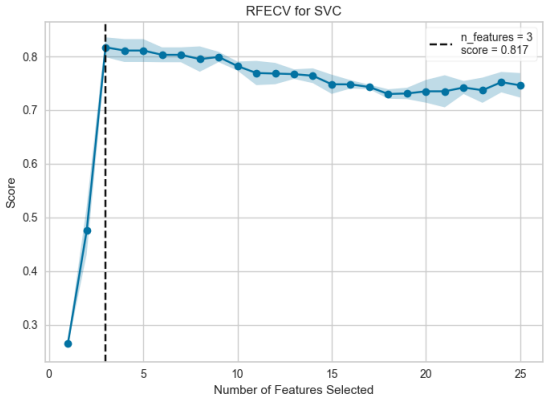

Figure 1: Recursive Feature Elimination - Image Credits

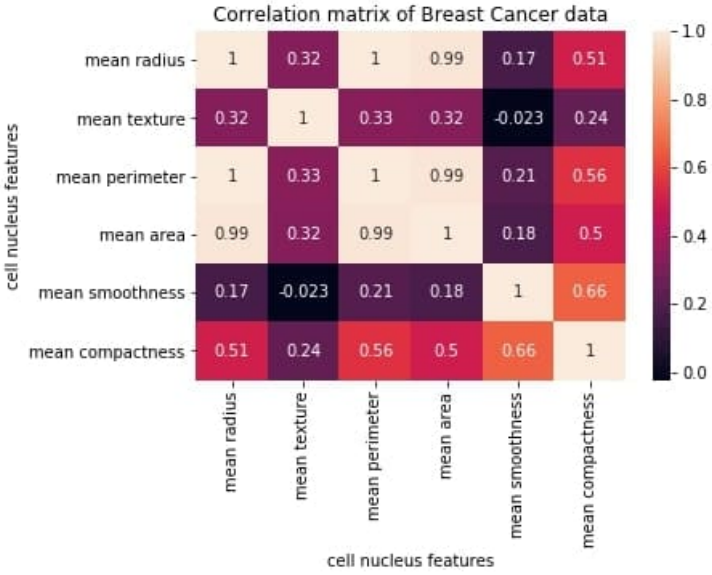

Figure 2: Correlation Matrix - Image Credits

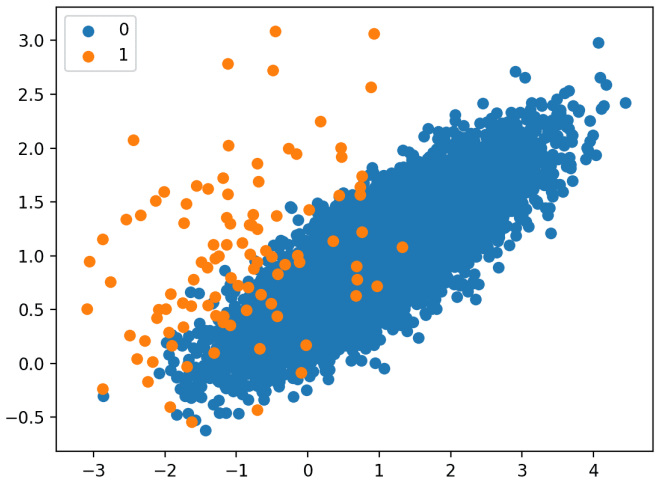

Figure 3: Scatter Plot of Imbalanced Classes - Image Credits

i. Which feature(s) should we remove from the table to avoid redundant information?

ii. There are some data with 1 correlation. Is this a coincide?

3. **Data Resampling Techniques**

   You are deciding on which data resampling methods to use for each of the following *imbalanced* datasets. Which of the data resampling methods should be applied? Briefly explain when and how the method(s) can be used in tandem with train-test split.

   (a) Dataset of 2 class labels. 80% majority class and 20% minority class.

   (b) Dataset of 3 class labels and discrete and continuous features. 45% majority class and 55% from minority classes.

   (c) Dataset with continuous output variable.

4. **Data Resampling: SMOTE**

   Refer the general Steps of SMOTE given to you below.

   1. From all the data points of your minority class, pick a random point.
   2. Find the k nearest neighbours to that point.
   3. Pick one of the neighbours randomly, now we have a pair from the minority class.
   4. Draw an imaginary line between the pair and pick a random point along the line.
   5. The new random point is added to the minority class.

   Figure 3 is a scatter plot of an imbalanced dataset to be used in a binary classification problem. Roughly illustrate how the transformed dataset will look like after SMOTE.