Positive-only semi-supervised classification

UCSD Data Mining Contest

Given unlabeled two-class training data and a few points belonging to one class, correctly classify a test set

Problems involved:

- 20-dimensional data – difficult to visualize
- 68,500 unlabeled training points
  - Any n^2 algorithm is out of the question
  - Just 60 positive points given (~0.1 % labeled)
  - No negative points!
- Real-world data from a physical experiment
- Roughly $1/8^{th}$ of the data is positive (inferred from quiz set results).

Progress measured by F1-score against a given quiz set

F1 = harmonic mean of precision and recall

= 2PR/(P+R)

2/F1 = 1+(All-TN)/TP

Initial plans of attack

- **Clustering**
  - see if there are clearly visible clusters of points
  - spectral clustering with local scaling can recover intricate patterns
- **SVD** and **visualization** to capture pattern in data
  - reduce dimensionality
- **Co-training**

  Make an assumption that two disjoint sets of features are sufficient for independently learning the concept. Use one classifier to suggest labels for the other iteratively
- **Learn a distance metric**

  With 20-dimensional data simple Euclidean distance may be snuffing out important differences in the data
- **Bayesian inference**

  Estimate independent feature-wise probability functions for given positive data and use Bayes Rule to infer a function for negative data

$$P(-|x) = \frac{P(x) - P(x|+)P(+)}{1 - P(+)}$$

- **Ranking**

  Rank the points by distance or another measure from positive points

The (Harsh?) Reality :)

- **Bayes**: Paucity of positive examples led to massively negative probabilities.

- **Clustering**: Most algorithms not feasible at this scale. Clustering on subsamples yielded no meaningful results.

- **SVD** and **dimensionality reduction**: No easily discernible separation or pattern distinguishing positives from the rest in top half of significant dimensions.

- **Sampling Negatives**: Needed for any conventional classifier.
  - Based on distance from positives and distance from each other.
  - Based on voting by different classifiers.

- **SVM**: Used seed negatives as those farthest from positives, and iteratively trained SVM classifiers adding to training set at each step.
  - Later used Non-linear kernels
  - Marginally improved results

Reality contd..

- **Nearest neighbor techniques**:
  NN can help detect a non-linear manifold structure in the data. Most helpful so far. Several variants
  - **one-at-a-time**
  - **batches**

    (Best results yet – in terms of F1 score)
  - **average distance**
  - **Clustering and then growing each clusters**

- **One-class SVM**:

  Used the libsvm implementation.

  Iteratively added positives to the given positive examples.

- **Hybrids**: Combining information from several techniques above.

Results

Quiz data: 11427 unlabeled examples

Method	Number of Positives submitted	Fraction of True Positives
NN one-at-a-time	1700	0.32
NN batch	4005	0.26
NN average	4020	0.12
NN 6 dimensions	1014	0.37
One-Class (RBF)	1307	0.31

Future Work!

- Trying other classifiers such as ANNs (e.g. one-class classifier using ANN instead of SVM)
- Identify a subset (might be large but not too large)

  covering all the positive examples,

  then Refine e.g. using Clustering
- Combining the approaches and utilizing the results to improve.
- If we knew anything more, we would have tried it :)

Sample Data in Reduced Dimensions