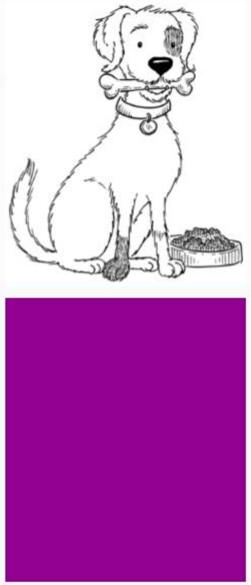
Machine Learning Techniques

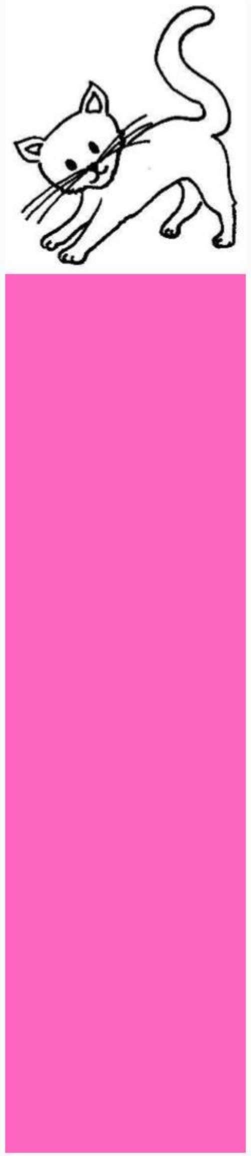
DATASCI 420

Lesson 02-2 Dealing with Class Imbalance

Distributions Matter…

Because the Internet is all about cute kittens





Resulting in highly skewed distribution in training set…

The Class Imbalance Problem I

- Data sets are said to be balanced if there are, approximately, as many positive examples of the concept as there are negative ones.
- Many domains that do not have a balanced data set.

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

- The problem with class imbalances is that standard learners are often biased towards the majority class.
- That is because these classifiers attempt to reduce global quantities such as the error rate, not taking the data distribution into consideration.
- As a result examples from the overwhelming class are well-classified whereas examples from the minority class tend to be 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

- Evaluating performance of a model on a class imbalance problem is not done appropriately with standard accuracy/error rate.
  - ROC Analysis is typically used, instead.

- There are three main ways to deal with class imbalances: re-sampling, re-weighing, and one-class learning (cover in SVMs)

- Re-sampling provides a simple way of biasing generalization process.
- It can do so by:
  - Generating synthetic samples accordingly biased
  - Controlling the amount and placement of the new samples

SMOTE: A State-of-the-Art Resampling Approach

- SMOTE stands for Synthetic Minority Oversampling Technique.
  - Technique designed by Chawla, Hall, & Kegelmeyer in 2002.

- It combines Informed **oversampling** of the **minority class** with **random undersampling** of the **majority class**.

- SMOTE currently yields the best results as far as re-sampling and modifying the probabilistic estimate techniques go (Chawla, 2003).

SMOTE’s Informed Oversampling Procedure II

For each minority Sample

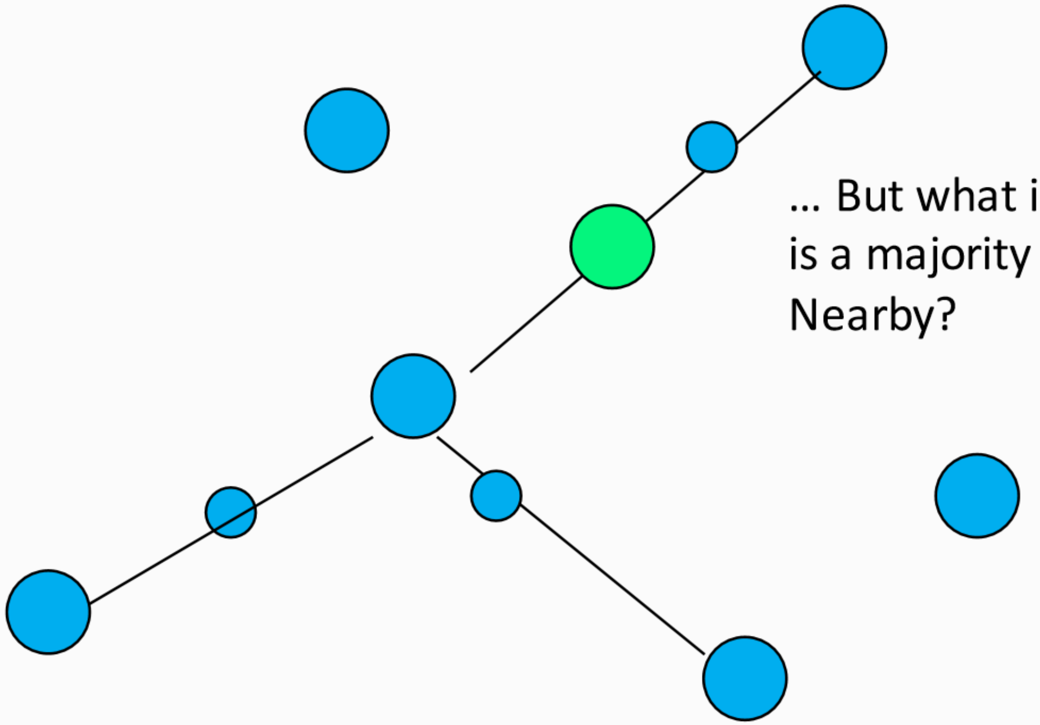
- Find its k-nearest minority neighbors
- Randomly select j of these neighbors
- Randomly generate synthetic samples along the lines joining the minority sample and its j selected neighbors

  (j depends on the amount of oversampling desired)

SMOTE’s Informed vs. Random Oversampling

- Random Oversampling (with replacement) of the minority class has the effect of making the decision region for the minority class very specific.
- In a decision tree, it would cause a new split and often leads to overfitting.
- SMOTE’s informed oversampling generalizes the decision region for the minority class.
- As a result, larger and less specific regions are learned, thus, paying attention to minority class samples without causing overfitting.

SMOTE’s Informed Oversampling Procedure I



… But what if there is a majority sample Nearby?

● : Minority sample

● : Synthetic sample

🟢 : Majority sample

SMOTE’s Shortcomings

- Overgeneralization
  - SMOTE’s procedure can be dangerous since it blindly generalizes the minority area without regard to the majority class.
  - This strategy is problematic in the case of highly skewed class distributions since, in such cases, the minority class is very sparse with respect to the majority class, thus resulting in a greater chance of class mixture.

- Lack of Flexibility
  - The number of synthetic samples generated by SMOTE is fixed in advance, thus not allowing for any flexibility in the re-balancing rate.