CSED 226
Introduction to Data Analysis
Final Exam

Problem

- You have learned ***supervised learning*** and ***unsupervised learning*** in this class.
- What if we have **a limited number of labeled data** for supervised learning tasks and addi tional **unlabeled data**?
  - You are given a limited number of labeled data and additional unlabeled data in this exam.
  - The goal is to improve your model's accuracy as much as possible using a given dataset consi sting of labeled and unlabeled data.

Strategies

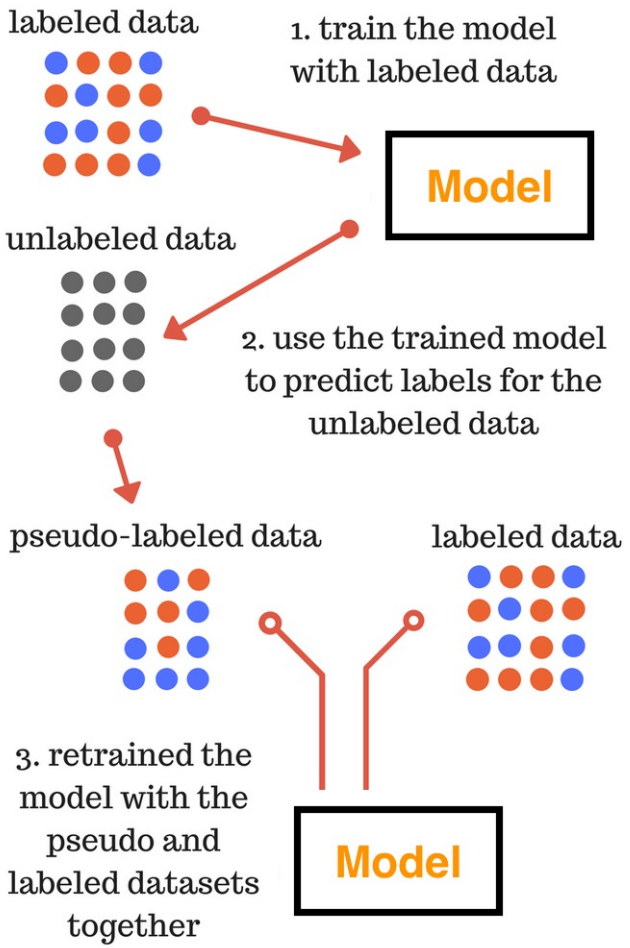
- You can **improve** your supervised learning task model's accuracy using additional unlabel ed data.
- **We suggest two simple strategies for this problem.**
  - Self-Training
  - Co-Training
- You can follow these strategies, or you can apply your own method.

Self-Training

- Train your classifier **with labeled data.**

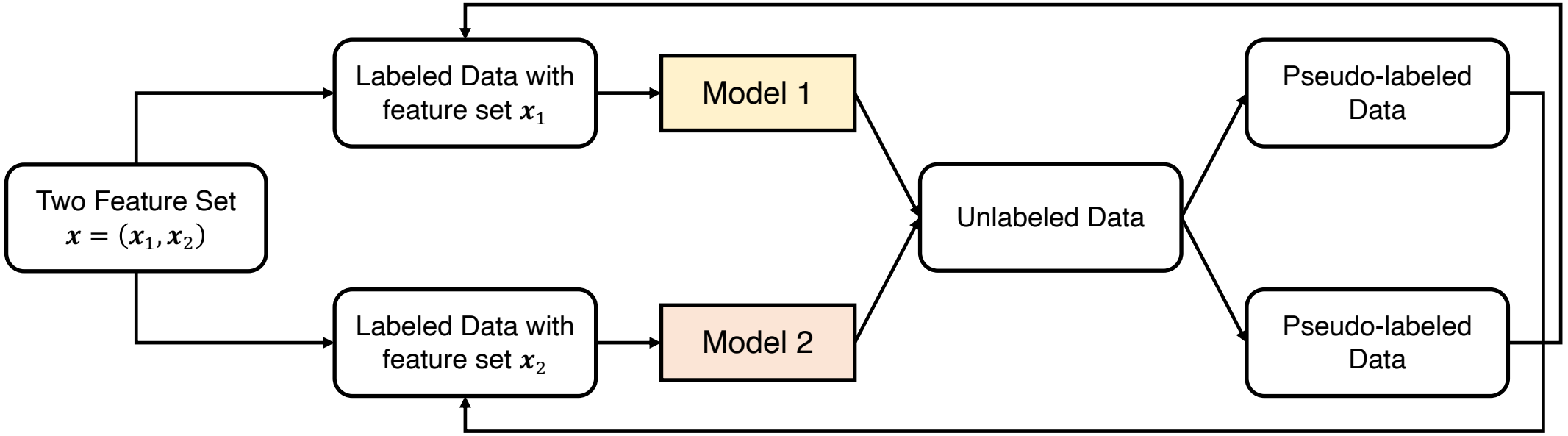
- Use the trained model to predict labels for the **unlabeled data.**
  - A confidence score can be used for predicting labels.
    - ✓ Score(x) > threshold
      - ✓ The model's prediction for x can be regarded as a confident label.
    - ✓ Score(x) < threshold
      - ✓ The model's prediction for x cannot be regarded as a confident label.

- Retrain the model with the **pseudo and labeled datasets tog ether.**



Co-Training

- Split features into two exclusive feature sets.
- Train two different classifiers using two different feature sets.
- Get pseudo-labels from unlabeled data.
- Expand labeled data using pseudo-labels from the different view models.



Libraries

- For the library, you can use the library used in the previous HW (i.e., seaborn, matplotlib, pandas, NumPy), sklearn, xgboost, kmodes, and mvlearn.
  - “sklearn.semi_supervised” and “mvlearn.semi_supervised” are allowed.