Bootstrapping

Krishna Kalyan

Introduction

The objective of this exercise is to learn a Maximum Entropy (ME) model for the corpus f50 with fewer annotated examples but with the same accuracy than the ME model learned by using the whole training data.

Implementation

- Passive Learning

The program `passive-learner.py` wraps system calls methods that learn a model from the whole traning dataset and returns an accuray on the test dataset. These results will be used to compare results obtained from active learner. We obtained an accuracy of around `78` percent with this method considering that we trained it on all examples.

- Margin Sampling and Entropy Sampling

The program `classifier-probs.py` implements methods that use `entropy sampling` and `margin sampling` while using `Active Learning`.

$$x^* = argmin_x P(y_1/x) - P(y_2/x)$$

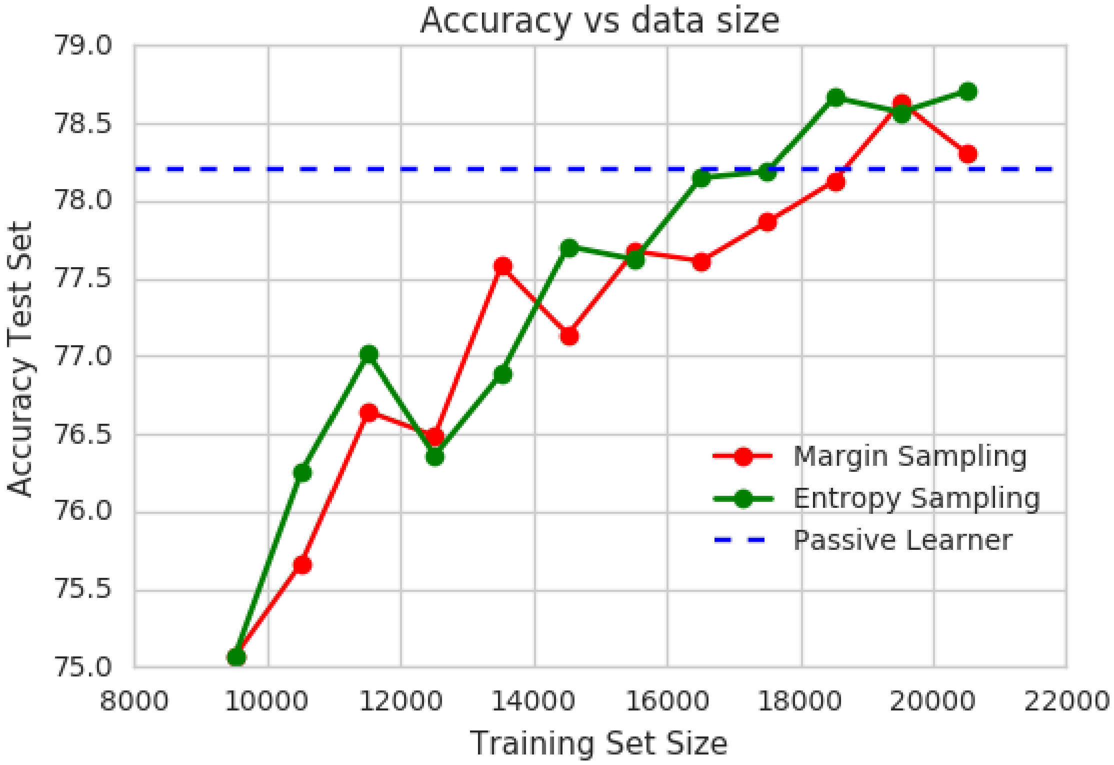
`Margin Sampling` is basically the difference between top two labels our observation. In this experiment we took top `k=1000` sorted examples that were obtaned by margin sampling.

$$x^* = argmax_x - \sum_{n=1}^{N} P(y_i/x) log P(y_i/x)$$

`Entropy Sampling` was implemented based on the formula above. In my program ,I decided to remove negative sign and minimise `x` so that the interface for `Entropy Sampling` and `Margin Sampling` remain same.

Conclusion

Based on our experiments we see that `Active Learning` with `Entropy Sampling` performs much better than `Passive Learning` with more examples. The performance of the `Active Learner` peaks at around `19514` training samples (`+- 1000`) which varies based on the type of sampler used for `Active Learning`. Maximum accuracy that was achieved with `Active Learning` was `79` around percent using a smaller subset of data.



Accuracy vs data size
79.0
78.5
78.0
77.5
77.0
76.5
76.0
75.5
75.0
Accuracy Test Set
8000
10000
12000
14000
16000
18000
20000
22000
Training Set Size
Margin Sampling
Entropy Sampling
Passive Learner