Note: All the questions and answer can be found in the book, the black color font is my own answer after reading the chapters, and the red one is the answer that author provided.

Ch1 Question:

1. **How would you define Machine Learning?**

   Machine Learning is the science and art of programming computers so they can learn from data.

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

2. **Can you name four types of problems where it shines?**

   Classification, Prediction,

   Ans: 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).

3. **What is a labeled training set?**

   This means for each column of the dataset; we have a y label to determine the type of the data.

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

4. **What are the two most common supervised tasks?**

   Spam email detestation, credit card fraud transaction detection.

   Ans: The two most common supervised tasks are regression and classification.

5. **Can you name four common unsupervised tasks?**

   Customer cluster analysis.

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

6. **What type of Machine Learning algorithm would you use to allow a robot to walk in various unknown terrains?**

   Supervised algorithm: give the label of each object like ‘human’, ‘building’ or ‘trees’ and then train them to avoid walking them.

   Ans: 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 semi-supervised learning problem, but it would be less natural.

7. **What type of algorithm would you use to segment your customers into multiple groups?**

   Unsupervised Learning algorithm

   Ans: 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.

8. **Would you frame the problem of spam detection as a supervised learning problem or an unsupervised learning problem?**

   Supervised Learning Problem.

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

9. **What is an online learning system?**

   Online learning system means a model can be trained by import small size batch of the data multiple times through the process, then the model will become more and more intelligence after learning.

   Ans: 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.

10. **What is out-of-core learning?**

    N/A

    Ans: 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 mini-batches.

11. **What type of learning algorithm relies on a similarity measure to make predictions?**

    Instance based model.

    Ans: An instance-based learning system learns the training data by heart; then, when given a new instance, it uses a similarity measure to find the most similar learned instances and uses them to make predictions.

12. **What is the difference between a model parameter and a learning algorithm's hyperparameter?**

Model parameter can't not change since it is defined by the model itself during the training process, however the hyperparameter usually is defined manually in order to improve the model performance.

Ans: 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).

13. **What do model-based learning algorithms search for? What is the most common strategy they use to succeed? How do they make predictions?**

The model search for the best parameter combination in order to minimize the prediction error.

They try all the parameters combination then compare the error in order to get the optimal one.

It makes predictions on the new testing dataset by using the model trained on the train data set.

Aws: Model-based learning algorithms search for an optimal value for the model parameters such that the model will generalize well to new instances.

We usually train such systems by minimizing a cost function that measures how bad the system is at making predictions on the training data, plus a penalty for model complexity if the model is regularized.

To make predictions, we feed the new instance's features into the model's prediction function, using the parameter values found by the learning algorithm.

14. **Can you name four of the main challenges in Machine Learning?**

    1) The dataset is too large which might need distributed computation power
    2) Train dataset sometimes in low quality.
    3) Train dataset become more complex nowadays such as the image and speech recognition and transformation.
    4) Too much people using the name of "Machine Learning/Deep Learning" for their blueprint or subject but actually know nothing even the basic idea.

Ans:

1） lack of data, poor data quality
2） nonrepresentative data
3） uninformative features

4）excessively simple models that underfit the training data, and excessively complex models that overfit the data.

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

It might be overfitting.

1) Import more training dataset for the model.
2) Decrease the training metrics be feature engineering or feature selection.
3) Remove some constraints so that the model could be general.

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.

16. **What is a test set, and why would you want to use it?**

The test set is proportion of the entire dataset, it didn't use for training by the model but use for testing the performance of the model as the "unknown dataset"

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.

17. **What is the purpose of a validation set?**

The validation set is use for find the better parameter combination of the model, ultimately improve the model performance on the training data set.

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

18. **What is the train-dev set, when do you need it, and how do you use it?**

N/A

When you have multiple data source such as the flower pictures from the internet which have multiple branches.

Pick the examples from all the type of flowers to avoid there is no repetitive dataset in training set.

The train-dev set is used when there is a risk of mismatch between the training data and the data used in the validation and test datasets (which should always be as close as possible to the data used once the model is in production).

The train-dev set is a part of the training set that's held out (the model is not trained on it). The model is trained on the rest of the training set, and evaluated on both the train-dev set and the validation set.

If the model performs well on the training set but not on the train-dev set, then the model is likely overfitting the training set. If it performs well on both the training set and the train-dev set, but not on the validation set, then there is probably a significant data mismatch between the training data and the validation + test data, and you should try to improve the training data to make it look more like the validation + test data.

19. What can go wrong if you tune hyperparameters using the test set?

So there will no data available to see the performance of the model.

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).