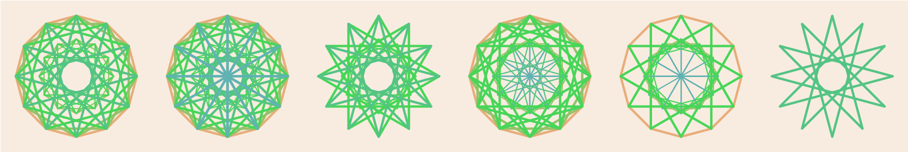
Review

CSC316 Machine Learning Professor Leon Tabak

17 February 2022



1. Which Greek letter are we using to represent a model's parameter vector?
2. Which Latin letter are we using to represent a model's feature vector?
3. How do the lengths of the parameter vector and a feature vector compare?
4. There is more than one way to multiply two vectors. Does a dot product give us a number or another vector?
5. What is $\vec{u} \cdot \vec{v}$?

$$\begin{aligned} \vec{u} &= [1, 2, 3] \\ \vec{v} &= [4, 5, 6] \\ \vec{u} \cdot \vec{v} &= ? \end{aligned}$$

6. What is the derivative of $f(x)$?

$$\begin{aligned} f(x) &= 2 + 3x \\ \frac{d\ f(x)}{dx} &= ? \end{aligned}$$

7. What does this expression signify?

$$\frac{\partial}{\partial \mathbf{\Theta_0}} \; MSE(\mathbf{\Theta})$$

8. What does this expression signify?

$$\nabla_{\mathbf{\Theta}} MSE(\mathbf{\Theta})$$

9. Let us suppose the $\mathbf{X}$ is a matrix with $70,000$ rows and 784 columns.

   $\mathbf{X}^T$ is the transpose of $\mathbf{X}$. It is also a matrix.

   $\mathbf{X}^T\mathbf{X}$ is a product of matrices. The product of two matrices is a matrix.

   $\mathbf{X}^T\mathbf{X}$ is a matrix with how many rows and columns?

10. What is a *local optimum*? Use a geographic example. What might a *local minimum* or *local maximum* look like to a hiker exploring a hilly region?

11. For those of you have studied calculus: What is the derivative of a function where the function reaches a minimum or maximum?

12. A gradient is a vector. Like other vectors, this vector has a norm—that is, it has a length.

    What is the relationship between a gradient's norm and the tolerance ϵ that a data scientist specifies in a call to the constructor of the SGDClassifier class?

13. Why are we using MSE rather than RMSE?

14. What is a *learning rate*?

15. What is a *learning schedule*?

16. Stochastic Gradient Descent resembles *simulated annealing*

17. How does the number of computations required to make a prediction with a linear regression model depend upon...

    (a) the number of feature vectors in the dataset?

    (b) the length of each feature vector?

18. SVD means Singular Value Decomposition. What does word *decomposition* mean in this context?

19. Which algorithm for linear regression performs poorly when the number of training instances is very large?