Machine Learning for Absolute Beginners

Machine Learning

Using data-driven mathematical algorithms or statistical models in computer systems instead of explicit rules

Today's Topics

- Representing data as vectors, feature extraction
- Supervised learning (classification)
- Unsupervised learning (clustering, anomaly detection)
- Summary

1. Representing Data

Data as Vectors

Dataset with N samples (row vectors) and M features (column vectors):

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots \\ \vdots & \ddots & \\ x_{n1} & & x_{nm} \end{pmatrix}$$

In supervised learning, the target (e.g. class in classification) is a column vector of length N:

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

Example: Wine Dataset

178 samples (wines), 13 features (chemical measurements). Type of the wine (0, 1, 2) as target (class).

Feature Extraction

Deriving informative and non-redundant values out of raw data to facilitate the learning

- Feature engineering:
    - Manual work to figure out good features for data
    - Requires domain knowledge, testing, brainstorming and a lot of trial and error
- Feature learning:
    - Learning features automatically
    - Common with deep learning (neural networks)
- Dimensionality reduction
    - Reducing non-redundant information from data to reduce computational complexity
    - Selecting a subset of features or applying further feature extraction methods

Example: Feature Engineering For Sound Data

2. Supervised Learning

A machine learning task where the desired output for new inputs is learnt from known input-output pairs.

- Regression: estimating the mapping from (continuous) input to output
- Classification: identifying which category a sample belongs to, based on previous samples (training set) whose category is known

Example: Classifying Sound Data with Support Vector Machine

Support Vector Machine

- Binary classifier that is based on finding the optimal hyperplane between the two classes that separates the classes with maximum margin
- New samples that fall on the same side of the hyperplane as the training samples in class y_i are classified to that class

3. Unsupervised Learning

Clustering

Finding groups of data (clusters) so that the samples within a cluster are more similar to each other than to the samples in other clusters

Example: Clustering Sound Data with k-means

k-means:

- Partition data into k clusters
- Iteratively move cluster *centroids* to the mean of the samples that are closest to it
- Minimizes variances of the clusters

Anomaly detection

Identifying samples that differ from the majority of samples

Example: Anomaly Detection For Sound Data

Local Outlier Factor

- Identify outliers (anomalies) by comparing their distance to k nearest neighbours to the corresponding distances of those neighbours

4. Summary

What Wasn't Discussed

- Tools:
    - Interactive Python environment (this thing I've been using) Jupyter (https://jupyter.org/)
    - Vectors and matrices: numpy (https://numpy.org/)
    - Higher-level dataframe API: pandas (https://pandas.pydata.org/)
    - Algorithms: scipy (https://www.scipy.org/), scikit-learn (https://scikit-learn.org/stable/), keras (https://keras.io/), ...
    - Visualization: matplotlib (https://matplotlib.org/), seaborn (https://seaborn.pydata.org/), ...
- Practical issues:
    - Data issues (lack of data, missing features, imbalanced classes)
    - Overfitting / underfitting
    - Model selection and evaluation
- Neural networks

Summary

- Data as vectors
    - Feature engineering, feature learning
- Supervised learning:
    - Classification: identifying which category a sample belongs to based on training set
    - SVM: classification by maximizing the margin between the classes
- Unsupervised learning:
    - Clustering: finding groups of data so that the samples within a cluster are more similar to each other than to the samples in other clusters
    - Anomaly detection: finding data that is different from the majority of data
- Machine learning: using data-driven mathematical algorithms or statistical models in computer systems instead of explicit rules