VSB TECHNICAL UNIVERSITY OF OSTRAVA | FACULTY OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE | DEPARTMENT OF COMPUTER SCIENCE

Machine Learning

Advanced Concepts

Jan Platoš

November 22, 2023

Department of Computer Science
Faculty of Electrical Engineering and Computer Science
VŠB - Technical University of Ostrava

Advanced Concepts

Advanced concepts

- Real world classification may bring nonstandard scenarios and special cases.
- Many of them was successfully addressed by different strategies.
- In general, these cases may be divided into two groups:
    - Difficult classification scenarios - multi-class classification, rare-class scenarios, managing very large training data, etc.
    - Enhancing classification - improving classification with various data-centric or user-centric approaches.

- Many classification models are naturally designed for binary class problems (SVM, neural networks, logistic regression).
- Multi-class generalization for these methods exists in algorithm-specific scenario.
- General meta-framework that take any binary classification algorithm $\mathcal{A}$ as inputs and allow multi-class classification is very useful.
- Several strategies were developed in the past.
- The k will denote number of classes in the data.

Advanced concepts - Multi-class learning

One-against-rest

- k different binary classification problems are created.
- One problem correspond to each class.
- Each i-th problem, i-th class is considered as positive case (+1) and all the remaining classes as the negative case (-1).
- The binary classifier $\mathcal{A}$ is applied to each of these training data sets.
- This results in k different models.
- The test instance is processed by each model.
- When a positive class is predicted, this class is rewarded with a vote.
- When a negative class is predicted, all classes are rewarded with a vote.
- The class with highest number of votes is predicted as a relevant one.
- The problem, when two or more classes have the same number of votes may be solved using numeric output/confidence of the classifier.

Advanced concepts - Multi-class learning

One-against-one

- $\binom{k}{2}$ different binary classification problems are created fo each pair of classes.
- The binary classifier $\mathcal{A}$ is applied to each of these training data sets.
- This results in $\frac{k\cdot(k-1)}{2}$ different models.
- For each model, prediction provides a vote for the winner.
- The class with highest number of votes is declared as the relevant one.
- The problem, when two or more classes have the same number of votes may be solved using numeric output/confidence of the classifier.
- This approach is usually less computationally expensive than the previous (when the time complexity increases super-linearly with the number of points).

Advanced concepts - Rare class learning

- The distribution of the classes in the dataset is usually not balanced (e.g. credit card frauds, network attacks, quality prediction).
- The rare classes are hidden under more frequent majority class/classes.
- Due to rareness of the class, even bad classifier produces very high accuracy models.
- The accuracy on the rare class is typically very low or zero.

$$P(c = Rare) = 2\% \Rightarrow min.\ accuracy = 98\%$$

- The misclassification of the rare class are much higher as a consequence.

Advanced concepts - Rare class learning

- The rare classes need to be emphasized due to its greater importance.
- This may be solved using *cost matrix* between classes of *cost* coefficient to each class.
- Example re-weighting
  - The samples are re-weighted according the misclassification costs.
  - Classification algorithm need to be adapted to work with weighted examples.
- Example re-sampling
  - The samples from different classes are re-samples to under-sample normal classes and oversample rare classes.
  - The classifiers need not to be modified.

Questions?