Active Learning

Name: Louise Baker Date: 03/12/23

Step 1: Initial Setup

1. **Import Libraries:**
   - Import the necessary libraries for your task. Common ones include sci-kit-learn for machine learning tasks.
2. **Load Data:**
   - Load your initial dataset. This should include both features (X) and labels (y).
3. **Split Data:**
   - Split your dataset into an initial labeled set and an unlabeled set. A common split is 80% labeled, 20% unlabeled.
4. **Train Initial Model:**
   - Train a machine learning model using the initial labeled dataset.

Step 2: Active Learning Loop

1. **Loop:**
   - Start a loop for active learning iterations.
2. **Query for Labels:**
   - Use your trained model to predict labels for the unlabeled data. Select the instances where the model is least certain (high uncertainty).
3. **Label Instances:**
   - Manually or automatically label the instances queried in the previous step.
4. **Update Labeled Set:**
   - Add the newly labeled instances to your labeled dataset.
5. **Retrain Model:**
   - Retrain your model using the updated labeled dataset.
6. **Repeat:**
   - Repeat steps 6-9 for a predefined number of iterations or until a certain performance threshold is reached.

Step 3: Evaluation (Optional)

1. **Validation:**
   - Optionally, evaluate your model on a validation set to monitor its performance during active learning.

2. **Test:**
    - After the active learning loop, evaluate your final model on a separate test set to assess its generalization.

Notes:

- **Model Choice:**
    - Choose a model suitable for your task, considering computational efficiency and ease of updating with new data.
- **Query Strategy:**
    - Define a strategy for querying instances. Common strategies include uncertainty sampling, query-by-committee, and diversity sampling.
- **Stopping Criteria:**
    - Decide on a stopping criterion for the active learning loop, such as reaching a certain accuracy or after a fixed number of iterations.
- **Data Annotation:**
    - Decide how new instances will be labeled. This can involve manual annotation, crowd-sourcing, or using pre-existing labeled datasets.