Machine Learning Topics

# The Machine Learning Landscape

## Types of Machine Learning Systems

### Supervised vs Unsupervised

* **Supervised Learning**  
  The training set you feed to the algorithm includes the desired solutions, called *labels*.
* **Unsupervised Learning**  
  The training set is unlabeled. The system tries to learn without a teacher
* **Semi-supervised Learning**  
  The algorithms can deal with data that’s partially labeled. Most of these types of algorithms are combinations of unsupervised and supervised algorithms
* **Reinforcement Learning**The learning system (*agent*) can observe the environment, select and perform actions, and get *rewards* or *penalties* in return. It must then learn by itself what is the best strategy, called a *policy*, to get the most reward over time. A policy defines what action the agent should choose when it is in a given situation.

### Batch and Online Learning

* **Batch Learning**  
  The system is incapable of learning incrementally: it must be trained using all the available data, usually done offline, and then once the model is trained, it is then launched into production and runs without learning anymore (*offline learning*)
* **Online Learning**  
  You train the system incrementally by feeding it data instances sequentially, either individually or in small groups called *mini batches*. Each learning step is fast and cheap, so the system can learn about new data on the fly, as it arrives.
  + Great for systems that receive data as a continuous flow (e.g., stock prices) and need to adapt to change rapidly or autonomously.
  + Also good with limited computing resources: once an online learning system has learned about new data instances, it does not need them anymore and you can discard them.
  + Online learning algorithms can be used to train systems on huge datasets that cannot fit in one machine’s main memory (*out-of-core learning*).
  + One important parameter of online learning systems is how fast they should adapt to changing data (*learning rate*)
  + If bad data is fed to the system, the system’s performance will gradually decline.

### Instance-Based vs. Model-Based learning

How do machine learning systems *generalize*?

* **Instance-based learning**  
  The system learns the examples by heart, then generalizes to new cases by using a similarity measure to compare them to the learned examples (or a subset).
* **Model-based learning**Build a model of these examples and then use that model to make *predictions*.
  + The model takes parameters.
  + To determine how you can know which values make your model perform best, you need to specify a performance measure.
    - Utility/fitness function measures how good your model is.
    - Cost function that measures how bad it is.

## Main Challenges of Machine Learning

* Insufficient Quantity of Training Data
* Nonrepresentative Training Data
* Poor Quality Data
* Irrelevant Features
* Overfitting
* Underfitting

## Testing and Validation

Split data into two sets: *training* and *test* sets. The error rate on new cases is called the *generalization error* (or *out-of-sample error*), and by evaluating your model on the test set, you get an estimate of this error.

# End-to-End Machine Learning Project