Learning with Deficient Supervision

CS771: Introduction to Machine Learning

Purushottam Kar

Outline of Discussion

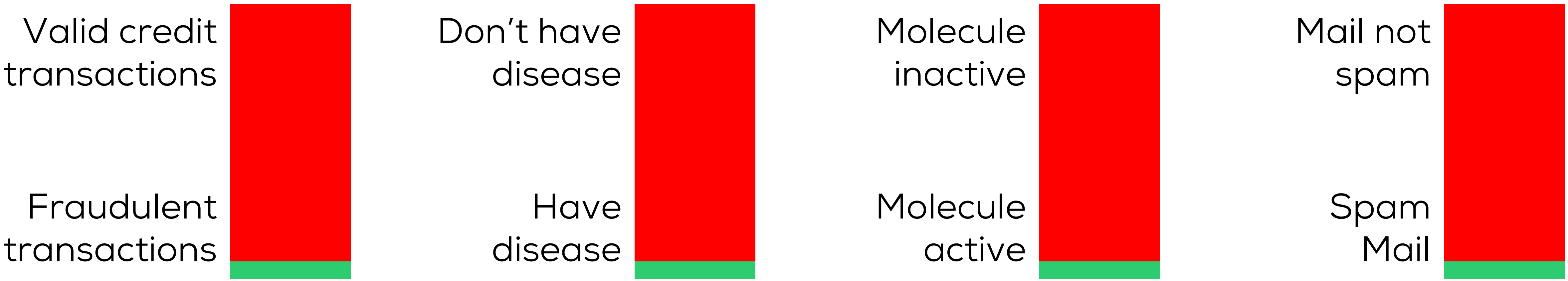
- How to perform supervised learning with a lazy teacher
- … may remind you of the current offering of CS771A
- Self-deprecating jokes aside, quite important in ML
- Learning with label imbalance
  - Fully labelled data but very less data for some/all classes
- Learning with weak supervision
  - Fully labelled data but "high-level" supervision
- Active learning
  - Fully unlabelled data but labels available on request
- Semi-supervised learning
  - Partly labelled data and no requests allowed

Increasing challenge

Learning with Label Imbalance

Label imbalance is very common

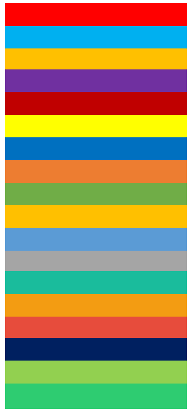
- Binary classification

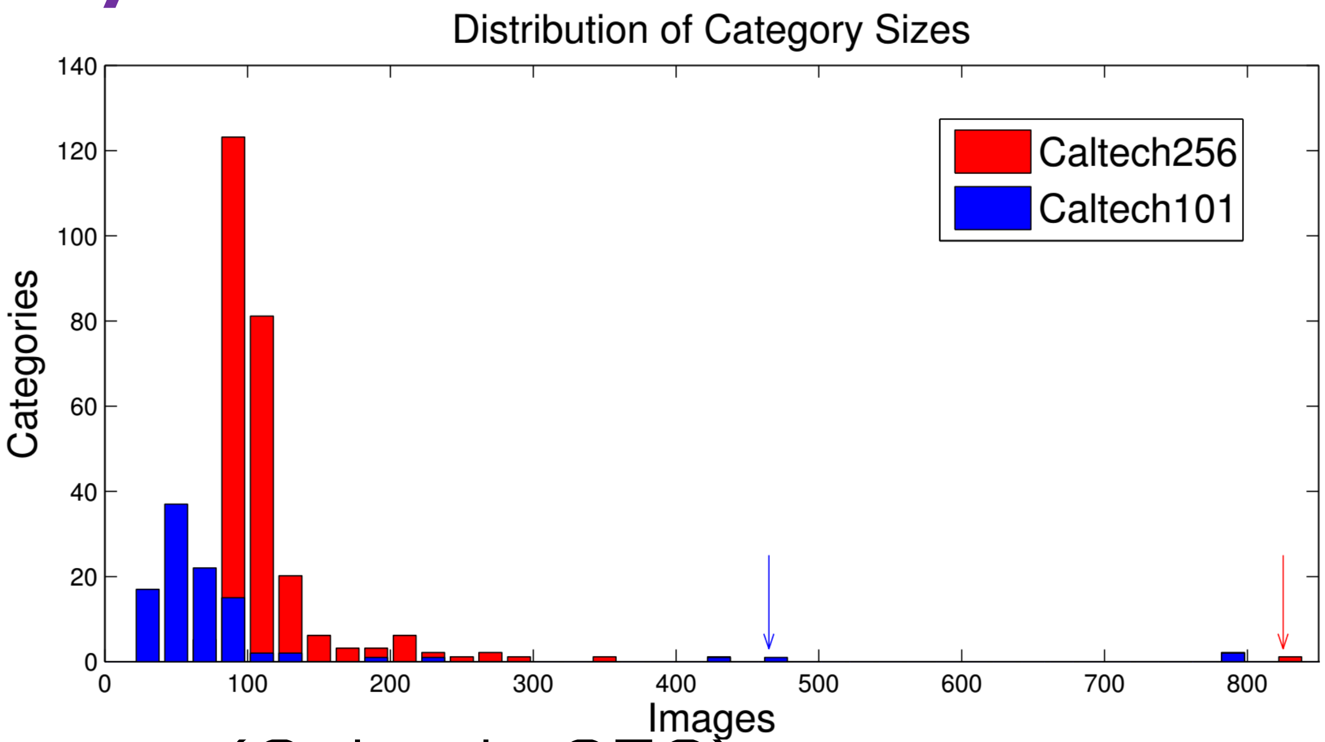


- In anomaly detection, medical diagnosis, drug discovery, spam filtering and many other domains, data is heavily biased
- Lots of data may be available but as low as 0.01% may belong to the "rare" class e.g. 100000 red vs 10 green

Label imbalance is very common

- Multi-classification

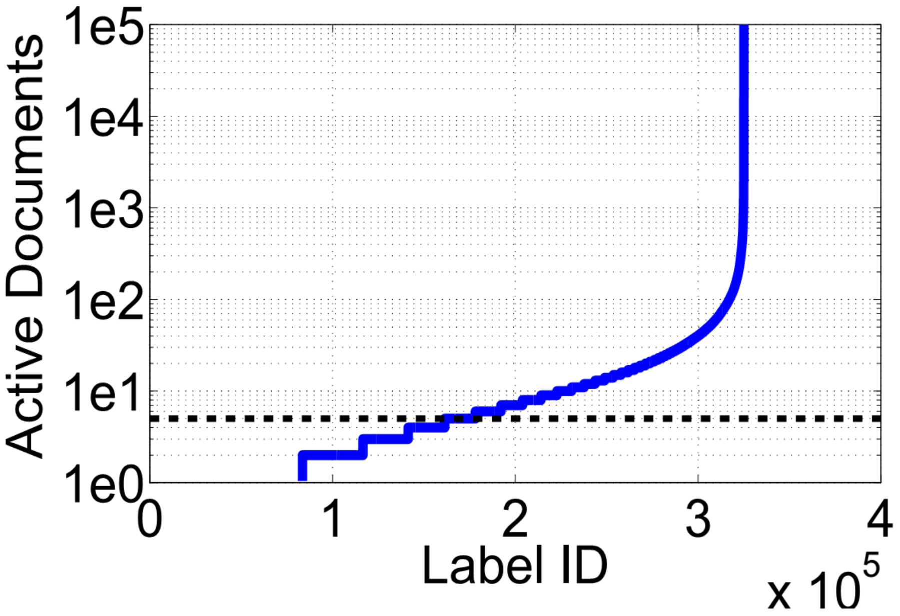




- Object detection with 256 classes (Caltech-256)
- … i.e. 50% of classes must have each less than 0.8% of the data
- Lots of data for few classes but most classes impoverished
- Rarest class has only 80 images, most popular has 800 images

Label imbalance is very common

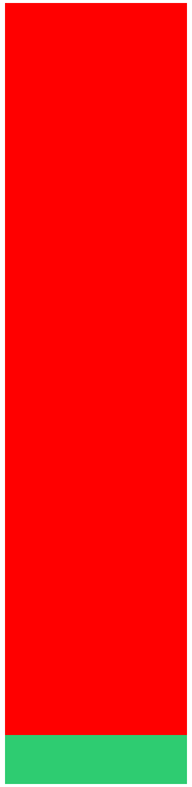
- Multi-label classification and recommendation



0	0	0	0	0	0	0	0	0	1	0	0
0	1	0	1	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	1	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	1	0	0	0	0	0	0	0	0

- Wikipedia LSHTC 300000 labels of which 150000 occur only in less than 5 documents. Most popular label in 100000 documents
- Most items in recommendation setting liked only by few people

Challenges with label imbalance



- When one class dominates data so much, it may dominate training and evaluation too
- If 99.99% data is red, a classifier that labels every data point as red gets 99.99% accuracy!
- Such a classifier is useless! Will let every spam mail through, and call every human healthy
- In the needle-in-a-haystack problem, calling everything hay is missing the point
- Even training is difficult – SVM may get tempted to minimize hinge loss only on red points
- Similar problems in multi class/label as well!

Solutions?

- Change the dataset
- Change the evaluation metric
- Change the learning algorithm
- In recent years, last two methods have become a single method since we now have advanced methods to train using even complex evaluation metrics
- Earlier people would train using simple metrics like hinge loss or logistic regression and then evaluate using something different
- First method is useful when changing algorithms or evaluation metric is not in our control
- For multi-label, refer to extreme classification literature – will not discuss that again here. Focus on binary classification now.

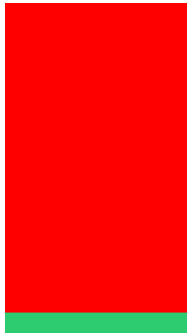
Dataset Modification

Dataset Modification

- Undersample the popular class (sample only 10 red points)

Dataset Modification

- Undersample the popular class (sample only 10 red points)



Dataset Modification

- Undersample the popular class (sample only 10 red points)

Dataset Modification

- Undersample the popular class (sample only 10 red points)

Dataset Modification

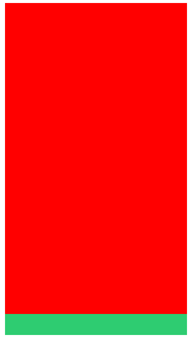
- Undersample the popular class (sample only 10 red points)

- Oversample the rare class (repeat each green point 9999 times)

Dataset Modification

- Undersample the popular class (sample only 10 red points)

- Oversample the rare class (repeat each green point 9999 times)



Dataset Modification

- Undersample the popular class (sample only 10 red points)

- Oversample the rare class (repeat each green point 9999 times)

Dataset Modification

- Undersample the popular class (sample only 10 red points)

- Oversample the rare class (repeat each green point 9999 times)

Dataset Modification

- Undersample the popular class (sample only 10 red points)

- Oversample the rare class (repeat each green point 9999 times)

Dataset Modification

- Undersample the popular class (sample only 10 red points)

- Oversample the rare class (repeat each green point 9999 times)

- Can get tricky to extend to multiclass/label settings

Dataset Modification

- Undersample the popular class (sample only 10 red points)

- Oversample the rare class (repeat each green point 9999 times)

- Can get tricky to extend to multiclass/label settings
- Undersampling throws away data, oversampling may slow training

Changing evaluation/training loss function

- Reweighted loss functions: add more weight to rare class points
- Similar effect to oversampling but dataset size remains same
- Classical SVM

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|_2^2 + C \cdot \sum_{i=1}^{n} \left[1 - y^i \cdot \langle \mathbf{w}, \mathbf{x}^i \rangle\right]_+$$

- Class-reweighted SVM

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|_2^2 + C_+ \cdot \sum_{i:y^i=+1} \left[1 - \langle \mathbf{w}, \mathbf{x}^i \rangle\right]_+ + C_- \cdot \sum_{i:y^i=-1} \left[1 + \langle \mathbf{w}, \mathbf{x}^i \rangle\right]_+$$

- C_+, C_- tuned. Often prop. to inverse of fraction of pos/neg. points
- **Exercise**: find the dual and kernelize the class-reweighted SVM
- Extends more gracefully to multiclass settings

Changing evaluation/training loss function

- More careful loss/evaluation functions used for critical apps
- Have already seen several of them in lecture
- **Precision**: of those predicted as green how many were green?
- **Recall**: of those actually green how many were predicted as green?
- Notice that completely disregards the red class since overpopulous
- No points for predicting red data points correctly as red!
- F-measure: harmonic mean of precision and recall
- Many other measures precision@k, nDCG, AUC, MRR
- Recent years have seen much progress in provable optimization of these complicated loss functions on huge datasets

Please give your Feedback

http://tinyurl.com/ml17-18afb