Chapter 1

Textbook Exercises

Q1: How would you define Machine Learning?

Q1: How would you define Machine Learning?

Machine Learning is about building systems that can learn from data. Learning means getting better at some task, given some performance measure.

Q2: Can you name four types of problems where it shines?

Q2: Can you name four types of problems where it shines?

Machine Learning is great for complex problems for which we have no algorithmic solution, to replace long lists of hand-tuned rules, to build systems that adapt to fluctuating environments, and finally to help humans learn (e.g., data mining).

Q3: What is a labeled training set?

Q3: What is a labeled training set?

A labeled training set is a training set that contains the desired solution (a.k.a. a label) for each instance.

Q4: What are the two most common supervised tasks?

Q4: What are the two most common supervised tasks?

The two most common supervised tasks are regression and classification.

Q5: Can you name four common unsupervised tasks?

Q5: Can you name four common unsupervised tasks?

Common unsupervised tasks include clustering, visualization, dimensionality reduction, and association rule learning.

Q6:What type of Machine Learning algorithm would you use to allow a robot to walk in various unknown terrains?

Q6:What type of Machine Learning algorithm would you use to allow a robot to walk in various unknown terrains?

Reinforcement Learning is likely to perform best if we want a robot to learn to walk in various unknown terrains, since this is typically the type of problem that Reinforcement Learning tackles. It might be possible to express the problem as a supervised or semisupervised learning problem, but it would be less natural.

Q7: What type of algorithm would you use to segment your customers into multiple groups?

Q7: What type of algorithm would you use to segment your customers into multiple groups?

If you don't know how to define the groups, then you can use a clustering algorithm (unsupervised learning) to segment your customers into clusters of similar customers. However, if you know what groups you would like to have, then you can feed many examples of each group to a classification algorithm (supervised learning), and it will classify all your customers into these groups.

Q8: Would you frame the problem of spam detection as a supervised learning problem or an unsupervised learning problem?

Q8: Would you frame the problem of spam detection as a supervised learning problem or an unsupervised learning problem?

Spam detection is a typical supervised learning problem: the algorithm is fed many emails along with their labels (spam or not spam).

Q9: What is an online learning system?

Q9: What is an online learning system?

An online learning system can learn incrementally, as opposed to a batch learning system. This makes it capable of adapting rapidly to both changing data and autonomous systems, and of training on very large quantities of data.

Q10: What is out-of-core learning?

Q10: What is out-of-core learning?

Out-of-core algorithms can handle vast quantities of data that cannot fit in a computer's main memory. An out-of-core learning algorithm chops the data into mini-batches and uses online learning techniques to learn from these minibatches.

Q12: What is the difference between a model parameter and a learning algorithm's hyperparameter?

Q12: What is the difference between a model parameter and a learning algorithm's hyperparameter?

A model has one or more model parameters that determine what it will predict given a new instance (e.g., the slope of a linear model). A learning algorithm tries to find optimal values for these parameters such that the model generalizes well to new instances. A hyperparameter is a parameter of the learning algorithm itself, not of the model (e.g., the amount of regularization to apply).

Q14: Can you name four of the main challenges in Machine Learning?

Q14: Can you name four of the main challenges in Machine Learning?

Some of the main challenges in Machine Learning are the lack of data, poor data quality, nonrepresentative data, uninformative features, excessively simple models that underfit the training data, and excessively complex models that overfit the data.

Q15: If your model performs great on the training data but generalizes poorly to new instances, what is happening? Can you name three possible solutions?

Q15: If your model performs great on the training data but generalizes poorly to new instances, what is happening? Can you name three possible solutions?

If a model performs great on the training data but generalizes poorly to new instances, the model is likely overfitting the training data (or we got extremely lucky on the training data). Possible solutions to overfitting are getting more data, simplifying the model (selecting a simpler algorithm, reducing the number of parameters or features used, or regularizing the model), or reducing the noise in the training data.

Q16: What is a test set, and why would you want to use it?

Q16: What is a test set, and why would you want to use it?

A test set is used to estimate the generalization error that a model will make on new instances, before the model is launched in production.

Q17: What is the purpose of a validation set?

Q17: What is the purpose of a validation set?

A validation set is used to compare models. It makes it possible to select the best model and tune the hyperparameters.

Q19: What can go wrong if you tune hyperparameters using the test set?

Q19: What can go wrong if you tune hyperparameters using the test set?

If you tune hyperparameters using the test set, you risk overfitting the test set, and the generalization error you measure will be optimistic (you may launch a model that performs worse than you expect).