ML Interview Book Answers

Mihai Anca

Contents

Chapter 1

Math

1.1 Algebra

1.1.1 Vectors

1. Dot product

   i. [E] What's the geometric interpretation of the dot product of two vectors?

   The dot product between two vectors a and b can be seen as the projection of a on b.

   ii. [E] Given a vector u, find vector v of unit length such that the dot product of u and v is maximum.

   The maximum dot product is achieved when the two vectors are going in the same direction. Since v is of unit length, the answer is $v = [1, 1, 1, ...]$.

2. Outer product

   i. [E] Given two vectors $a = [3, 2, 1]$ and $b = [-1, 0, 1]$. Calculate the outer product $a^T b$?

$$\begin{bmatrix} -3 & 0 & 3 \\ -2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix}$$

   ii. [M] Give an example of how the outer product can be useful in ML.

   Incomplete Answer: Error back propagation in multi-layer perceptrons.

3. [E] What does it mean for two vectors to be linearly independent?

Two vectors $v1, v2$ are linearly independent if for any scalars $c1, c2$, $c1 * v1 + c2 * v2 != 0$.

4. [M] Given two sets of vectors $A = a_1, a_2, a_3, ..., a_n$ and $B = b_1, b_2, b_3, ..., b_m$. How do you check that they share the same basis?

Potential (sharing a basis?) incomplete answer: The two vectors would *form* a basis if they are linearly independent with each other. This can be checked by taking the dot product between the two and verifying that it does not equal 0.

5. [M] Given n vectors, each of d dimensions. What is their dimensionality span?

You can treat the vectors as the rows of a matrix. The dimensionality span would be given by the rank of this matrix. The rank is equal to the number of linearly independent rows.

6. Norms and metrics

    i. [E] What's a norm? What is L_0, L_1, L_2, L_{norm}?

    Answer

    ii. [M] How do norm and metric differ? Given a norm, make a metric. Given a metric, can we make a norm?

    Answer

1.1.2 Matrices

1. [E] Why do we say that matrices are linear transformations?

Answer

2. [E] What's the inverse of a matrix? Do all matrices have an inverse? Is the inverse of a matrix always unique?

Answer

3. [E] What does the determinant of a matrix represent?

Answer

4. [E] What happens to the determinant of a matrix if we multiply one of its rows by a scalar $t \times R$?

Answer

5. [M] A 4×4 matrix has four eigenvalues $3, 3, 2, -1$. What can we say about the trace and the determinant of this matrix?

Answer

6. [M] Given the following matrix:

$$\begin{bmatrix} 1 & 4 & -2 \\ -1 & 3 & 2 \\ 3 & 5 & -6 \end{bmatrix}$$

Without explicitly using the equation for calculating determinants, what can we say about this matrix's determinant?

Hint: rely on a property of this matrix to determine its determinant.

Answer

7. [M] What's the difference between the covariance matrix $A^T A$ and the Gram matrix AA^T?

Answer

8. Given $A \in R^{n \times m}$ and $b \in R^n$

   i. [M] Find x such that: $Ax = b$.

   Answer

   ii. [E] When does this have a unique solution?

   Answer

   iii. [M] Why is it when A has more columns than rows, $Ax = b$ has multiple solutions?

   Answer

   iv. [M] Given a matrix A with no inverse. How would you solve the equation $Ax = b$? What is the pseudoinverse and how to calculate it?

   Answer

9. Derivative is the backbone of gradient descent.

   1. [E] What does derivative represent?

   Answer

   1. [M] What's the difference between derivative, gradient, and Jacobian?

   Answer

10. [H] Say we have the weights $w \in R^{d \times m}$ and a mini-batch x of n elements, each element is of the shape $1 \times d$ so that $x \in R^{n \times d}$. We have the output $y = f(x; w) = xw$. What's the dimension of the Jacobian $\frac{\delta y}{\delta x}$?

Answer

11. [H] Given a very large symmetric matrix A that doesn't fit in memory, say $A \in R^{1M \times 1M}$ and a function f that can quickly compute $f(x) = Ax$ for $x \in R^{1M}$. Find the unit vector x so that $x^T Ax$ is minimal.

    Hint: Can you frame it as an optimization problem and use gradient descent to find an approximate solution?

Answer

1.1.3 Dimensionality reduction

1. [E] Why do we need dimensionality reduction?

Answer

2. [E] Eigendecomposition is a common factorization technique used for dimensionality reduction. Is the eigendecomposition of a matrix always unique?

Answer

3. [M] Name some applications of eigenvalues and eigenvectors.

Answer

4. [M] We want to do PCA on a dataset of multiple features in different ranges. For example, one is in the range 0-1 and one is in the range 10 - 1000. Will PCA work on this dataset?

Answer

5. [H] Under what conditions can one apply eigendecomposition? What about SVD?

    i. What is the relationship between SVD and eigendecomposition?

    Answer

    ii. What's the relationship between PCA and SVD?

    Answer

6. [H] How does t-SNE (T-distributed Stochastic Neighbor Embedding) work? Why do we need it?

Answer

1.1.4 Calculus and convex optimization

1. Differentiable functions

    i. [E] What does it mean when a function is differentiable?

    Answer

    ii. [E] Give an example of when a function doesn't have a derivative at a point.

    Answer

    iii. [M] Give an example of non-differentiable functions that are frequently used in machine learning. How do we do backpropagation if those functions aren't differentiable?

    Answer

2. Convexity

i. [E] What does it mean for a function to be convex or concave? Draw it.

Answer

ii. [E] Why is convexity desirable in an optimization problem?

Answer

iii. [M] Show that the cross-entropy loss function is convex.

Answer

3. Given a logistic discriminant classifier:

$$p(y=1|x) = \sigma(w^T x)$$

where the sigmoid function is given by:

$$\sigma(z) = (1 + \exp(-z))^{-1}$$

The logistic loss for a training sample x_i with class label y_i is given by:

$$L(y_i, x_i; w) = -\log p(y_i|x_i)$$

i. Show that $p(y=-1|x) = \sigma(-w^T x)$.

Answer

ii. Show that $\Delta_w L(y_i, x_i; w) = -y_i(1 - p(y_i|x_i))x_i$.

Answer

iii. Show that $\Delta_w L(y_i, x_i; w)$ is convex.

Answer

4. Most ML algorithms we use nowadays use first-order derivatives (gradients) to construct the next training iteration.

i. [E] How can we use second-order derivatives for training models?

Answer

ii. [M] Pros and cons of second-order optimization.

Answer

iii. [M] Why don't we see more second-order optimization in practice?

Answer

5. [M] How can we use the Hessian (second derivative matrix) to test for critical points?

Answer

6. [E] Jensen's inequality forms the basis for many algorithms for probabilistic inference, including Expectation-Maximization and variational inference.. Explain what Jensen's inequality is.

Answer

7. [E] Explain the chain rule.

Answer

8. [M] Let $x \in R_n$, $L = crossentropy(softmax(x), y)$ in which y is a one-hot vector. Take the derivative of L with respect to x.

Answer

9. [M] Given the function $f(x, y) = 4x^2 - y$ with the constraint $x^2 + y^2 = 1$. Find the function's maximum and minimum values.

Answer