Active Learning: Definition and Concepts

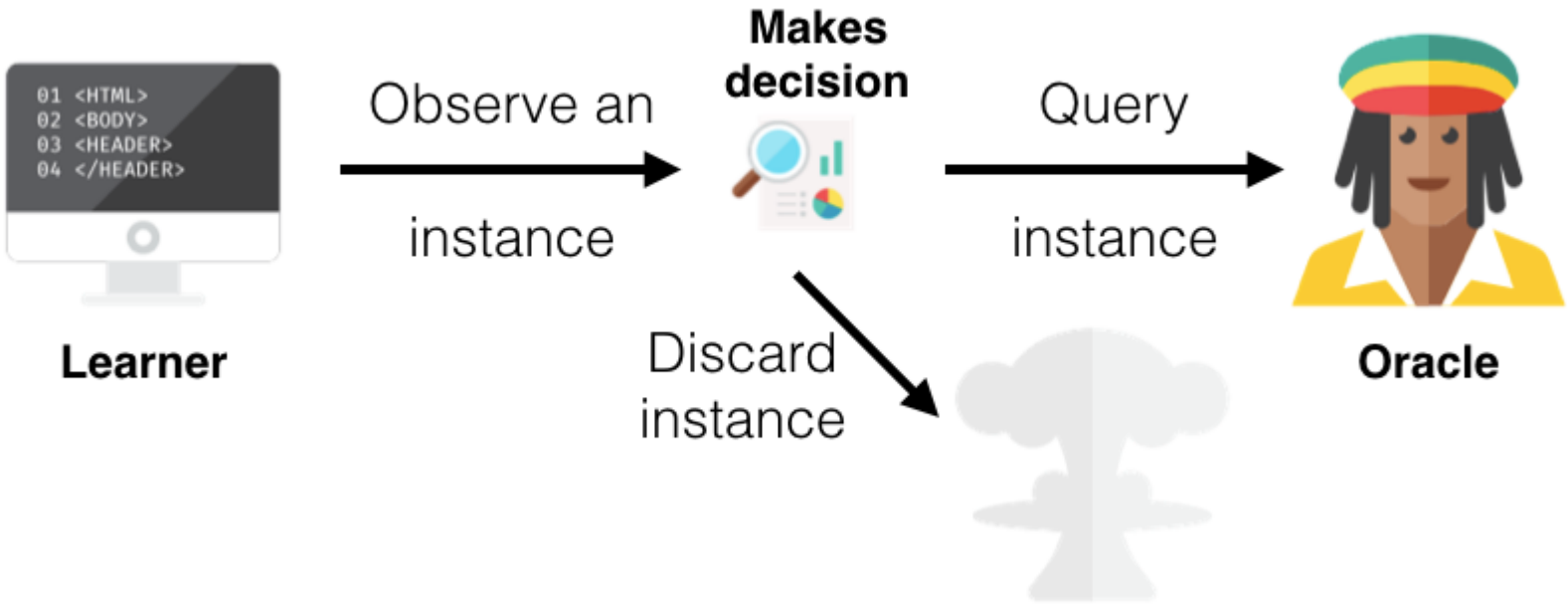
The main hypothesis in active learning is that if a learning algorithm can choose the data it wants to learn from, it can perform better than traditional methods with substantially less data for training.

Scenarios

- Pool-Based sampling

- Membership Query Synthesis

- Stream-Based Selective Sampling

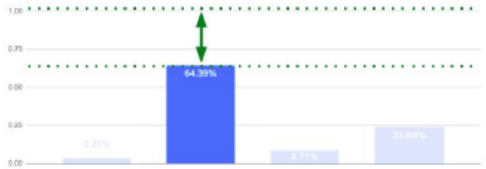


Query Strategies

Instances	Label A	Label B	Label C
d_1	0.9	0.09	0.01
d_2	0.2	0.5	0.3

- **Least Confidence (LC)**
- **Margin Sampling**
- **Entropy Sampling**

Least Confidence: *difference between the most confident prediction and 100% confidence*



$$\frac{n\,(1 - P_\theta(y^*_1 \mid x))}{n - 1}$$

```
most_conf = torch.max(prob)
num_labels = prob.numel()
numerator = (num_labels * (1 - most_conf))
denominator = (num_labels - 1)

least_conf = numerator / denominator
```

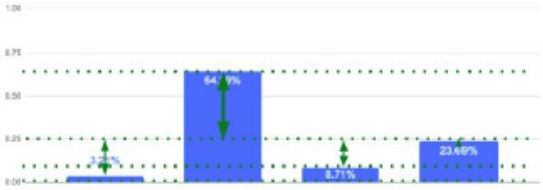
Margin of Confidence: difference between the top two most confident predictions

$$1 - (P_\theta(y^*_1 \mid x) - P_\theta(y^*_2 \mid x))$$

```
prob, _ = torch.sort(prob, descending=True)
difference = (prob.data[0] - prob.data[1])

margin_conf = 1 - difference
```

Entropy: difference between all predictions, as defined by information theory



$$\frac{-\sum_y P_\theta(y \mid x)\,\log_2 P_\theta(y \mid x))}{\log_2(n)}$$

```
prbslogs = prob * torch.log2(prob)
numerator = 0 - np.sum(prbslogs)
denominator = math.log2(prob.numel())

entropy = numerator / denominator
```

Active learning process

Procedure: Active Learning Process
Input: initial small training set L, and pool of unlabeled data set U
Use L to train the initial classifier C
Repeat
1. Use the current classifier C to label all unlabeled examples in U
2. Use uncertainty sampling technique to select m^2 most informative unlabeled examples, and ask oracle H for labeling
3. Augment L with these m new examples, and remove them from U
4. Use L to retrain the current classifier C

Until the predefined stopping criterion SC is met.

Refrences

- https://www.datacamp.com/community/tutorials/active-learning
- https://www.aclweb.org/anthology/C08-1143
- http://robertmunro.com/Uncertainty_Sampling_Cheatsheet_PyTorch.pdf