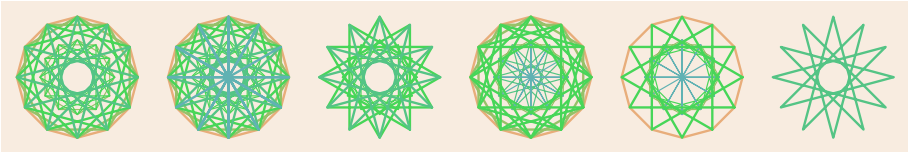
# Quiz 0

CSC316 Machine Learning
Professor Leon Tabak

11 February 2022

This work is licensed under CC BY 4.0. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.



1. People and computer programs can learn from examples. Do people or programs need more examples?

---

Machine learning appears to require more examples than do people.

2. What is an early and widely used application of machine learning that Aurélien Géron identifies in Chapter 1 of *Hands-On Machine Learning*?

---

Géron presents spam filters as an early application of machine learning. The technology dates from the 1990s.

He also mentions Optical Character Recognition (OCR).

3. What is the file format for Jupyter notebooks?

---

Jupyter uses the JSON (JavaScript Object Notation) format.

The file extension is "ipynb." That is an "Interactive Python Notebook."

4. The program in Chapter 2 downloads the California Housing Prices dataset, uncompresses the data, extracts all files, and then copies the data into a Pandas DataFrame..
   (a) What is the format of the downloaded file?
   (b) What is the format of the file that the program extracts from the downloaded file? .

---

(a) It is a "tgz" file. This is a tar file that has been compressed with the GNU zip (gzip) program.

"Tar" is the Unix and Linux 'Tape Archive program. It is a means of packing several or many files in a single files.

GNU zip is a means of reducing the size of a file by reducing the redundancy in the data within the file.

(b) A "csv" file is a "comma separated values" file. It is a text file. It is a way of storing tabular data. Each line in a csv file contains one row of a table. Within a line, commas separate the data in adjacent columns.

5. In a discussion of poor quality data, Géron identifies two ways that data scientists may clean data. What are his two examples?

---

(a) "Poor quality" could mean outliers. An outlier is a suspect, extreme value. A data scientist can remove outliers. It might also be possible to replace outliers in some cases.

(b) "Poor quality" could mean missing values. A data scientist can remove records in a dataset (that is, rows in a table) that contain missing values, remove the variable (that is, a column in a table), or replace missing values (with a mean, median, mode, or constant).

6. The first step in Géron's 8 step checklist for machine learning projects is **Frame the Problem and Look at the Big Picture**.

   (a) Find the word "current" in this section of Appendix B. What does Géron tell us in that sentence?

   (b) Find the word "reuse" in this section of Appendix B. What does Géron tell us in that sentence?

   (c) Find the word "expertise" in this section of Appendix B. What does Géron tell us in that sentence?

---

   (a) A data scientist should look for *current* solutions—has someone else already solved the problem (or part of the problem)?

   (b) A team of data scientists might be able to *reuse* experience and tools (for example, software) that the team (or other teams) acquired and developed on other projects.

   (c) The team should look for human *expertise*. The team might find other people who have studied the same or similar problems. Those other people might be able to help the team build a better solution to a problem.

7. The next to last step (just before **Launch!**) in Géron's 8 step checklist for machine learning projects is **Present Your Solution.**

   (a) Find the word "highlight" in this section of Appendix B. What does Géron tell us in that sentence?

   (b) Find the word "explain" in this section of Appendix B. What does Géron tell us in that sentence?

   (c) Find the word "list" in this section of Appendix B. What does Géron tell us in that sentence?

   (d) Find the word "ensure" in this section of Appendix B. What does Géron tell us in that sentence?

(a) Data scientists should *highlight* the big picture—begin by telling the audience how their efforts support the work of other people in the organization.

(b) Data scientists should *explain* how their work helps the organization move toward its goals.

(c) Data scientists should *list* the assumptions that they made when they built their model, and thereby help the audience understand the model's limitations.

Assumptions that hold at one time might not hold at later times—the world changes. Assumptions that hold in one place or in one application might not hold in others.

Data scientists should advise clients on the circumstances in which a model is most likely to be reliable.

(d) Data scientists should *ensure* that their presentation engages an audiences that will include people who do not posses the data scientists' knowledge of mathematics and computer science, and who may not share the data scientists' fascination with technical and statistical details.

Data scientists should clearly state the most important results.

Data scientists may find the images convey important ideas more effectively than tables of numbers.

Data scientists should prefer plain language and concise statements. They should work hard to compose statements that members of the audience will easily understand and remember.

8. Data scientists may describe RMSE as a cost function or a measure of distance or a vector norm. The symbol $\ell_2$ (that is a small letter L) is a synonym for RMSE.

RMSE is the square root of a sum of differences.

What are the differences? That is, what is subtracted in this function?

---

The function computes a differences between computed predictions and known values.

The known values are the labels in the dataset.

The function computes the square root of the sum of the squares of these differences.

RMSE is a generalization of the formula for finding the length of the hypotenuse of a right triangle. That is also very much like the formula used to find a distance "as the crow flies." RMSE is the Euclidean norm. The distance as the crow flies is the Euclidean distance.

(RMSE differs from the "as the crow flies" formula by the fact that it divides the sum of squares by the number of differences in the sum. RMSE works in any number of dimensions.)

9. The California Housing Prices dataset contains information about the locations of neighborhoods, the ages and sizes of the homes in the those neighborhoods, and the incomes of the people who live in those neighborhoods.

   Are there any hyperparameters in that dataset?

---

There are no hyperparameters in the dataset.

A hyperparameter is value given to a function to control how that function computes the values it returns to its caller.

A hyperparameter is not a part of the description of that part of the world that we are trying to model.

10. We are studying a machine learning model that predicts the prices of homes in California. What marks this as an example of supervised learning?

---

The dataset includes a information about locations, incomes, numbers of rooms in houses, and other factors that might influence the cost of houses.

It also includes the cost of houses in many U.S. Census Bureau *block groups*. These block groups are districts (or neighborhoods) with populations of a few hundred to a few thousand people.

Machine learning in this case means a computer program that examines the descriptions of many thousands of block groups together with the median prices of houses in those districts, recognizes patterns, and then generalizes in ways that allow the prediction of prices in other block groups.

This is supervised learning because the data is labeled—the labels are the median prices.