教育數據探勘與應用 #3

陸嘉康 110590045

8.12 Threshold: 0.5

Tuple 1: P (0.95 >= 0.5) → TP
Tuple 2: N (0.85 >= 0.5) → FP
Tuple 3: P (0.78 >= 0.5) → TP
Tuple 4: P (0.66 >= 0.5) → TP
Tuple 5: N (0.60 >= 0.5) → FP
Tuple 6: P (0.55 >= 0.5) → TP
Tuple 7: N (0.53 >= 0.5) → FP
Tuple 8: N (0.52 >= 0.5) → FP
Tuple 9: N (0.51 >= 0.5) → FP
Tuple 10: P (0.40 < 0.5) → FN

True Positives (TP): 4
False Positives (FP): 5
True Negatives (TN): 0
False Negatives (FN): 1

TPR: 4 / (4 + 1) = 0.8
FPR: 5 / (5 + 0) = 1

8.16 Balance the training set,
Oversampling: Oversample the minority class.
Under-sampling: Randomly eliminate tuples from majority class
Synthesizing: Synthesize new minority classes

At the algorithm level,
Threshold-moving: Move the decision threshold, t, so that the rare class tuples are easier to classify, and hence, less chance of costly false negative errors
Class weight adjusting: Since false negative costs more than false positive, we can give larger weight to false negative
Ensemble techniques: Ensemble multiple classifiers introduced in the following chapter

I would use oversampling to balance the data set, and use class weight adjusting to find potential fraudulent cases.Those reported cases could then be manually reviewed by human.

9.4 Eager Classification:

Advantages:

- **Efficiency during prediction**: Eager classifiers construct a model during the training phase. Prediction for new instances is typically faster as the model is already built.
- **Regularization**: Eager classifiers often have built-in regularization mechanisms to prevent overfitting.

Disadvantages:

- **Static model**: Once trained, the model is static and doesn't adapt to changes in the data distribution without retraining.
- **Resource intensive training**: Training eager classifiers can be computationally expensive.

Lazy Classification:

Advantages:

- **Adaptability**: Lazy classifiers adapt to changes in the data distribution without retraining. They are suitable for dynamic or evolving datasets.
- **No upfront training cost**: Lazy classifiers have no upfront training cost. Prediction time is the only time cost, and it depends on the size of the training dataset.

Disadvantages:

- **Computational cost during prediction**: Prediction time for lazy classifiers can be higher because they need to compute distances or similarities during prediction.
- **Sensitivity to noise**: Lazy classifiers can be sensitive to noise and irrelevant features in the dataset.