Group Learning in Active Learning

Zhipeng Luo & Dr. Hauskrecht
Dec. 11th
Subjects of Summary

- Stage 1: Basic Setting
  - General process
  - Baselines
- Stage 2: Which one to split?
- Stage 3: How to split?
- About ‘noisy’ data
For a pool of unlabeled training data:
- Initialize $K$ clusters ($K=10$), Query each cluster’s ratio of $-/+$, and softly label each instance by creating duplicates.
- Iterate until each cluster is pure (ratio=1 or 0)
  - Learn decision boundary using basic learner
  - Apply this boundary to test data and record AUC score.
  - Which one?
  - Choose one cluster to split
  - Split this cluster into two sub-clusters
  - Query the ratio of either one sub-cluster and update the other’s
  - Update soft labels of the affected instances

How to split?
Which one?
If we choose the cluster with the largest projection and split it roughly half-half by its projection center, then we have the yellow line.

**Baseline 1:** traditional uncertainty sampling method

**Baseline 2:** initialize labels by using cluster labeling but utilize uncertainty sampling on later phase.

Much fewer # of queries! Good and steady start!
Which one to split?

- Choose the cluster that has:
  - 1. largest projection variance (presented just now) or
  - 2. largest entropy or
  - 3. largest size

- **Note:** exps are based on ‘center’ in the 2\textsuperscript{nd} phase.
How to split it?

- Given the cluster we want to partition, we can split it by:
  - Using its projection center as boundary (presented just now) or
  - Using its medium projection value as boundary or
  - Using **Partition by Variance**
  - Using **Partition by Variance, after filtering noise by Gaussian Distribution**
  - Using **Partition by Ratio Variance**

- **Note:** exps are based on ‘projection’ in 1\textsuperscript{st} phase.

![ROC Curves](image)
How to split it?

Partition by Variance:
- Given the set of one-dimensional projection values of a cluster $S$,
- E.g. $\text{Variance}(S) = \{1.21, 2.32, 1.43, -1.90, 0.23, -1.23, -2.12, \ldots\}$
- We split it into 2 sub-clusters $A$ and $B$ such that:
  - $\text{Variance}(A) + \text{Variance}(B)$ is minimum.
- i.e. $\arg\min_{\{A, B\}} \sum_{x \in A} \frac{||x - u_A||^2}{\text{size}(A)} + \sum_{x \in B} \frac{||x - u_B||^2}{\text{size}(B)}$
- It’s NP-hard but I used an approximate algorithm:
  - Sort $\text{Variance}(S)$
  - Linearly try the boundary one by one
  - And find the boundary with the minimum sum of variances.

Mistakes I have made:
- 1. Rudely conduct K-means directly on $\text{Variance}(S)$
  - i.e. $\arg\min_{\{A, B\}} \sum_{x \in A} ||x - u_A||^2 + \sum_{x \in B} ||x - u_B||^2$
- 2. Forgot to sort $\text{Variance}(S)$ at first…

Partition by Ratio Variance:
- i.e. $\arg\min_{\{A, B\}} \sum_{x \in A} \frac{||x - u_A||^2}{\text{size}(A)^2} + \sum_{x \in B} \frac{||x - u_B||^2}{\text{size}(B)^2}$
About ‘noisy’ data

- Given the Variance($S$), its distribution is probably as follows:

- If we use Hist function to plot, it is shown as Gaussian Distribution
How to deal with the ‘noisy’ data

- After Yanbing’s presentation, we can apply 3-sigma rule to filter the noise data.

- Then, use ‘normal’ data remained to find the splitting boundary using **Partition by Variance** algorithm.

  *Best performance so far.*