Editorial-W3A1: Gradient Descent, Learning Rate, PCA, Eigenvectors, and Overfitting

Assignment 1

MCQ

Question 1:

A startup is training a machine learning model using gradient descent. They observe that their model sometimes overshoots the optimal solution and oscillates wildly. What is the most likely reason behind this behavior?

A. Too small a learning rate

B. Incorrect loss function

C. Too high a learning rate

D. Not using regularization

Answer: C

Explanation: In gradient descent, the learning rate determines how large a step the algorithm takes in the direction of the negative gradient. If it's too high, the algorithm might overshoot the minimum and land on the other side, repeating this behavior and causing oscillations. This happens because the updates are too aggressive, and the algorithm keeps jumping back and forth over the minimum, rather than settling down to it. This not only prevents convergence but can also make the model diverge entirely.

Question 2:

During model training, an engineer visualizes the loss surface and notices that the model is stuck at a flat region where gradients are nearly zero. What is this region likely to be?

A. A global minimum

B. A saddle point

C. A local maximum

D. A noisy data point

Answer: B

Explanation: A saddle point is a point on the loss surface where the gradient is zero in all directions, but the point is not a minimum. In some directions, the surface curves upwards (like a maximum), and in others, it curves downwards (like a minimum). Gradient-based optimization algorithms like gradient descent can get stuck here because they rely on gradient direction to make updates, and with gradients close to zero, the model sees no clear path to move, leading to very slow or no progress.

Question 3:

Why might Stochastic Gradient Descent be preferred over Batch Gradient Descent in real-time learning systems?

A. It's more accurate

B. It converges instantly

C. It updates faster using fewer data points

D. It never gets stuck in local minima

Answer: C

Explanation: Stochastic Gradient Descent (SGD) updates the model's parameters using a single randomly selected data point at a time. This allows it to make quick updates, which is ideal for real-time systems where decisions need to be made on-the-fly. While it may not always follow a smooth path to the minimum due to the noise introduced by individual samples, it often escapes shallow local minima and can begin learning immediately from streaming data without needing to process the entire dataset first.

Question 4:

What is the goal of using a loss function in model training?

A. To calculate training speed

B. To predict labels

C. To measure model’s prediction error

D. To apply regularization

Answer: C

Explanation: The loss function serves as a quantitative measure of how far off the model's predictions are from the true values. It guides the learning process by giving the optimization algorithm a value to minimize. Without a loss function, the model wouldn’t have any feedback mechanism to know whether it’s improving. It forms the core of training—every update to the model’s parameters aims to reduce this loss.

Question 5:

During training, Deep uses a very small learning rate and takes a long time to reach the minimum. What is the likely benefit of this approach?

A. Fast convergence

B. Less risk of overfitting

C. Stability in training

D. Reduced regularization

Answer: C

Explanation: Using a small learning rate means the model updates its parameters very slowly. While this results in slower convergence, it helps ensure that each step is carefully taken, reducing the chance of overshooting or skipping over the minimum. This makes training more stable, especially in complex landscapes with many local minima and saddle points, where large steps might destabilize the model or cause erratic behavior.

Question 6:

During a machine learning workshop, a participant uses PCA on a customer dataset. After calculating the eigenvalues, they decide to keep only the top two principal components. What is the most likely reason for this choice?

A. The top two eigenvalues always provide the most accurate results

B. PCA only works with two components

C. The top two eigenvectors represent directions that capture the most variance

D. It's computationally easier to visualize 2D data

Answer: C

Explanation: Principal Component Analysis (PCA) transforms the dataset into a new coordinate system defined by the eigenvectors of the data's covariance matrix. The associated eigenvalues represent the amount of variance captured by each direction. The directions (eigenvectors) with the highest eigenvalues correspond to the most informative components. Keeping the top two components helps simplify the data while retaining most of its structure.

Question 7:

In a crime prediction model, analysts use matrix transformations to understand patterns in spatial data. Over multiple matrix multiplications, a particular direction becomes dominant, regardless of initial conditions. What concept is this illustrating?

A. Overfitting

B. Eigenvector convergence

C. Data normalization

D. Loss minimization

Answer: B

Explanation: When a matrix is repeatedly multiplied by a vector, the result tends to align with the dominant eigenvector—i.e., the one associated with the largest eigenvalue. This is because components in the direction of the dominant eigenvector get amplified more than others. This principle is the basis of the power iteration method and highlights the significance of principal directions in understanding data structure.

Question 8:

A student analyzes stock market trends and uses PCA to extract patterns. They observe that one principal component explains 90% of the data variance. What should they conclude?

A. The dataset is too simple

B. One feature dominates all others

C. Most variability in the data is captured by this direction

D. PCA failed to work

Answer: C

Explanation: A principal component that explains a large proportion of the variance indicates that most of the data's variability is concentrated along one direction. This often means that the data can be effectively represented in a lower-dimensional space without significant loss of information. It's a sign that PCA worked well in finding a meaningful pattern in the dataset.

MSQ

Question 9:

Which of the following statements are true about eigenvalues and eigenvectors in the context of PCA?

A. Eigenvectors define the directions of maximum variance

B. Eigenvalues determine how important each direction is

C. All eigenvalues are always equal in PCA

D. PCA helps in dimensionality reduction using eigenvectors

Answers: A, B, D

Explanation: PCA involves computing the eigenvectors and eigenvalues of the covariance matrix of the data. Eigenvectors give the directions (new axes), while eigenvalues tell how much variance is captured along those axes. PCA uses this information to rank and select the most meaningful directions, enabling dimensionality reduction. It's incorrect to say that all

eigenvalues are equal—variations in eigenvalues are precisely what allow PCA to prioritize features.

Question 10:

What happens during repeated matrix multiplication involving eigenvectors in data analysis?

A. Random outputs are generated

B. The vector converges to the direction of the dominant eigenvector

C. The smallest eigenvalue becomes more important

D. The vector grows in the direction of the eigenvector with the largest eigenvalue

Answers: B, D

Explanation: Repeatedly multiplying a vector by a matrix amplifies components along the eigenvectors of the matrix, with the one associated with the largest eigenvalue growing the fastest. Eventually, the result aligns with the dominant eigenvector. This is key in PCA and other spectral methods, where principal directions are identified through such convergence.

Question 11:

Which situations may indicate that a model is overfitting?

A. Low training loss, high test loss

B. Model performs equally well on all data

C. Model has very large weight values

D. Model uses L2 regularization

Answers: A, C

Explanation: Overfitting happens when a model learns the noise and peculiarities of the training data instead of generalizing to unseen data. This is evidenced by good performance on training data but poor generalization to test data. Large weight values also suggest the model is overly complex, which is a common sign of overfitting. L2 regularization, in contrast, helps prevent overfitting by penalizing large weights.

Question 12:

Which of these techniques help avoid overfitting?

A. Increasing the dataset

B. L1 Regularization

C. Very high learning rate

D. L2 Regularization

Answers: A, B, D

Explanation: Adding more training data gives the model a broader view of the underlying distribution, improving its ability to generalize. L1 and L2 regularizations add penalties to the loss function to discourage overly complex models. L1 can also lead to sparsity, which aids interpretability. A very high learning rate, however, can cause instability and is not a regularization method.

Numeric Type

Question 13:

Assume initial w = 0, learning rate = 0.1, and gradient = -10. What will be the updated w?

Answer: 1.0

Explanation: Using the gradient descent update rule:

w_new = w_old - (learning_rate * gradient)

Substituting the values: w_new = 0 - (0.1 * -10) = 1.0.

The negative gradient means the slope is negative, so the weight is increased in that direction.

Question 14:

Suppose a model took 5 steps with learning rate 0.2 and constant gradient 3. What is the total change in weight magnitude?

Answer: 3.0

Explanation: Each step updates the weight by learning_rate * gradient = 0.2 * 3 = 0.6.

After 5 steps, the total change = 0.6 * 5 = 3.0.

This calculation helps in understanding how parameters evolve during training.

Question 15:

Suppose the learning rate is 0.1 and the gradient at a step is 5. By how much will the weight be updated (magnitude only)?

Answer: 0.5

Explanation: The update magnitude is calculated as learning rate * gradient = 0.1 * 5 = 0.5.

This tells how significantly the model is adjusting in one iteration, helping control the pace of learning.