**Global AI Intern - ML EBOOK Write-up**

Cheng-Yun Tsai

**Introduction to Machine Learning**

* Machine learning algorithms use computational methods to “learn” information directly from data without relying on a **predetermined** equation as a model (eBOOK Section1).
* Machine Learning algorithms can be separated into two categories, **supervised learning** and **unsupervised learning**.
* Supervised learning uses the known features (the input data) and the labels (the output data) to solve **classification** and **regression** problems.
* Unsupervised learning extracts hidden pattern from features and solve problems by **clustering** techniques.
* Since there are dozens of algorithms in each category, selecting a suitable algorithm under different situations is a critical problem.
* Fig. 1 shows the classification of ML algorithms.

Diagram

Description automatically generated

Fig. 1 ML classification

**Model Selection and Workflow**

* More flexible models may cause **overfitting**, simple models may **assume too much**. We need to try out different models and find the suitable point between the tradeoffs.
* The **workflow** of constructing ML models can be summarized below.

1. **Load the data**: Get the data from database/ API/ …
2. **Preprocess the data**: Data Cleaning/ Feature Engineering/ EDA.
3. **Derive Features**: Feature Engineering/ Data Augmentation by domain knowledge
4. **Build and Train the model**: building models/ visualizing results
5. **Improve the model**: Simplify the models by PCA/ Correlation matrix/ … or make the model more complicated by adding features/ model combination.

**Unsupervised Learning**

* Most unsupervised learning algorithms are clustering analysis, which classify data into groups have similar features (e.g., Gaussian Mixture).
* Clustering can be classified into **hard clustering** and **soft clustering** (each point belongs to more than one cluster).
* Common hard clustering methods include k-Means, k-Medoids, Hierarchical Clustering and Self-Organizing Map.
* Common soft clustering methods include Fuzzy c-Means and Gaussian Mixture Model.
* And to improve the model, dimensionality reduction techniques such as PCA, Factor analysis and Nonnegative matrix factorization are common methods.
* PCA can be used in datasets with many variables, groups of variables often move together.
* Factor analysis measures the underlying correlations of variables (features).
* Nonnegative matrix factorization generates nonnegative results.
* Unsupervised learning might be the final algorithm of your model but can also be a step of feature engineering. Fig. 2 shows the workflow of unsupervised learning.

Diagram

Description automatically generated

Fig. 2 Unsupervised and Supervised Learning Map

**Supervised Learning**

* Selection of Machine Learning algorithm is a trade-off between:

1. Speed of training
2. Memory usage
3. Testing accuracy
4. Interpretability

* Common **Classification** algorithms include Logistic regression, KNN, SVM, NN, Naïve Bayes, Discriminant analysis, Decision tree, Bagged and Boosted Decision Trees
* Common **Regression** algorithms include Linear Regression, Nonlinear regression, Gaussian Process Regression (GPR), SVM Regression, Generalized Linear Model, Regression Tree.
* We can improve models by feature engineering and fine tunning.