National University of Singapore
School of Computing
CS3244: Machine Learning
Solution to Week - 08 - Tutorial 06

Data Processing and Feature Engineering

1. **Curse of Dimensionality**

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

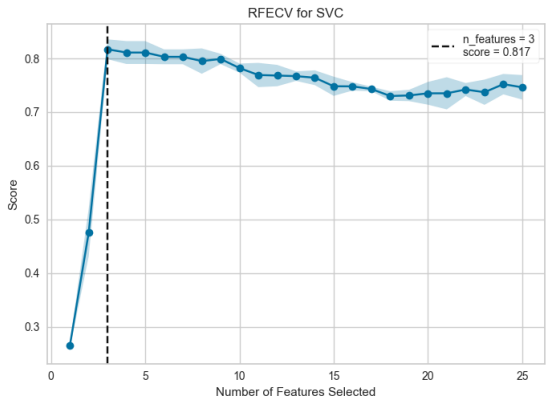

Figure 1: Recursive Feature Elimination - Image Credits

   i. Describe the general trend in the graph.
   The graph peaked when only the top 3 features are considered. The score slowly decreases as additional features are added. This is due to the fact that the latter features do not provide any further information. It is possible that they are a source of noise, explaining the decrease in the score as more features are added.

   ii. Is this graph theoretically possible if we implement the high-level RFE algorithm described in the lecture? Explain your answer.
   The graph is theoretically possible. Even though RFE is to eliminate the features with least decrease, that does not mean that it has to be monotonically increasing. RFE retrains the model each time after a feature is removed. As such, the next feature to be removed (new least decrease) can always have either a net positive or negative impact on performance after retraining.

   (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.

   i. Which feature(s) should we remove from the table to avoid redundant information?
   We see that all three of mean radius, mean perimeter, and mean area are strongly correlated, with value very near to 1. Hence, we need to only keep one of them. The feature we end up keeping will be not very relevant, as shown below.

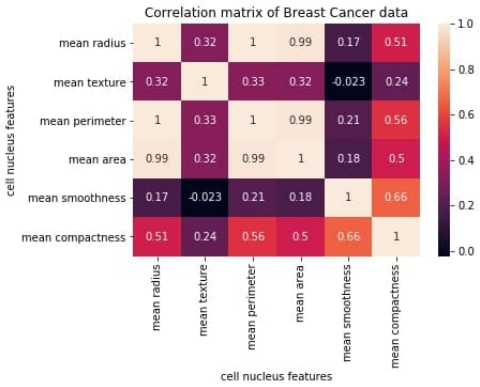

Figure 2: Correlation Matrix - Image Credits

ii. There are some data with 1 correlation. Is this a coincide?
Note that there are a relation between radius r and area A of a circle, $A = \pi r^2$ and also a relation between radius r and perimeter P with $P = 2\pi r$. Hence, we should expect that the correlation is very high.

2. **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.
Oversampling and SMOTE are plausible methods. Undersampling should not be used as we lose 75% of the majority class, which makes up 60% of the overall dataset. Data resampling methods are always done after the train-test split, and only on the training set.

(b) Dataset of 3 class labels and discrete and continuous features. 45% majority class and 55% from minority classes.
Here, the plausible methods depends on the composition of the 55% minority. If the minority classes are split 30%-25%, all three methods are applicable. However, if the minority classes are split 40%-15%, undersampling should be avoided for the same reason as in (a). Extra care also has to be taken to determine if there are any logical errors when using SMOTE as we have discretized features. If the features are intended to be strictly categorical variables, SMOTE should be avoided due to its interpolation.

(c) Dataset with continuous output variable.
After train-test split, we assign the training instances to bins based on their output values and plot a simple histogram. With dataset domain knowledge, whether resampling is required or not is determined. If required, we choose the resampling method based on the overall distribution of the bins and the type of the dataset features. When is undersampling preferred over oversampling?

3. **Data Resampling: SMOTE**

Refer the general Steps of SMOTE given to you below.

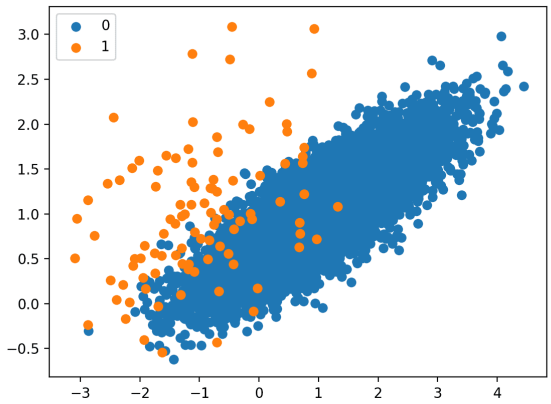

Figure 3: Scatter Plot of Imbalanced Classes - Image Credits

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.

Taking k to be the default value of 5, we can find out all combinations of the minority class pairs. Add points along the lines connecting each minority class pair and we would get a result similar to what we see in Figure 4 after SMOTE. (We will get varying results as k changes)

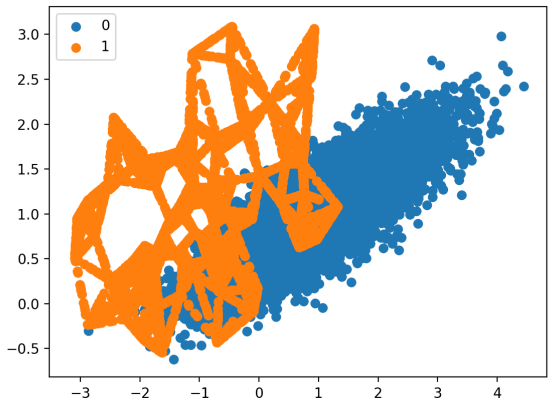

Figure 4: Scatter Plot after SMOTE transformation - Image Credits

4. **Feature Engineering (Image features, and Text Features)**

(a) Alice is building an animal image classification system. The images need to be classified into one of the five different dogs. All the input images are of size 256x256x3 ($H \times W \times C$, where H is the height, W is the width, and C is the number of channels). She makes a vector(size 196608) of the image by flattening it. Describe two key issues if we use this vector as input to our learning model?
One key issue is that number of dimensions is high. Hence fitting it to a function needs a highly complex model. It results in a long training time. The second issue is that by flattening, we throw the spatial relationship of images. The loss of spatial relationships makes it harder to find patterns. (In week 10, you will study Convolutional Neural Networks.)

(b) Alice is given with a feature extractor which outputs 100 dimensional feature vector for every image. Alice uses this new vector to build the machine learning model. However, the model performance is poor. Describe possible reasons for this poor performance. Recommend some techniques where Alice can improve the classification performance.
The followings are possible reasons.

i. 100 *features may not be enough to represent the original images.*
ii. *The filter may produce features which may not be good for discriminating the dogs. For example, some color features may be bad at differentiating the dog classes.*

The poor performance can be mitigated using a number of ways. First, more features can be used instead of 100 features. i.e. using some other feature reduction techniques (PCA, LDA) can retain useful information in varying number of features. Secondly, using an ensemble of features will help to get a better model.

(c) TF-IDF(Term Frequency–Inverse Document Frequency) is a well-known extension of Bag of Words(BoW). Give two advantages and disadvantages of it.
Advantages

i. Simple and Easy to compute.
ii. A good way to measure the similarity between documents.

Disadvantages

i. It cannot capture semantics. Hence, it may not be useful for natural language understanding.
ii. It cannot capture the position of words.