Linear Algebra / Statistics Review

Stats 503

Prof. Liza Levina

Vectors and Matrices

Vector

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \qquad y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

Matrix

$$A_{[n\times m]} = \begin{pmatrix} A_{11} & \cdots & A_{1m} \\ \vdots & \ddots & \vdots \\ A_{n1} & \cdots & A_{nm} \end{pmatrix}$$

The identity matrix

$$I_{[n\times n]} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ & & \ddots & \\ 0 & \cdots & 0 & 1 \end{pmatrix}$$

Matrix multiplication

$$
\begin{aligned}
A_{[n\times m]}B_{[m\times k]} &= \begin{pmatrix} A_{11} & \cdots & A_{1m} \\ \vdots & \ddots & \vdots \\ A_{n1} & \cdots & A_{nm} \end{pmatrix} \cdot \begin{pmatrix} B_{11} & \cdots & B_{1k} \\ \vdots & \ddots & \vdots \\ B_{m1} & \cdots & B_{mk} \end{pmatrix} \\
&= \begin{pmatrix} \cdots & \cdots & \cdots \\ \cdots & \sum_{\ell=1}^{m} A_{i\ell}B_{\ell j} & \cdots \\ \cdots & \cdots & \cdots \end{pmatrix}_{[n\times k]}
\end{aligned}
$$

Example:

Matrix transpose

$$A' = A^T_{[n\times m]} = \begin{pmatrix} A_{11} & A_{21} & \cdots & A_{n1} \\ A_{12} & A_{22} & \cdots & A_{n2} \\ \vdots & \ddots & \ddots & \vdots \\ A_{1m} & A_{2m} & \cdots & A_{nm} \end{pmatrix}_{[m\times n]}$$

Properties:

$$(A^T)^T = A$$
$$(A+B)^T = A^T + B^T$$
$$(AB)^T = B^T A^T$$

Example:

Norm and inner product

- Vector norm (length): $\|x\| = \sqrt{x_1^2 + \cdots + x_n^2} = \sqrt{x^T x}$
- Inner product: $x^T y = x_1 y_1 + x_2 y_2 + \cdots x_n y_n$
- Geometric interpretation: $\cos\theta = \frac{x^T y}{\|x\| \cdot \|y\|}$
- x and y are orthogonal if $x^T y = 0$
- A set of vectors $x_1, \ldots, x_m$ is orthonormal if $x_i^T x_j = 0$ if $i \neq j$ and $\|x_i\| = 1$ for all i.

Example:

Vector spaces

- Vectors $x_1, \ldots x_m$ are linearly dependent if there exist scalars $a_1, \ldots a_m$ such that at least one $a_j \neq 0$ and

$$a_1 x_1 + \cdots + a_m x_m = 0$$

- A vector space spanned by vectors $x_1, \ldots x_m$ is

$$\begin{aligned} \mathscr{X} &= \text{span}(x_1, \ldots, x_m) \\ &= \{x : x = a_1 x_1 + a_2 x_2 + \cdots + a_m x_m, a_j \in \mathbb{R}\} \end{aligned}$$

- A minimal set of such vectors is called a basis (all linearly independent)

Basis examples

Give a basis of the space spanned by

$$x_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \quad x_2 = \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix} \quad x_3 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$$

Give a basis of R^2:

Functions of matrices: determinant, trace, norms

Determinant (square matrix): $\det(A)$ (complicated formula)

- For diagonal matrices, $\det(A) = A_{11} \times \cdots \times A_{nn}$
- $\det(AB) = \det(A)\det(B)$

Trace (square matrix): $\mathrm{trace}(A) = A_{11} + \cdots + A_{nn}$

- $\mathrm{trace}(A+B) = \mathrm{trace}(A) + \mathrm{trace}(B)$
- $\mathrm{trace}(AB) = \mathrm{trace}(BA)$

Matrix norms: many choices. For example, the Frobenius matrix norm

$$\|A\|_F^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} A_{ij}^2 = \mathrm{trace}(AA^T)$$

Example:

Matrix inverse (square matrix)

- Defined by $A^{-1}A = AA^{-1} = I$
- If the columns (rows) of A are linearly independent, then A is invertible, i.e. A^{-1} exists; otherwise, A is singular, and $\det(A) = 0$.

Properties

- $(AB)^{-1} = B^{-1}A^{-1}$
- $(A^{-1})^T = (A^T)^{-1} = A^{-T}$
- $(AB)^{-T} = A^{-T}B^{-T}$
- $\det(A^{-1}) = (\det(A))^{-1}$

Orthogonal matrix: $A^{-1} = A^T$. Orthogonal matrices have orthonormal columns.

Eigendecomposition

- Eigenvectors and eigenvalues: there exists a vector $x \neq 0$

$$A_{[n\times n]}x_{[n\times 1]} = \lambda x_{[n\times 1]}$$

- All eigenvalues are roots (real or complex) of the equation

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

  There are exactly n eigenvalues (not necessarily distinct)
- If A is real-valued and symmetric ($A^T = A$), all eigenvalues are real
- A is singular iff at least one of the $\lambda = 0$
- A is positive definite ($x^T Ax > 0$ for all $x \neq 0$)$\Leftrightarrow$ all eigenvalues $\lambda > 0$. Non-negative definite: all $\lambda \geq 0$.
- $\det(A) = \lambda_1 \times \cdots \times \lambda_n$
- $\text{trace}(A) = \lambda_1 + \cdots + \lambda_n$

- If x is an eigenvector, so is cx for any $c \neq 0$, so require $\|x\| = 1$
  Check:

- Collecting all eigenvectors into a $n \times n$ matrix Q (each column is an eigenvector) and writing $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, we have

$$AQ = Q\Lambda$$

- If Q is invertible, we say A is diagonalizable. For real symmetric matrices, this is always true, and the eigenvectors form a basis of R^n. The canonical eigenvectors are orthonormal, and $Q^{-1} = Q^T$. Then we have the eigendecomposition

$$A = Q\Lambda Q^T$$

Examples

Find the eigendecomposition of

1. $\begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$

2. $\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$

Singular value decomposition

For any real rectangular $m \times n$ matrix A, there exist orthogonal square matrices $U_{[m\times m]}$ and $V_{[n\times n]}$ such that $A = U\Sigma V^T$, where

$$\Sigma = \begin{pmatrix} \mathrm{diag}(d_1, \ldots, d_r) & 0_{[r\times(n-r)]} \\ 0_{[(m-r)\times r]} & 0_{[(m-r)\times(n-r)]} \end{pmatrix},$$

- By convention, $d_1 \geq d_2 \geq \cdots \geq d_r > 0$.
- The rank of A is r
- Diagonal elements of Σ are square roots of the eigenvalues of $A^T A$.
- Columns of V (right singular vectors) are the eigenvectors of $A^T A$.
- Columns of U (left singular vectors) are the eigenvectors of AA^T.
- The best rank k approximation to A in Frobenius norm can be obtained by keeping the first k singular values and replacing the rest by 0 in Σ.

Example: the Procrustes problem

- Goal: apply translation and rotation to data matrix Y to make it as similar as possible to the "target" data matrix X
- Applications: the registration problem (imaging, brain scans, etc)
- After translation so that X and Y have the same "center" (i.e. mean), the problem is to minimize

$$\|X - YR\|_F \text{ subject to } R^T R = I$$

- R is a rotation matrix
  - Rx is the rotated vector x
  - $R^{-1} = R^T$, or $RR^T = I$ (rotating there and back leaves you in the same place)
  - Any orthonormal matrix represents a rotation
- Turns out the Procrustes solution is given by $R = UV^T$, where $U\Sigma V^T$ is the SVD of Y^TX (see Michailidis "Linear algebra review" chapter for derivation, p. 129).

Mean and variance

For a random variable X,

$$
\begin{aligned}
E(X) &= \int x dP(x) \quad \left[= \int x f(x) dx \text{ or } \sum_i x_i P(X = x_i)\right] \\
Var(X) &= E(X - E(X))^2 \\
Cov(X_1, X_2) &= E[(X_1 - E(X_1))(X_2 - E(X_2))] \\
Corr(X_1, X_2) &= \frac{Cov(X_1, X_2)}{\sqrt{Var(X_1)Var(X_2)}}
\end{aligned}
$$

If X_1 and X_2 are independent, $Cov(X_1, X_2) = Corr(X_1, X_2) = 0$.

$$
\begin{aligned}
E(aX) &= aE(X) \\
Var(aX) &= a^2 Var(X) \\
Cov(aX_1, bX_2) &= abCov(X_1, X_2)
\end{aligned}
$$

Mean and variance of vectors

Vector of r.v.s: $X = (X_1, \ldots X_p)'$
Mean: $E(X) = (E(X_1), \ldots E(X_p))'$
Variance-covariance matrix:

$$\begin{aligned} Var(X) &= Cov(X) = E[(X - E(X))(X - E(X))'] = \\ &= \begin{pmatrix} Var(X_1) & Cov(X_1, X_2) & \cdots & Cov(X_1, X_p) \\ Cov(X_2, X_1) & Var(X_2) & \cdots & Cov(X_2, X_p) \\ \vdots & & \ddots & \vdots \\ Cov(X_p, X_1) & \cdots & \cdots & Var(X_p) \end{pmatrix} \end{aligned}$$

If X_j's are independent, $Var(X)$ is diagonal.
If A is a constant (non-random) matrix, then

$$\begin{aligned} E(AX) &= AE(X) \\ Var(AX) &= AVar(X)A^T \end{aligned}$$

Derivatives of vector-valued functions

- Argument $x_{[m\times 1]} = (x_1, \dots x_m)'$
- Function $y_{[n\times 1]} = f(x_{[m\times 1]})$;
- Derivative $D_{[m\times n]} = \partial y / \partial x$ is defined by

$$D_{ij} = \frac{\partial y_j}{\partial x_i}$$

- For a rectangular matrix A,

$$\frac{\partial (Ax)}{\partial x} = A^T$$

- For a square matrix B,

$$\frac{\partial (x^T Bx)}{\partial x} = (B + B^T)x$$

Multivariate normal distribution

A vector $X = (X_1, \ldots, X_p)'$ is multivariate normal if any of the below holds:

- Every linear combination $\sum_i a_i X_i$ has a univariate normal distribution
- $X = \mu + AZ$, where Z is a vector of i.i.d. univariate normals $N(0,1)$. Then $E(X) = \mu$, $Var(X) = AA^T$.
- The joint density of X is given by

$$f(x) = \frac{1}{(2\pi)^{p/2} \det(\Sigma)^{1/2}} \exp\left(-\frac{1}{2} (X-\mu)^T \Sigma^{-1} (X-\mu) \right)$$

  where μ is any p-vector and Σ is symmetric and non-negative definite. Then we write $X \sim N_p(\mu, \Sigma)$, and $E(X) = \mu$, $Var(X) = \Sigma$.

Example: bivariate normal

$$X \sim N_2\left(\begin{pmatrix} 1 \\ -1 \end{pmatrix}, \begin{pmatrix} 1 & -1 \\ -1 & 4 \end{pmatrix}\right)$$

Properties of the multivariate normal

- Contours of the multivariate density are ellipsoids
- $\Sigma = I$ correponds to a spherical distribution
- $\Sigma_{ij} = 0$ iff X_i and X_j are independent
- The multivariate normal is the only distribution where correlation 0 is equivalent to independence
- $\Sigma_{ij}^{-1} = 0$ iff X_i and X_j are *conditionally* independent given all other variables.
- For any vector a and matrix $B \neq 0$, $a + BX$ is also multivariate normal.

Conditional distribution for the multivariate normal

- Partition $X_{[p\times 1]}$ into two parts $(X_{1_{[p_1\times 1]}}, X_{2_{[p_2\times 1]}})$, corresponding to

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

- Then the conditional distribution of X_1 given $X_2 = x$ is also multivariate normal
- The conditional mean is

$$E(X_1|X_2 = x) = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x - \mu_2)$$

- The conditional variance-covariance matrix is

$$Var(X_1|X_2 = x) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

Example

Let X be bivariate normal where both variables have mean 1, variance 2, and their correlation is 0.5. Find
(a) The distribution of $X_1 + X_2$
Scalar solution:

Vector solution:

(b) Find the conditional distribution of X_1 at $X_2 = 0$

Practice

- Pair up with a neighbor
- Make up a value for the mean (2×1 vector) and the covariance matrix (2×2 symmetric positive definite matrix) of a bivariate normal.
- Swap the parameters you made up with your neighbor. Sketch an ellipsoid representing a contour of **their** distribution. Make sure you show the scale of the axes.