Concepts in Machine Learning

Prof.Dr. Bahadır AKTUĞ
Machine Learning with Python

**Compiled from sources given in the references.*

Machine Learning

- Machine learning is a subfield of computer science that is concerned with building algorithms which, to be useful, rely on a collection of examples of some phenomenon.
- These examples can come from nature, be handcrafted by humans or generated by another algorithm.
- Machine learning can also be defined as the process of solving a practical problem by
  - 1) gathering a dataset, and
  - 2) algorithmically building a statistical model based on that dataset.

Machine Learning

- Learning can be
  - Supervised
  - Semi-supervised
  - Unsupervised
  - Reinforcement

Supervised vs. Unsupervised Learning

- The goal of a supervised learning algorithm is to use the dataset to produce a model that takes a feature vector x as input and outputs information that allows deducing the label for this feature vector.

Unsupervised Learning

- In unsupervised learning, the dataset is a collection of unlabeled examples {xi}Ni=1.
- Again, x is a feature vector, and the goal of an unsupervised learning algorithm is to create a model that takes a feature vector x as input and either transforms it into another vector or into a value that can be used to solve a practical problem.
- For example, in clustering, the model returns the id of the cluster for each feature vector in the dataset.
- In dimensionality reduction, the output of the model is a feature vector that has fewer features than the input x; in outlier detection, the output is a real number that indicates how x is different from a "typical" example in the dataset.

Semisupervised Learning

- In semi-supervised learning, the dataset contains both labeled and unlabeled examples.
- Usually, the quantity of unlabeled examples is much higher than the number of labeled examples. The goal of a semi-supervised learning algorithm is the same as the goal of the supervised learning algorithm. The hope here is that using many unlabeled examples can help the learning algorithm to find (we might say "produce" or "compute") a better model.
- It could look counter-intuitive that learning could benefit from adding more unlabeled examples. It seems like we add more uncertainty to the problem. However, when you add unlabeled examples, you add more information about your problem: a larger sample reflects better the probability distribution the data we labeled came from.
- Theoretically, a learning algorithm should be able to leverage this additional information.

Reinforcement Learning

- Reinforcement learning is a subfield of machine learning where the machine "lives" in an environment and is capable of perceiving the state of that environment as a vector of features. The machine can execute actions in every state.
- Different actions bring different rewards and could also move the machine to another state of the environment.
- The goal of a reinforcement learning algorithm is to learn a policy.

Definitions

- Model: the collection of parameters you are trying to fit
- Data: what you are using to fit the model
- Target: the value you are trying to predict with your model
- Features: attributes of your data that will be used in prediction
- Methods: algorithms that will use your data to fit a model

Supervised Learning

- There are two main tasks of supervised learning:
  - Classification:The systems learns from already labeled data (training set) and tries to predict the class of the unknown samples (test set).
  - Regression: If the desired output consists of one or more continuous variables, then the task is called regression.

Unsupervised Learning

- There are three main tasks of unsupervised learning:
  - Clustering: find groups of similar samples (clustering)
  - Density Estimation: determine the data distribution (density estimation)
  - Dimension Reduction: project data from high-dimensional space to a low-dimensional space

Classification

- Classification is a problem of automatically assigning a label to an unlabeled example.
- Spam detection is a famous example of classification.
- In machine learning, the classification problem is solved by a classification learning algorithm that takes a collection of labeled examples as inputs and produces a model that can take an unlabeled example as input and either directly output a label or output a number that can be used by the analyst to deduce the label. An example of such a number is a probability.
- In a classification problem, a label is a member of a finite set of classes. If the size of the set of classes is two ("sick"/"healthy", "spam"/"not_spam"), we talk about binary classification (also called binomial in some sources).
- Multiclass classification (also called multinomial) is a classification problem with three or more classes

Regression

- Regression is a problem of predicting a real-valued label (often called a target) given an unlabeled example.
- Estimating house price valuation based on house features, such as area, the number of bedrooms, location and so on is a famous example of regression.
- The regression problem is solved by a regression learning algorithm that takes a collection of labeled examples as inputs and produces a model that can take an unlabeled example as input and output a target.

Model-Based vs. Instance-Based Learning

- Most supervised learning algorithms are model-based.
- We have already seen one such algorithm: SVM. Model-based learning algorithms use the training data to create a model that has parameters learned from the training data. In SVM, the two parameters we saw were w∗ and b∗. After the model was built, the training data can be discarded.
- Instance-based learning algorithms use the whole dataset as the model. One instance-based algorithm frequently used in practice is k-Nearest Neighbors (kNN).
- In classification, to predict a label for an input example the kNN algorithm looks at the close neighborhood of the input example in the space of feature vectors and outputs the label that it saw the most often in this close neighborhood

Shallow vs. Deep Learning

- A shallow learning algorithm learns the parameters of the model directly from the features of the training examples. Most supervised learning algorithms are shallow.
- The notorious exceptions are neural network learning algorithms, specifically those that build neural networks with more than one layer between input and output. Such neural networks are called deep neural networks.
- In deep neural network learning (or, simply, deep learning), contrary to shallow learning, most model parameters are learned not directly from the features of the training examples, but from the outputs of the preceding layers.

Supervised Learning Algorithms

- Linear Models
- Linear and Quadratic Discriminant Analysis
- Kernel ridge regression
- Support Vector Machines
- Stochastic Gradient Descent
- Nearest Neighbors
- Gaussian Processes
- Cross decomposition
- Naive Bayes
- Decision Trees
- Ensemble methods
- Multiclass and multioutput algorithms

- Feature selection
- Semi-supervised learning
- Isotonic regression
- Probability calibration
- Neural network models (supervised)

Unsupervised Learning Algorithms

- Gaussian mixture models
- Manifold learning
- Clustering
- Biclustering
- Decomposing signals in components (matrix factorization problems)
- Covariance estimation
- Novelty and Outlier Detection
- Density Estimation
- Neural network models (unsupervised)

References

1 https://scikit-learn.org/

2 https://towardsdatascience.com/

3 McKinney, W. (2017). Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython 2nd Edition.

4 Albon, C. (2018). Machine Learning with Python Cookbook: Practical Solutions from Preprocessing to Deep Learning

5 Géron, A. (2017). Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems 1st Edition

6 Müller, A. C., Guido, S. (2016). Introduction to Machine Learning with Python: A Guide for Data Scientists

7 Burkov, A. (2019). The Hundred-Page Machine Learning Book.

8 Burkov, A. (2020). Machine Learning Engineering.

9 Goodrich, M.T., Tamassia, R., Goldwasser, M.H. (2013). Data Structures and Algorithms in Python, Wiley.

10 https://towardsdatascience.com

11 https://docs.python.org/3/tutorial/

12 http://www.python-course.eu

13 https://developers.google.com/edu/python/

14 http://learnpythonthehardway.org/book/