Week 1

Thursday, December 30, 2021 12:02 PM

What is Machine Learning?

* Intro to ML

- What is ML?
  - ML is the subfield of CS that gives "computers the ability to learn without being explicitly programmed."

- Major ML techniques
  - Regression / Estimation
    - Predicting continuous values
  - Classification
    - Predicting the item class / category of a case
  - Clustering
    - Finding the structure of data; summarization
  - Associations
    - Associating frequent co-occurring items / events
  - Anomaly detection
    - Discovering abnormal and unusual cases
  - Sequence mining
    - Predicting next events; click-stream (Markov Model, HMM)
  - Dimension Reduction
    - Reducing the size of data (PCA)
  - Recommendation systems
    - Recommending items

- Difference b/w AI, ML, and DL
  - AI components:
    - CV
    - Language Processing
    - Creativity, etc.
  - ML:
    - Classification
    - Clustering
    - Neural network, etc.
  - Revolution in ML:
    - Deep learning

* Python for ML

- Python libraries for ML
  - NumPy - SciPy - matplotlib - Pandas
  - scikit-learn

- More about scikit-learn

- Free software machine learning library
- Classification, Regression and Clustering algorithms
- Works with NumPy and SciPy
- Great documentation
- Easy to implement

Data Preprocessing → Train/Test split → Algorithm setup → Model fitting → Prediction → Evaluation → Model export

Question

Why Scikit is a proper library for Machine Learning (select all the options that are correct)?

- [x] Scikit-learn is a free machine learning library that works with Numpy and Scipy.

✓ Correct

- [x] Scikit-learn has most of machine learning algorithms.

✓ Correct

- [ ] Scikit-learn support all data science languages such as Python, R and Java.

- scikit-learn functions

```
from sklearn import preprocessing
X = preprocessing.StandardScaler().fit(X).transform(X)
```

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33)
```

```
from sklearn import svm
clf = svm.SVC(gamma=0.001, C=100.)
```

```
clf.fit(X_train, y_train)
```

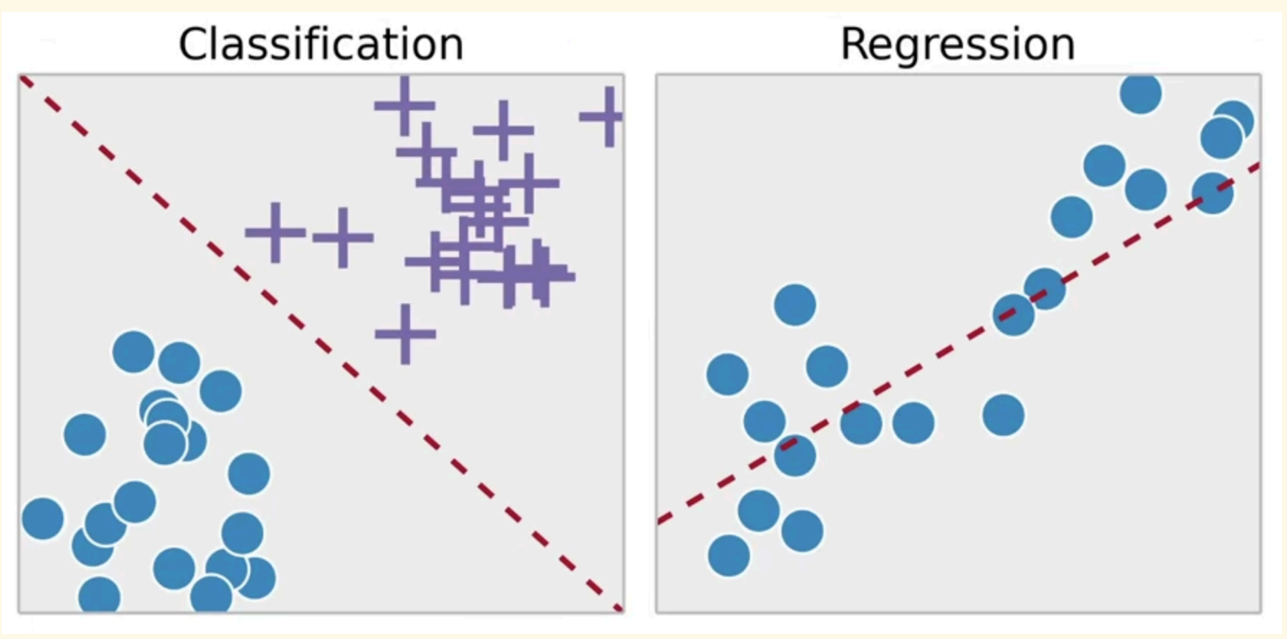
```
clf.predict(X_test)
```

```
from sklearn.metrics import confusion_matrix
print(confusion_matrix(y_test, yhat, labels=[1,0]))
```

```
import pickle
s = pickle.dumps(clf)
```

* Supervised vs Unsupervised

- Types of supervised learning



- What is classification?
  - Classification is the process of predicting discrete class labels or categories.

- What is regression?
  - Regression is the process of predicting continuous values.

- What is unsupervised learning?

Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio
1	41	2	6	19	0.124	1.073	NBA001	6.3
2	47	1	26	100	4.582	8.218	NBA021	12.8
3	33	2	10	57	6.111	5.802	NBA013	20.9
4	29	2	4	19	0.681	0.516	NBA009	6.3
5	47	1	31	253	9.308	8.908	NBA008	7.2
6	40	1	23	81	0.998	7.831	NBA016	10.9
7	38	2	4	56	0.442	0.454	NBA013	1.6
8	42	3	0	64	0.279	3.945	NBA009	6.6
9	26	1	5	18	0.575	2.215	NBA006	15.5
10	47	3	23	115	0.653	3.947	NBA011	4
11	44	3	8	88	0.285	5.083	NBA010	6.1
12	34	2	9	40	0.374	0.266	NBA003	1.6

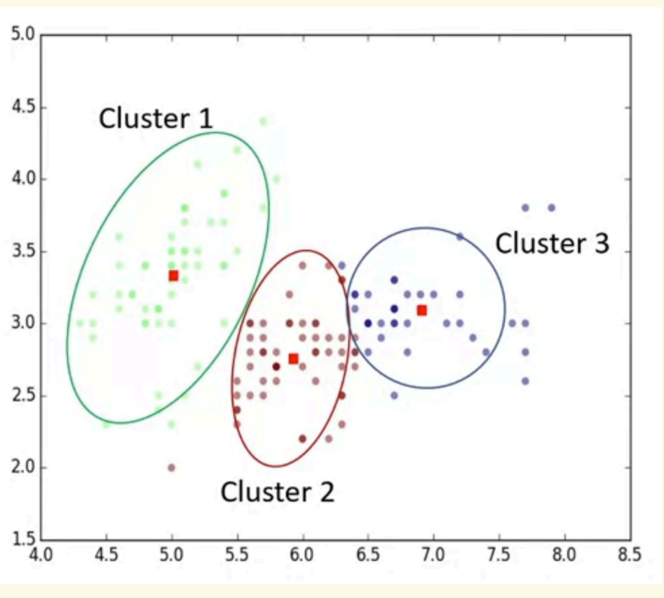
Unsupervised learning techniques:
- Dimension reduction
- Density estimation
- Market basket analysis
- Clustering

ALL OF THIS DATA IS UNLABELED

The model works on its own to discover information.

- What is clustering?

Clustering is grouping of data points or objects that are somehow similar by:
- Discovering structure
- Summarization
- Anomaly detection



- Supervised vs unsupervised learning

Supervised Learning
- **Classification:** Classifies labeled data
- **Regression:** Predicts trends using previous labeled data
- Has more evaluation methods than unsupervised learning
- Controlled environment

Unsupervised Learning
- **Clustering:** Finds patterns and groupings from unlabeled data
- Has fewer evaluation methods than supervised learning
- Less controlled environment