Machine Learning

COGS 108 Spring 2025
Jason Chen
Week 9

xic007@ucsd.edu

OH: Thu, 3-5 PM

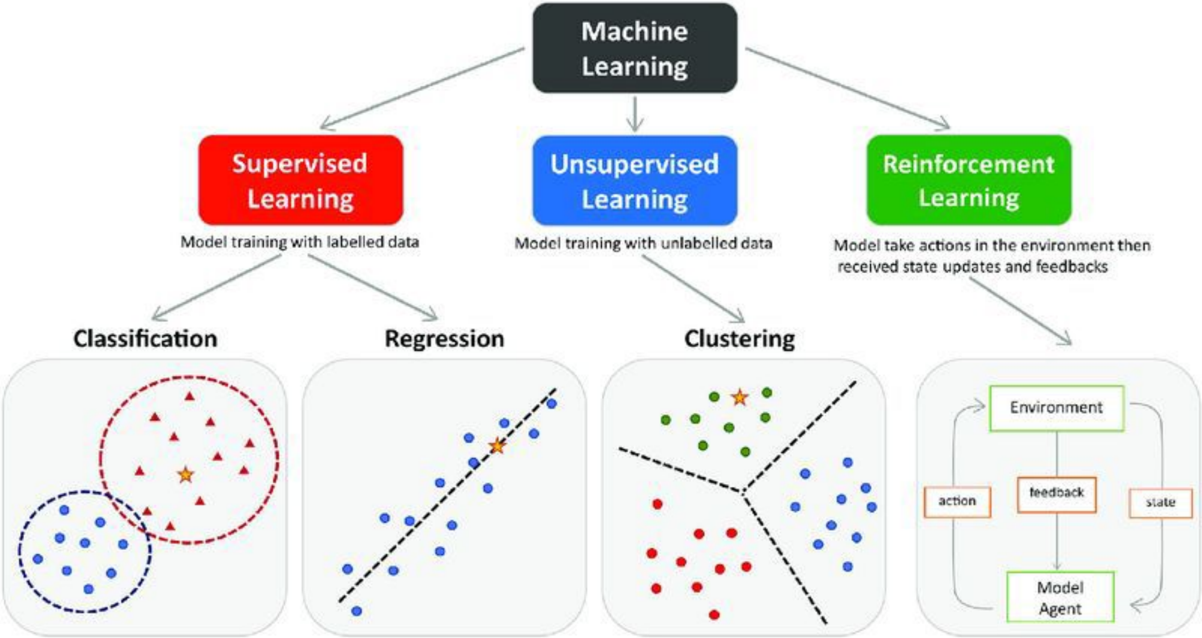
Discussion slides and materials adapted from Ruby Ying & previous quarter

Due dates

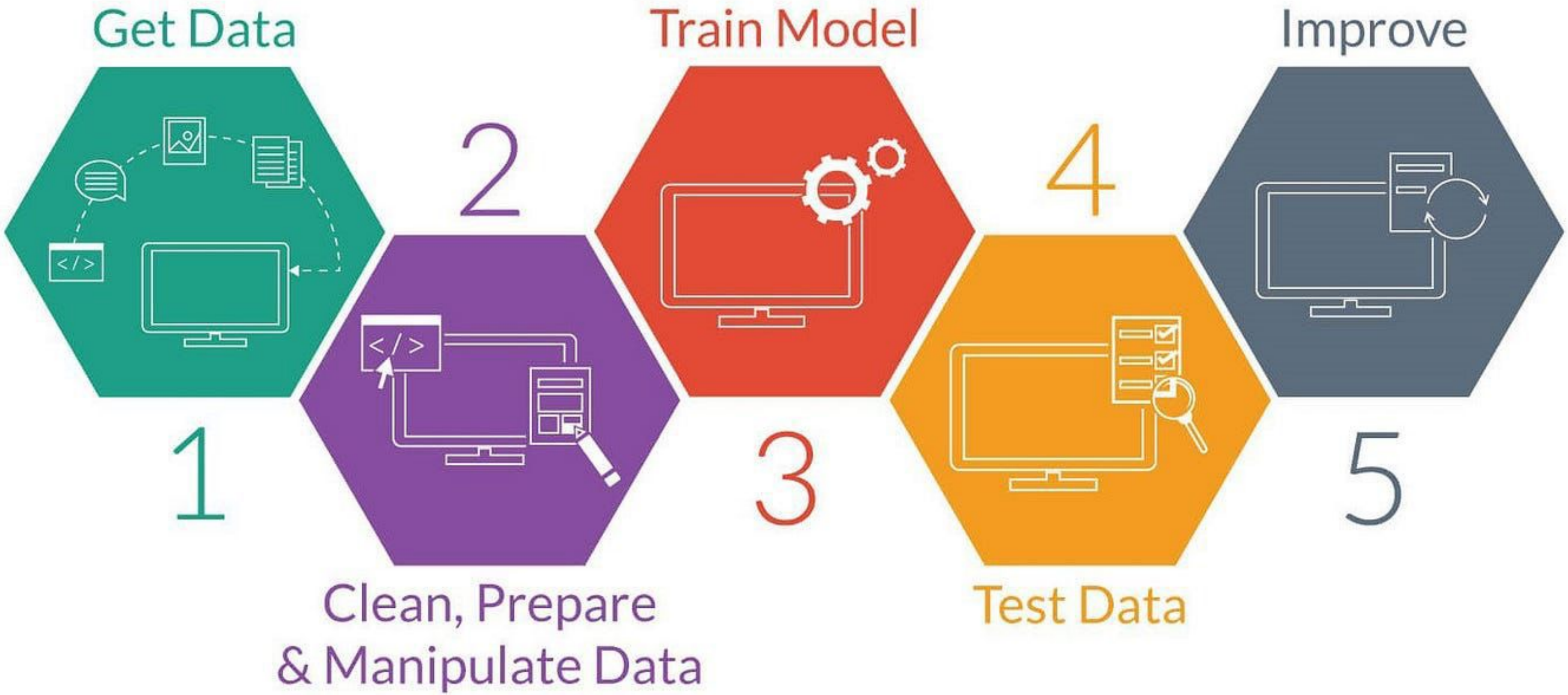
- Checkpoint #2: EDA is due Wednesday, May 28
- Discussion lab 8 is due Friday May 30

What is Machine Learning

- Machine learning is a subset of artificial intelligence that focuses on the development of algorithms and statistical models that enable computers to learn and make predictions or decisions without being explicitly programmed.



Steps Involved in ML



1 Get Data

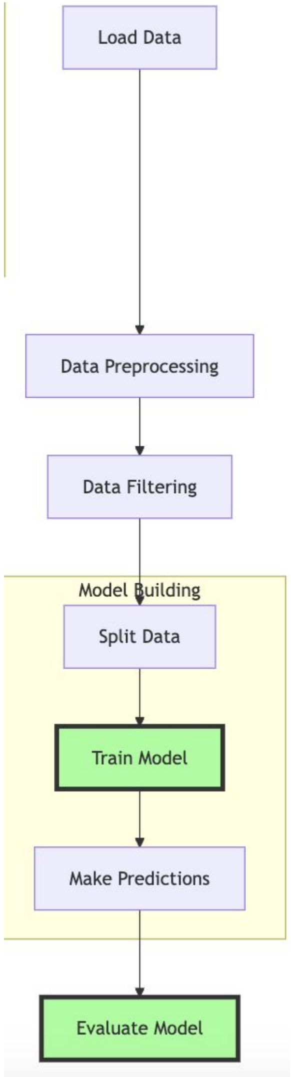
2 Clean, Prepare & Manipulate Data

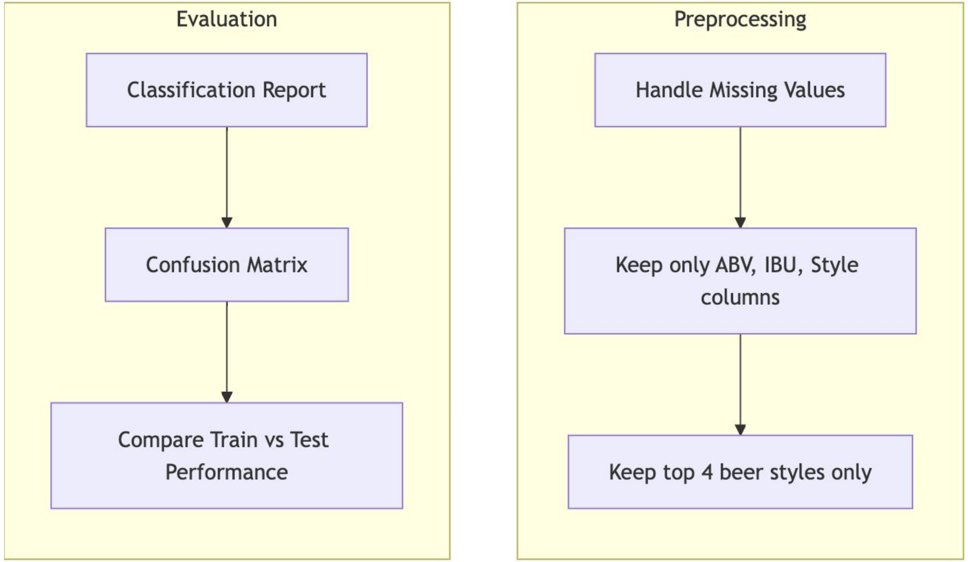
3 Train Model

4 Test Data

5 Improve

D8: Machine Learning




Evaluation
Classification Report
Confusion Matrix
Compare Train vs Test Performance
Preprocessing
Handle Missing Values
Keep only ABV, IBU, Style columns
Keep top 4 beer styles only

1. Data preprocessing: Handle missing values and extract ABV, IBU features
2. Filter data to keep only top 4 most common beer styles
3. Split data into training (80%) and test (20%) sets
4. Train SVM model and generate predictions
5. Evaluate model using classification reports and confusion matrices for both training and test sets

Part I: Data, Wrangling, & EDA

1. Analyze missing values(**.isnull().sum(axis=0)**)

```
   Name   ABV   IBU        Name   ABV    IBU        Name   1
0  Beer1  5.0  45.0     0  False  False  False      ABV    1
1  Beer2  NaN  60.0     1  False  True   False      IBU    2
2  Beer3  7.5   NaN     2  False  False  True       dtype: int64
3   None  4.8   NaN     3  True   False  True
```

2. Remove rows with missing values in style, abv, ibu(**dropna(subset=[])**)

3. Merge beer and brewery datasets(left join)
   - How does left join works?
4. Filter dataset to keep only top 4 styles (**.value_counts()[:].index.tolist()**)

Part II : Prediction Model

1. **Extract features (X: ABV, IBU) and labels (Y: Style)**

```
data_x = beer_df[['abv','ibu']]
data_y= np.array(beer_df['style'])
```

2. **Split data into train/test sets**

```
beer_train_X, beer_test_X, beer_train_ Y, beer_test_ Y = train_test_split(
   beer_X, beer_ Y, test_size=0.2, random_state=42, stratify=beer_ Y
)
```

3. Train SVM model and generate predictions

```
beer_clf =  Pipeline([
   ('scaler', StandardScaler()),
   ('svm', SVC(kernel='linear'))
])
beer_clf.fit()
beer_predicted_train_ Y = beer_clf.predict()
```

Part III : Model Assessment

1. Generate classification reports (precision, recall, f1-score)

2. Create confusion matrices

3. Compare training vs testing performance
   Train accuracy vs test accuracy

4. Evaluate potential overfitting

Section Materials

https://github.com/JasonC1217/COGS_108_B03-B04_Sp25/tree/master

or:

https://tinyurl.com/4d8wx3ne

SCAN ME

THANKS!

Questions on EdStem or office hours

Office hours: Thu, 3-5 PM