Guide to the Reading

CSC316 Machine Learning
Professor Leon Tabak

07 February 2022
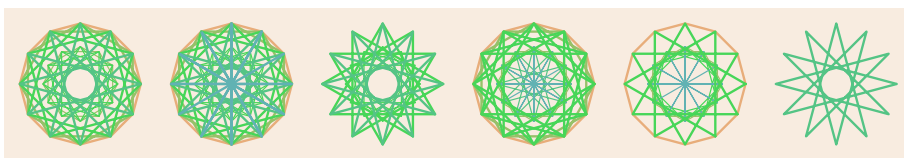
1. The first chapter of *Hands-On Machine Learning* differs from the chapters that follow by the fact that it contains...

   (a) more code?

   (b) less code?

2. We will find in the first chapter of *Hands-On Machine Learning*...

   (a) a timeline of the development of machine learning

   (b) a review of the Python programming language

   (c) listings of the most important functions in the Python libraries that we will use

   (d) quotations and anecdotes to inspire us and motivate our study

   (e) ideas and vocabulary with which we must make ourselves thoroughly familiar

3. OCR is an example of an early application of machine learning. What is OCR?

4. Knowingly or unknowingly, we are all using spam filters. This application of machine learning appeared when?

5. To build a spam filter, a data scientist needs a *training set*. The training set contains *training instances*. These instances are also called *samples*.

   What might we find in the *training set*?

6. Here are some ways to categorize machine learning methods:

   - supervised vs. unsupervised
   - batch vs. online
   - instance-based vs. model-based

   (a) What does *supervised* mean in the case of the spam filter? Hint: the input to the machine learning algorithm includes both messages and labels. What are the labels?

   (b) In the case of the spam filter, what do we gain by choosing online learning over batch learning?

   (c) What is an example of a *measure of similarity* that we might use in an instance-based machine learning system for the detection of spam?

7. Why might a data scientist choose a low *learning rate*?

8. What might make *batch learning* impractical?

9. Keep a browser open while you read *Hands-On Machine Learning.* When you encounter unfamiliar words in the book, search on the Internet for alternative or complementary definitions. This will help you gain a fuller understanding of the new ideas.

   (a) Who is Ke Jie?

   (b) What is *AlphaGo*?

   (c) What is *DeepMind*?

   (d) What kind of machine learning algorithm does *AlphaGo* use?

10. The editors of a magazine wanted to predict the winner of the 1936 U.S. presidential election. To this end, they surveyed a large number of people.

    (a) Who was the Democratic candidate for president in 1936?

    (b) Who was the Republican candidate for president in 1936?

    (c) Who won the election?

    (d) The magazine failed to predict the winner. Where did the editors go wrong?

11. A Lagrange interpolating polynomial describes a curve that passes through all N points that we use to construct the polynomial.

    Although this curve perfectly predicts all values in a training set, it makes a very poor machine learning model. Why?

12. Distinguish between *feature selection* and *feature extraction*.

13. The quality of our data might be poor. Describe two ways of cleaning data.

14. What does it mean to say that a training set is *noisy*?

15. Regularization means directly changing the values of...

    (a) dataset?
    (b) model parameters?
    (c) hyperparameters?

16. Adjusting the values of hyperparameters can help us...

    (a) solve problems of overfitting?
    (b) solve problems of underfitting?
    (c) solve problems of overfitting and underfitting?

17. Our success in building a machine learning system depends upon...

    (a) the characteristics of our dataset?
    (b) our choice of a machine learning model?
    (c) both the dataset and the model?

18. A *validation set* is a subset of the...

    (a) *training set*?
    (b) *test set*?

19. *Cross-validation* is a means of trying to solve what problem?

    You will likely need more words to answer this question than you need to answer other questions in this exercise.

20. Aurélien Geron, the author of *Hands-On Machine Learning*, gave developed a model to predict happiness of a nation's people in Chapter 1 of the book. He used a dataset that is available online here.

    (a) The model supposes that there might be correlation between happiness and which other variable?
    (b) The dataset comes from the OECD. What is the OECD?
    (c) The model is linear. Write a mathematical expression that shows us the form of a linear model.