Dealing with Class Imbalance

Lesson 7 – Section 3

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Distributions Matter…

Because the Internet is all about cute kittens

Resulting in highly skewed distribution in training set

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

The Class Imbalance Problem I

Balanced Data set:

- Approximately positive examples = negative examples

Some domains do not have balanced data sets

- Examples:
- Helicopter Gearbox Fault Monitoring
- Discrimination between Earthquakes and Nuclear Explosions
- Document Filtering
- Detection of Oil Spills
- Detection of Fraudulent Telephone Calls

The Class Imbalance Problem II

Standard learners biased toward majority class

- Classifiers attempt to global quantities (like error rate) regardless of the data distribution
- Examples from main class are well classified
- Examples from minority class are misclassified

$$LOSS = \sum_{i=1}^{n} (y_i - f_\theta(\mathbf{x}_i))^2 = \sum_{i=1}^{n_{pos}} (y_i - f_\theta(\mathbf{x}_i))^2 + \sum_{i=1}^{n_{neg}} (y_i - f_\theta(\mathbf{x}_i))^2$$

If $n_{neg} >> n_{pos}$, the LOSS function benefits more on making the negative cases accurate, than on making the positive cases accurate

Some Generalities

Standard accuracy/error rate does not catch class imbalance.

–Use ROC Analysis instead

3 ways to deal with class imbalances:

- Re-sampling
- Re-weighing
- One-class Learning (see SVMs)

SMOTE: A state-of-the-art Resampling Approach

SMOTE: Synthetic Minority Oversampling Technique

–Designed by Chawla, Hall, & Kegelmeyer in 2002

- Combines informed **oversampling** of the **minority** class with random **undersampling** of the **majority** class.

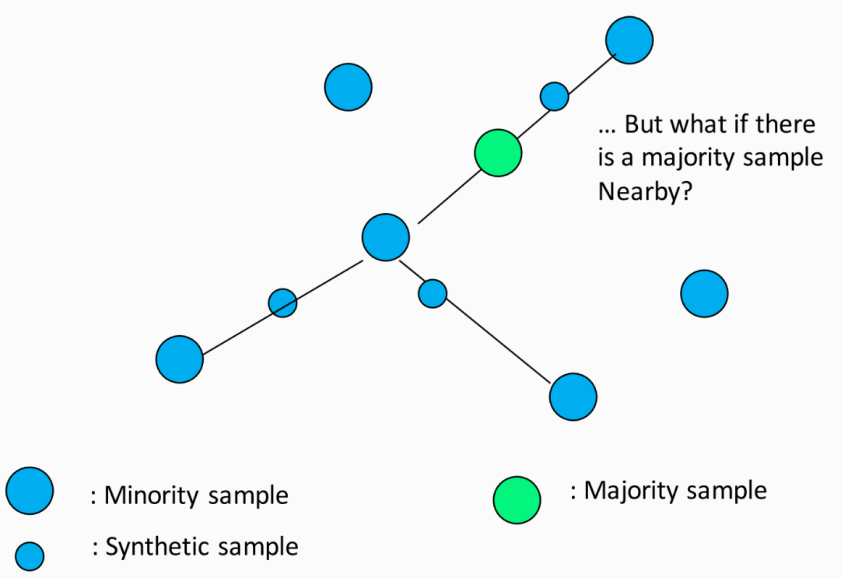
PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

SMOTE's Informed Oversampling Procedure

For each minority sample:

- Find its k-nearest minority neighbors
- Randomly select j of these neighbors
- Randomly generate synthetic samples along the lines joining the minority sample and its j selected neighbors.
  - j depends on the amount of oversampling desired

SMOTE'S Informed Oversampling Procedure I



SMOTE'S Shortcomings

Overgeneralization

- Blinding generalizes minority without regard to majority class
  - In highly-skewed class distributions, with sparse minority class, end up with greater chance of class mixture

Lack of Flexibility

- Number of synthetic samples is fixed
  - No flexibility in re-balancing rate

Summary

>Introduced unbalanced classification, and its impact on classification models

>Described SMOTE (Synthetic Minority Oversampling Technique) for oversampling minority class