Hands-on Machine Learning with Scikit-Learn, Keras and Tensorflow

Solutions to Chapter 1 Exercises

1.

Machine learning is a field of computing science that incorporates statistical models into computational algorithms that perform predictions or mappings on data.

2.

- When your program requires a lot of hard-coded rules
- Complex problems with not traditional solution
- Unpredictable environments or problems
- Drawing patterns from a lot data

3.

labels are the correct predictions assigned to items in a training set that a machine learning algorithm should try to classify or predict data into.

4.

Classification and value prediction

5.

Clustering Anomaly detection Visualization and dimensionality reduction Association learning

6.

This problems sounds appropriate for reinforcement learning. This model will equip the robot with the ability to explore unknown environments

7.

A clustering algorithm could be used for this purpose. We can find groups of similar customers this way.

8.

Supervised learning problem. We can create a model using a large of set of example spam and non-spam emails and train (fit) the model according to the labels.

9.

Online learning is a method of training an ML system in real time as the system receives data incrementally or in new batches.

10.

Out-of-core learning is a method of online learning for handling large datasets that cannot be loaded all at once into the systems memory. Instead part of the data is loads, system is trained and the process is repeated until all the data is used.

11.

Clustering

12.

Model parameters are set by the system being trained. Hyperparameters are set by the human conducting the training prior to training.

13.

They look for a model that best generalizes an example training set. They use a cost function to keep track of how well a model is performing. They make predictions by taking input data and passing it to the model which is often a type of polynomial and get a numerical output data.

14.

- Insufficient Training Data
- Non-representative training data
- Overfitting
- Underfitting

15.

It's likely due to overfitting the training data. To overcome this you can:

- simplify the model
- increase the number of training data available
- reduce noise in the training data

16.

A test set is used to measure the generalization error of a system to see how well it performs when encountering new data.

17.

Validation set is a subset of the training set that is being used to compare the performance of multiple plausible models. You train multiple models on the remainder of the training set and run the models with the validation set and pick the best performing one.

18.

A train-dev set is a subset of a training set used to measure the performance of a system on training data. If the error is high it indicates an overfitting of training data. A poor performance could indicate poor training data that requires preprocessing.

19.

Performing the tuning over and over on the test set indicates we are trying to find the best fit of that particular set of data and thus could mean our model would not generalize well and handle new data.