

NPTEL
NPTEL ONLINE CERTIFICATION COURSE

Introduction to Machine Learning

Linear Algebra-2

(Refer Slide Time: 00:15)

Eigenvalues & Eigenvectors

- Given a square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, λ is said to be an eigenvalue of \mathbf{A} and vector \vec{x} the corresponding eigenvector if
$$\mathbf{A}\vec{x} = \lambda\vec{x}$$
- Geometrical interpretation**
We can think of the eigenvectors of a matrix A as those vectors which upon being operated by A are only scaled but not rotated.
- Example**
$$A = \begin{bmatrix} 6 & 5 \\ 1 & 2 \end{bmatrix}, \vec{x} = \begin{bmatrix} 5 \\ 1 \end{bmatrix}$$
$$A\vec{x} = \begin{bmatrix} 35 \\ 7 \end{bmatrix} = 7\vec{x}$$

NPTEL
Abhinav Ghorashi, Varun Gopal
Linear Algebra Tutorial
January 19, 2018 16 / 31

The eigenvector and eigenvalue of A spectral wave. So note here then eigenvectors and eigenvalues are tied together which means that any eigenvector has an associated eigenvalue. You often characterize square meters as head down of their eigenvector one way of looking at eigenvector is as follows.

X can be count out for the vector in \mathbb{R}^n and the square meter is A acts like an operator which transforms X into another N dimensional vector AX . Now the eigenvectors of A are those vectors which are being transformed by A or operated upon by A are only scaled by λ but not rotated. In other words that direction does not change.

We can have a look at this example here, the 2x2 matrix A on multiplying the vector X/1 clicks back the vector X multiplied by the real value 7. So here X is an eigenvector of A and 7 is an eigenvalue of A.

(Refer Slide Time: 01:43)

Characteristic Equation

- Trivially, the $\vec{0}$ vector would always be an eigenvector of any matrix. Hence, we only refer only to non-zero vectors as eigenvectors.
- Given a matrix A , how do we find all eigenvalue-eigenvector pairs?

$$A\vec{x} = \lambda\vec{x}$$

$$A\vec{x} - \lambda\vec{x} = 0$$

$$(A - \lambda I)\vec{x} = 0$$

The above will hold iff

$$|(A - \lambda I)| = 0$$

This equation is also referred to as the characteristic equation of A . Solving the equation gives us all the eigenvalues λ of A . Note that these eigenvalues can be **complex**.

NPTEL

Abhinav Garlapati, Varun Goyal Linear Algebra Tutorial January 19, 2016 17 / 31

We can see that 0 would always be an eigenvector of any matrix it reasonably go by the $AX=\lambda X$ definition. Hence we only refer to nonzero vectors and eigenvectors. So the question is given a matrix A out as 1 mild all the eigenvalue, eigenvector bits, by simplifying a sequence λx we get $A-\lambda I$ to $X=0$.

Now since we are only looking at nonzero vectors \det of x cannot be zero and x can be a zero vector which means that \det of $A-\lambda I$ should be zero. So the equation $\det A-\lambda I=0$ is called a characteristic equation of A . So one designation gives this all the eigenvalues of A , one thing you should notice that even though all the value of A are real, A is a real matrix, the eigenvalues can complex.

(Refer Slide Time: 02:55)

Properties

- 1 The trace $\text{tr}(A)$ of a matrix A also equals the sum of its n eigenvalues.
$$\text{tr}(A) = \sum_{i=1}^n \lambda_i$$
- 2 The determinant $|A|$ is equal to the product of the eigenvalues.
$$|A| = \prod_{i=1}^n \lambda_i$$
- 3 The rank of a matrix is equal to the number of non zero eigenvalues of A .
- 4 If A is invertible, then the eigenvalues of A^{-1} are of form $\frac{1}{\lambda_i}$, where λ_i are the eigenvalues of A .

NPTEL
Abhinav Gargapati, Varun Iyengar
Linear Algebra Tutorial
January 19, 2016 31 / 31

There are interesting relations between some properties of a matrix and its eigenvalues. For instance, the trace of a matrix is equal to the sum of its eigenvalue by the determinant is equal to the product. The rank of a matrix is equal to the number of nonzero eigenvalue. Note that if a eigenvalue has multiplicity greater than 1.

For instance, if two distinct eigenvectors X_1 and X_2 both have eigenvalue λ we would count λ twice. Also we can describe the eigenvalues of A^{-1} in terms of the eigenvalues of A provided of course, A is invertible. The eigenvalues of A^{-1} maybe of the form $1/\lambda_i$, where λ_i is an eigenvalue of A .

(Refer Slide Time: 03:55)

Proof

$$\sum_{i=1}^{i=k} a_i v_i = \vec{0}$$

$$(A - \lambda_k I) \sum_{i=1}^{i=k} a_i v_i = \vec{0}$$

$$\sum_{i=1}^{i=k} (A - \lambda_k I) a_i v_i = \vec{0}$$

$$\sum_{i=1}^{i=k} a_i (\lambda_i - \lambda_k) v_i = \vec{0}$$

Since the eigenvalues are distinct, $\lambda_i \neq \lambda_k \forall i \neq k$. Thus the set of $(k-1)$ eigenvectors is also linearly dependent, violating our assumption of it being the smallest such set. This is a result of our incorrect starting assumption. Hence proved by contradiction.

Abhinav Karlapati, Varun Gargal Linear Algebra Tutorial January 19, 2018 20 / 31

Now let us have a look at an interesting theorem about eigenvalues and eigenvectors. The theorem goes as follows. And for matrix as all its eigenvalue is distinct when its eigenvectors are linearly independent which is proof is by what is called proof by contradiction. This theorem does not hold that means there is a set of A eigenvectors such that it is linearly dependent.

That ith vector in the set BBI and the corresponding eigenvalue will have the i. Note that we are considering the smallest such set. Since the set is linearly dependent this means there exists real consonants Ai such set summation Ai Vi=0. Now let us multiply both sides of the equation by A-λk(i). Since Vk is an eigenvector of A, A-λk(i)Vk will be equal to 0, we can understand this from the characteristic equation.

Hence the term corresponding to V_k disappears from the equation since it goes to zero. Now for the remaining eigenvalues since we know they are distinct the term $\lambda_i - \lambda_k$ cannot be equal to 0. Note that $A - \lambda_k I$ applied to V_i simplifies to $(\lambda_i - \lambda_k)V_i$ since $AV_i = \lambda_i V_i$. For we would now, we can think of $(\lambda_i - \lambda_k)V_i$ as a new constant V_i this means now that we have a summation running from $i=1$ to $i=k-1$ such that $\sum_{i=1}^{k-1} V_i = 0$.

However, we can assume that this is the, that the sake of size k was the smallest set of linearly dependent eigenvector. However, now we have an even smaller set, this contradicts starting assumption. Hence, such a set of k linearly dependent eigenvectors cannot exist for any k greater than equal to 2. Hence all are eigenvectors are linearly independent, hence our theorem stand to.

(Refer Slide Time: 06:53)

Diagonalization

Given a matrix A , we consider the matrix S with each column being an eigenvector of A

$$S = \begin{bmatrix} | & | & \dots & | \\ v_1 & v_2 & \dots & v_n \\ | & | & \dots & | \end{bmatrix}$$

$$AS = \begin{bmatrix} | & | & \dots & | \\ \lambda_1 v_1 & \lambda_2 v_2 & \dots & \lambda_n v_n \\ | & | & \dots & | \end{bmatrix}$$

$$AS = \begin{bmatrix} | & | & \dots & | \\ v_1 & v_2 & \dots & v_n \\ | & | & \dots & | \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \dots \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_n \end{bmatrix}$$

NPTEL | Anil Kumar, IIT Madras | Linear Algebra Tutor | January 14, 2016 | 23 / 31

Diagonalization gives us a way of representing a matrix in terms of its eigenvalues and eigenvectors. Let us consider a $N \times N$ matrix A where the amount of matrix where every column is an eigenvector of A/S . On multiplying S/A each column would get multiplied by λ_i since the column itself is an eigenvector of A .

This right hand side can then be simplified and the product of two matrices. The first one means itself by the second one B the diagonal matrix where the i^{th} diagonal matrix the eigenvalue λ_i . Remember the DLH is AS .

(Refer Slide Time: 07:50)

Diagonalization

$$AS = S\Lambda$$

$$A = S\Lambda S^{-1}$$

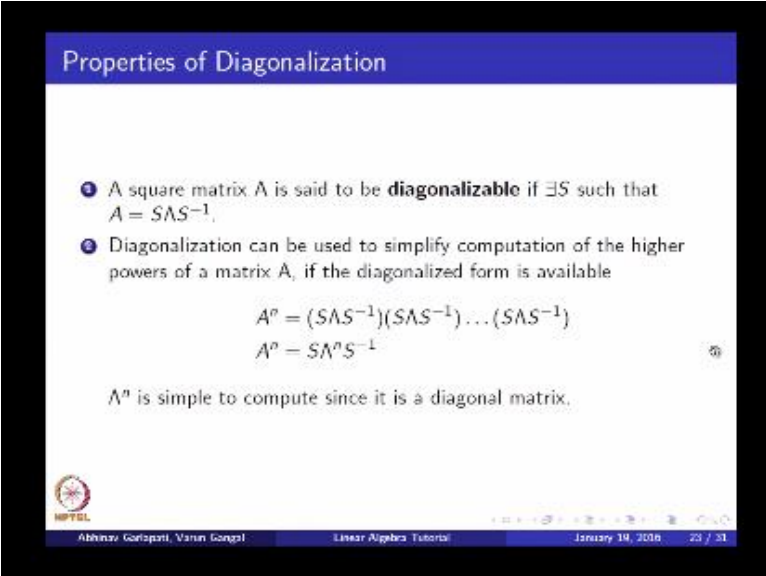
- $S^{-1}AS$ is diagonal
- Note that the above result is dependent on S being invertible. In the case where the eigenvalues are distinct, this will be true since the eigenvectors will be linearly independent

NPTEL
 Abhinav Gorlapati, Varun Goyal Linear Algebra Tutorial January 18, 2016 22 / 33

Now we have the equation $AS = S\lambda$ where λ is the diagonal matrix of eigenvalues. On simplifying this we get $A = S\lambda S^{-1}$, this is a diagonalization of A . Note that $S^{-1}AS$ is a diagonal matrix since $S^{-1}AS$ is nothing but λ the diagonal matrix of eigenvalues. This result is dependent on S being invertible.

It will be clear with the eigenvalue of a matrix at distinct. Since the eigenvectors would then be linearly independent. This would mean the columns of S would be linearly independent, and hence S would be invertible and as a consequence invertible.

(Refer Slide Time: 08:59)



The slide is titled "Properties of Diagonalization" and contains the following content:

- 1 A square matrix A is said to be **diagonalizable** if $\exists S$ such that $A = SAS^{-1}$.
- 2 Diagonalization can be used to simplify computation of the higher powers of a matrix A , if the diagonalized form is available

$$A^n = (SAS^{-1})(SAS^{-1}) \dots (SAS^{-1})$$
$$A^n = S\Lambda^n S^{-1}$$

A^n is simple to compute since it is a diagonal matrix.

At the bottom of the slide, there is a logo for NPTEL (National Programme on Technology Enhanced Learning) and the text "Abhinav Garg, VIT-IT, VIT-IT". The footer also includes "Linear Algebra Tutor" and "January 19, 2006 25 / 31".

Then do we see that the square matrix is diagonalizable. Well when such a diagonalization exist we saw that we needed S to be invertible for the diagonalization to exist. Another advantage of diagonalization is that it simplifies the process of computing A^n , the first represent every A in diagonalized form.

Now you can see that the S^{-1} of the first term and the S of the second term would multiply to give us I . Similarly for the second toward, third, fourth and so on, in this way by regrouping the terms we get $A^n = S\Lambda^n S^{-1}$. Note that it is very easy to compute the n th power of a diagonal matrix. Since you just have to realize every diagonal element to the power of n . In this way the

diagonalization has left us simply by the process of computing A^n without the simplification we would have needed to multiply a non-diagonal matrix 10 times.

(Refer Slide Time: 10:30)

The slide is titled "Eigenvalues & Eigenvectors of Symmetric Matrices". It contains the following content:

- Two important properties for a symmetric matrix A :
 - All the eigenvalues of A are real
 - The eigenvectors of A are orthonormal, i.e., matrix S is orthogonal. Thus, $A = SAS^T$.
- Definiteness of a symmetric matrix depends entirely on the sign of its eigenvalues. Suppose $A = SAS^T$, then

$$x^T Ax = x^T SAS^T x = y^T \Lambda y = \sum_{i=1}^n \lambda_i y_i^2$$
- Since $y_i^2 > 0$, sign of expression depends entirely on the λ_i 's. For example, if all $\lambda_i > 0$, then matrix A is positive definite.

At the bottom of the slide, there is an NPTEL logo and footer text: "Abhinav Gargapati, Varun Gargal | Linear Algebra Tutorial | January 19, 2018 | 26 / 31".

If a matrix is symmetric then all its eigenvalues are real numbers. Also if its eigenvectors are also normal that is they are mutually orthogonal and normalized. This means that the matrix of eigenvectors S is also orthogonal. We have seen that for orthogonal matrices that inverse and a transpose are the same.

Hence we can write $A = S\Lambda S^T$ as when the diagonalization we defined earlier. For symmetric matrices the definiteness can be inferred from the signs of their eigenvalue. Suppose that $A = S\Lambda S^T$ now taking the quadratic form with respect to A for the vector X , $X^T A X$ simplifies to $Y^T \lambda y$, thereby is $S^T X$.

This further simplifies to sum over $\lambda_i y_i^2$. Now for a matrix to be positive definite this term must always be positive. Since y_i^2 is always greater than 0 anyway this λ_i of this term depends on the eigenvalue. All the eigenvalues are positive, the matrix is positive definite.

(Refer Slide Time: 12:12)

Eigenvalues of a PSD Matrix

Consider a positive semi definite matrix A . Then, $\forall \vec{x}$ which are eigenvectors of A ,

$$\vec{x}^T A \vec{x} \geq 0$$

$$\lambda \vec{x}^T \vec{x} \geq 0$$

$$\lambda \|\vec{x}\|^2 \geq 0$$

Hence, all eigenvalues of a PSD matrix are non-negative.

NPTEL

Abhinav Choudhary, Varun Iyengar Linear Algebra Tutorial January 19, 2016 25 / 31

If we know that the matrix is positive semi definite or PSD then what can we say about its eigenvalues. Since the quadratic form of a PSD matrix is non-negative for any vector X this should hold for the eigenvectors too. Now since $AX = \lambda X$ $X^T AX$ simplifies to λ norm of X^2 greater than equal to 0.

Since eigenvectors are nonzero by definition the square of the norm is always positive. This means that every eigenvalue of A is non-negative.

(Refer Slide Time: 13:03)

Singular Value Decomposition

- 1 We saw that diagonalization is applicable only to square matrices. We need some analogue for rectangular matrices too, since we often encounter them, e.g the Document-Term matrix. For a rectangular matrix, we consider left singular and right singular vectors as two bases instead of a single base of eigenvectors for square matrices.
- 2 The Singular Value Decomposition is given by $A = U\Sigma V^T$ where $U \in R^{m \times m}$, $\Sigma \in R^{m \times n}$ and $V \in R^{n \times n}$.

NPTL

Abhinav Garlapati, Varun Goyal Linear Algebra Tutorial January 19, 2016 26 / 31

We looked about diagonalization which stood in our square matrix of size $n \times n$ and represented it in terms of its eigenvectors. However, we cannot directly apply by the same diagonalization for rectangular matrices, since the notion of eigenvector is defined only for the square matrix. They need a diagonalization for rectangular matrices since it come to them often.

For instance, the matrix of N data points out integers or the matrix of n documents and R terms. For the rectangular matrix A or size $m \times n$ we can represent it in terms of the eigenvectors of AT^T

and $A^T A$ out of which our square matrices. This is known as the singular value decomposition A is represented as $U\Sigma V^T$ where U is a $m \times m$ matrix Σ is a $m \times n$ matrix and V is a $n \times n$ matrix.

(Refer Slide Time: 14:20)

Singular Value Decomposition

- 1 U is such that the m columns of U are the eigenvectors of AA^T , also known as the left singular vectors of A .
- 2 V is such that the n columns of V are the eigenvectors of $A^T A$, also known as the right singular vectors of A .
- 3 Σ is a rectangular diagonal matrix with each element being the square root of an eigenvalue of AA^T or $A^T A$.

Significance: SVD allows us to construct a lower rank approximation of a rectangular matrix. We choose only the top r singular values in Σ , and the corresponding columns in U and rows in V^T .

NPTEL
Abhinav Gargapat, Varun Gargal Linear Algebra Tutorial January 10, 2016 27 / 31

The three N is $U\Sigma V$ are as follows. In U every column represent an eigenvector of AA^T , in V every column represents as eigenvector or $A^T A$ Σ is a rectangular diagonal matrix if each elopement being described of an eigenvalue of AA^T or $A^T A$. Now note that AA^T and $A^T A$ have different eigenvectors with the set of eigenvalues is the same.

This is because suppose $A^T AX = \lambda X$ for some eigenvector X and eigenvalue λ . Now multiplying both side by A we get AA^T whereas $AX = \lambda AX$ hence AX is an eigenvector of AA^T while λ is also an eigenvalue of AA^T , this is why AA^T and $A^T A$ have the same set of eigenvalues. The

significance of this decomposition is that we all know in U , V and Σ such that the eigenvalue is larger come first both in U and V at the column or and also along the diagonal in Σ .

Then we can drop everything greater than index R to get a R dimension and load and approximation of the original matrix A . Since approximate form of A we represented as U which is an $m \times r$ matrix, Σ which is a $r \times r$ matrix, and V which is a $n \times r$ matrix.

(Refer Slide Time: 16:26)

Matrix Calculus

- 1 **The Gradient**
Consider a function $f : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}$. The gradient $\nabla_A f(A)$ denotes the matrix of partial derivatives with respect to every element of the matrix A . Each element is given by $(\nabla_A f(A))_{ij} = \frac{\partial f(A)}{\partial A_{ij}}$.
- 2 **The Hessian**
Suppose a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ takes in vectors and returns real numbers. The Hessian, denoted as $\nabla_x^2 f(x)$ or H is the $n \times n$ matrix of partial derivatives. $(\nabla_x^2 f(x))_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}$. Note that the Hessian is always symmetric.
- 3 Note that the Hessian is not the gradient of the gradient, since the gradient is a vector, and we cannot take the gradient of the vector. However, if we do take elementwise gradients of every element of the gradient, then we can construct the Hessian.

NPTEL
Abhinav Ghoropati, Varun Goyal Linear Algebra Tutorial January 21, 2018 28 / 31

Consider function F which takes in matrix systems of dimension $m \times n$ and outputs real of course. The gradient is the matrix of partial derivatives. The i, j element of $\Delta F(A)$ or the gradient of $F(A)$ is the partial derivative of $F(A)$ with respect to A_{ij} . Consider it with time of function which takes in at the in dimensional vector and returns a real number.

The Hessian for this function is defined as follows, the i, j the element of the Hessian is given by first differentiating $F(X)$ with respect to the j^{th} component of X , X_j and then the i^{th} component X_i . We can see that the Hessian would be $n \times n$ matrix.

(Refer Slide Time: 17:27)

Differentiating Linear and Quadratic Functions

If $f(x) = b^T x$, for some constant $b \in \mathbb{R}^n$. Let us find the gradient of f .

$$f(x) = \sum_{i=1}^n b_i x_i$$
$$\frac{\partial f(x)}{\partial x_k} = b_k$$

We can see that $\frac{\partial b^T x}{\partial x} = b$. We can intuitively see how this relates to differentiating $f(x) = ax$ with respect to x when a and x are real scalars.

NPTEL
Abhinav Ghorupati, Varun Iyengar | Linear Algebra Tutorial | January 21, 2016 | 28 / 31

Now let us study how will you find the gradient for some simple vector functions. Consider the function $F(X) = B^T X$ where X is an n dimensional vector and B is also an n dimensional vector. $F(X)$ can be written down as sum over $i=1$ to $i=n$ $B_i X_i$. On differentiating this with respect to the k^{th} component of the vector X we can do $F(X)$ by $\partial F(X) = B_k$.

The gradient of $F(X)$ is given by the vector V , you can see how this intuitively remains to the first derivative of the scalar function $F(X) = AX$ which is equal to A .

(Refer Slide Time: 18:27)

Differentiating Linear and Quadratic Functions

Consider the function $f(x) = x^T A x$ where $x \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n \times n}$ is a known symmetric matrix.

$$f(x) = \sum_{i=1}^n \sum_{j=1}^n A_{ij} x_i x_j$$

$$\frac{\partial f(x)}{\partial x_k} = \frac{\partial}{\partial x_k} \left[\sum_{i \neq k} \sum_{j \neq k} A_{ij} x_i x_j + \sum_{i \neq k} A_{ik} x_i x_k + \sum_{j \neq k} A_{kj} x_k x_j + A_{kk} x_k^2 \right]$$

$$\frac{\partial f(x)}{\partial x_k} = \sum_{i \neq k} A_{ik} x_i + \sum_{j \neq k} A_{kj} x_j + 2A_{kk} x_k$$

$$\frac{\partial f(x)}{\partial x_k} = \sum_{i=1}^n A_{ik} x_i + \sum_{j=1}^n A_{kj} x_j$$

$$\frac{\partial f(x)}{\partial x_k} = 2 \sum_{i=1}^n A_{ki} x_i$$

NPTEL
Abhinav Garg, Varun Iyengar | Linear Algebra Tutorial | January 21, 2016 | 30 / 31

We had earlier looked at a type of function called the quadratic form defined for an $n \times n$ matrix A . The quadratic form with respect to matrix A is a function $F(X) = X^T A X$ so it takes in an n -dimensional vector X . Now let us have a look at how one can find the gradient and Hessian for the quadratic form of a known symmetric matrix A .

They can write down $F(X)$ as $\sum_{i=1}^n \sum_{j=1}^n A_{ij} x_i x_j$. We can split up this summation into four terms based on whether i and j are equal or not equal to k . Finally we get

$\partial F(X)$ for $\partial X K = Y$ sum over $i=1, i=n$ $A_{ki} X_i$. Note that the simplification from the second last step with the last step can only be done if A is symmetric.

(Refer Slide Time: 19:41)

Differentiating Linear and Quadratic Functions

Thus $\nabla_x(x^T Ax) = 2Ax$. Now, let us find the Hessian H .

$$\frac{\partial}{\partial x_k} \frac{\partial f(x)}{\partial x_l} = \frac{\partial}{\partial x_k} \left(2 \sum_{i=1}^{i=n} A_{li} x_i \right) = 2A_{kl}$$

Hence, $\nabla_x^2(x^T Ax) = 2A$.

NPTEL
Abhinav Goyal, Varun Goyal | Linear Algebra Tutorial | January 23, 2016 | 31 / 31

Thus we get the gradient of $X^T AX = AX$. Similarly, on further differentiating every element of the gradient by XK we can drive the Hessian of the function. The Hessian of this function comes out to be $2A$.

IIT Madras Production

Funded by

Department of Higher Education

Ministry of Human Resource Development

Government of India

www.nptel.ac.in

Copyrights Reserved