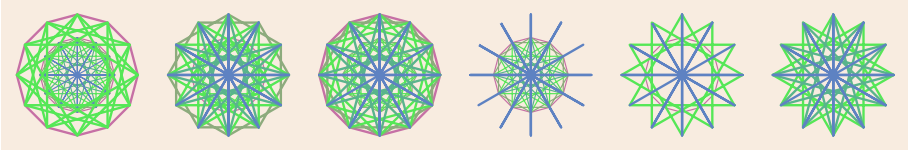
Review

Leon Tabak

24 February 2022



1. Software engineers strive to use three resources economically. These resources are...

   - time—if all other things are equal, we prefer a program that produces a result quickly over one that is slower
   - space—if all other things are equal, we prefer a program that requires less memory over one that requires more memory
   - and what other resource?

   Hints—check these articles:

   - here
   - here
   - here
   - here

2. Read the paragraph that begins at the bottom of page 289 of *Hands-On Machine Learning* and continues on the top of page 290.

   What problem did David RumelHart, Geoffrey Hinton, and and Ronald Williams solve in *Learning Internal Representations by Error Propagation*?

   Briefly, how does their solution work?

3. Here is Equation 4-13 in *Hands-On Machine Learning*. (The equation is in Chapter 4 on page 143.)

$$\begin{aligned}\hat{p} &= h_{\mathbf{\Theta}}(\mathbf{x}) \\ &= \sigma(\mathbf{x}^T\mathbf{\Theta}\end{aligned}$$

   (a) What does $\hat{p}$ signify?

   (b) What does $\mathbf{x}$ signify?

   (c) What does $\mathbf{\Theta}$ signify?

4. Let's stay with Equation 4-13.

   σ is the logistic function. You will also see it refered to as the sigmoid function.

   (a) When $\mathbf{x}^T\mathbf{\Theta} \ll 0.0$ what is the value of $\sigma(\mathbf{x}^T\mathbf{\Theta})$?

   (b) When $\mathbf{x}^T\mathbf{\Theta} = 0.0$ what is the value of $\sigma(\mathbf{x}^T\mathbf{\Theta})$?

   (c) When $\mathbf{x}^T\mathbf{\Theta} \gg 0.0$ what is the value of $\sigma(\mathbf{x}^T\mathbf{\Theta})$?

5. This expression appeared in the explanation of logistic expression and in the explanation of neural networks.

$$\frac{1}{1+e^{-x}}$$

   How might we use this function in a neural network?

   (There is a word that identifies the role of this function in a neural network.)

6. Now look at Equation 4-15 on page 143.

$$\hat{y} = \begin{array}{l} 0 \text{ if } \hat{p} < 0.5 \\ 1 \text{ if } \hat{p} \geq 0.5 \end{array}$$

   (a) What does $\hat{y}$ signify?

   (b) What does $\hat{p}$ signify?

7. Equation 4-17 is the logistic regression cost function (also called the log loss function).

   Here it is written in a different way.

$$f(y^{(i)}, \log \hat{p}^{(i)}) = y^{(i)} \log \hat{p}^{(i)}$$
$$g(y^{(i)}, \log \hat{p}^{(i)}) = (1 - y^{(i)})(1 - \log \hat{p}^{(i)})$$

$$J(\mathbf{\Theta}) = -\frac{1}{m} \sum_{i=1}^{m} [f(y^{(i)}, \log \hat{p}^{(i)}) + g(y^{(i)}, \log \hat{p}^{(i)})]$$

We are using this cost function in a binary classifier. The label on each instance in our dataset is either 0 or 1.

Here, we are using $y^{(i)}$ to signify the label on the i^{th} instance.

We are using $\hat{p}^{(i)}$ to signify the probability that the model predicts that given instance $\mathbf{x}^{(i)}$ is a positive instance ($y = 1$).

(a) Can you see why it never be that both functions $f()$ and $g()$ produce non-zero values for any instance?

(b) Let's say that the label on an instance is 1 and our model predicts a value of 1.

Can you see how this makes the cost zero?

(c) Let's say that the label on an instance is 0 but the model predicts a value of 1.

Can you see how this formula produces a cost with a very large magnitude?

(d) Let's say that the label on an instance is 0 and our model predicts a value of 0.

Can you see how this makes the cost zero?

(e) Let's say that the label on an instance is 1 but the model predicts a value of 0.

Can you see how this formula produces a cost with a very large magnitude?

8. What is the purpose of regularization?

9. Here is the formula for Ridge Regression.

(It is in Equation 4-8 on page 135 in *Hands-On Machine Learning.*)

$$J(\mathbf{\Theta}) = \text{MSE}(\mathbf{\Theta}) + \alpha \frac{1}{2} \sum_{i=1}^{n} \Theta_i^2$$

(a) What is MSE?

(b) Where is the hyperparameter in this function?

(c) You have seen $\sum_{i=1}^{n} \Theta_i^2$ before. What is it?

10. How does the formula for Lasso Regression differ from the formula for Ridge Regression?

11. "Lasso" sounds like a cowboy's tool. In fact, it is an acronym. We often capitalize every letter in an acronym (for example, NASA and CAT scan), but not always (for example, radar).

    From which words did the creators of this method construct this acronym?

    (I am asking you to look this up just once. This is just trivia, not something you need to remember.)

12. How does Ridge Regression affect the elements of $\boldsymbol{\Theta}$, the parameter vector?

13. How is Elastic Net related to Ridge Regression and Lasso Regression?

14. The formula for Elastic Net regression contains two hyperparameters. The value of one of these hyperparameters will always lie in the interval $[0.0, 1.0]$.

    Explain.

15. When should we use regression?

    (You may have to look a little harder for an answer to this question. You will not find a complete answer in one place in *Hands-On Machine Learning*.)

16. *Early stopping* is a kind of regularization. Géron refers to early stopping in his discussion of regularized linear models and in his discussion of neural networks.

    (a) How does early stopping work?

    (b) How did Geoffrey Hinton characterize this method?

17. Describe the iris dataset.

18. Describe the relationship between bias and variance.

19. Identify a relationship between anomaly detection and density estimation.

    (Both anomaly detection and density estimation are problems that can be solved with unsupervised learning.)

20. *Hands-On Machine Learning* contains some code that a data scientist might use to determine the best number of dimensions to choose when reducing dimensions. Find it. Explain the idea—what is the criterion for selecting the "right" number of dimensions?