Recitation: A Review of Linear Algebra and Probability Theory

6.864 Advanced Natural Language Processing

Wei Fang

MIT CSAIL

Intro

- In this recitation we will *very quickly* go over the basic mathematical tools needed in the course:
  linear algebra and *probability theory*
- They will mostly be definitions and statements, and we will not go through any derivation/proofs.
- Recitation will be recorded; slides & notebook will be uploaded
- Feel free to skip if already familiar
- Resources at MIT:
  - 18.06 Linear Algebra
  - 18.05 Intro to Probability and Statistics

Table of contents

Scalars & Vectors

Scalar

a single number, e.g. $s \in \mathbb{R}$

Vector

1-d array of numbers of the form $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n$

Matrices

Matrix

2-d rectangular array of numbers of the form

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \in \mathbb{R}^{m \times n}$$

The entry in the i-th row and j-th column of $\mathbf{A}$ is most commonly denoted as a_{ij},$a_{i,j}$,or $(\mathbf{A})_{i,j}$.

A vector $\mathbf{x} \in \mathbb{R}^n$ can also be represented in the form of a column vector, which is an $n \times 1$ matrix, or a row vector, which is an $1 \times n$ matrix. Generally we use the column vector format and write them in the form of

$$\mathbf{x} = [x_1 \ \cdots x_n]^T.$$

Basic Matrix Operations

Addition

$A + B$: $(A + B)_{i,j} = A_{i,j} + B_{i,j}$

Scalar Multiplication

cA: $(cA)_{i,j} = c \cdot A_{i,j}$

Transposition

A^T: $(A^T)_{i,j} = (A)_{j,i}$

Matrix Multiplication

$C = AB$: $(C)_{i,j} = \sum_k (A)_{i,k}(B)_{k,j}$

Dot Product

Dot Product

The dot product between vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ is defined as:

$$\mathbf{x} \cdot \mathbf{y} = \sum_{i=1}^{n} x_i y_i.$$

If the vectors are identified with column vectors ($n \times 1$), it can also we written as a matrix product $\mathbf{x}^T\mathbf{y}$.

Commutative

$$\mathbf{x}^T\mathbf{y} = \mathbf{y}^T\mathbf{x}.$$

System of Linear Equations

For $\mathsf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{x} \in \mathbb{R}^{n \times 1}$, and $\mathbf{b} \in \mathbb{R}^{m \times 1}$,

$$\mathsf{A}\mathbf{x} = \mathbf{b}$$

is equivalent to the system of linear equations

$$\begin{aligned} \mathsf{A}_{1,1}x_1 + \mathsf{A}_{1,2}x_2 + \cdots + \mathsf{A}_{1,n}x_n &= b_1 \\ &\vdots \\ \mathsf{A}_{m,1}x_1 + \mathsf{A}_{m,2}x_2 + \cdots + \mathsf{A}_{m,n}x_n &= b_m \end{aligned}$$

Matrix Inverse

Identity Matrix

square matrix with 1 in diagonal entries and 0 otherwise:

$$\mathbf{I}_n \in \mathbb{R}^{n \times n},\ \forall\, x \in \mathbb{R}^n,\ \mathbf{I}_n x = x.$$

Matrix Inverse

The matrix inverse of $\mathbf{A}$, if it *exists*, is the matrix such that

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_n$$

Now we can solve the linear system $\mathbf{A}\mathbf{x} = \mathbf{b}$:

$$\mathbf{A}^{-1}\mathbf{A}\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$$
$$\mathbf{I}_n\mathbf{x} = \mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$$

Vector Space

So far we have talked about these array-like objects, some of its operations and properties, and its use in solving linear equations. Now we are going to connect them to the notion of vector spaces:

Vector Space

A vector space is a set of vectors on which two operations are defined:

- *addition*: takes vectors $\mathbf{x}$ and $\mathbf{y}$, gives unique element $\mathbf{x} + \mathbf{y}$ in the set.
- *scalar multiplication*: takes scalar a and vector $\mathbf{x}$, gives unique element $a\mathbf{x}$ in the set.

There are 8 properties such as associativity and commutativity that must hold (we won't discuss this here). An example of a vector space is the Euclidean space.

Span

Linear Combination

Given vectors $\mathbf{x}_1, \dots, \mathbf{x}_n$ and scalars $a_1, \dots, a_n$, then the linear combination of those vectors with scalars as coefficients is

$$a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \cdots + a_n\mathbf{x}_n.$$

Span

Given a set of vectors $S = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, the span of S is

$$\operatorname{span}(S) = \{a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \cdots + a_n\mathbf{x}_n : \; a_1, \dots, a_n \in \mathbb{R}\}.$$

Linear Dependence & Independence

Linear Dependence

A sequence of vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is linearly dependent if there exists scalars $a_1, \ldots, a_n$, not all zero, such that

$$a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \cdots + a_n\mathbf{x}_n = \mathbf{0}.$$

Linear Independence

A sequence of vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is linearly independent if such nonzero scalars $a_1, \ldots, a_n$ do not exist.

Subspace, Bases & Dimensions

Subspace

A subset W of a vector space V is a subspace of V if W is a vector space under the operations of V.

E.g. Let V be $\mathbb{R}^3$ and W be the set of vectors in V whose last component is 0, then W is a subspace of V.

Basis

A basis B of vector space V is a *linearly independent* subset of V that spans V.

E.g. Following example above, $B = \{(1,0,0),(0,1,0)\}$ is a basis for W.

Dimension

The number of unique vectors in each basis of V is called the dimension of V, denoted by $\dim(V)$.

E.g. $\dim(W) = 2$ from example above.

Linear Transformations

Linear transformation

Let V and W be vector spaces. Then function $T : V \to W$ is a linear transformation from V to W if for all $\mathbf{x}, \mathbf{y} \in V$ and scalar c,

1. $T(\mathbf{x} + \mathbf{y}) = T(\mathbf{x}) + T(\mathbf{y})$ and
2. $T(c\mathbf{x}) = cT(\mathbf{x})$.

Turns out we can represent T with matrices, if we have the bases of V and W (won't show here). Thus we can consider a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ as a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$. Furthermore, matrix multiplication of multiple matrices, e.g. $\mathbf{AB}$, is associated to the composition of linear transformations represented by $\mathbf{A}$ and $\mathbf{B}$, respectively.

Null Space, Range & Rank

Nullspace

The nullspace of a matrix $A \in \mathbb{R}^{m\times n}$ is the vector space that consists of the all vectors $\mathbf{x} \in \mathbb{R}^n$ such that $\mathbf{Ax} = \mathbf{0}$. In the view of linear transformation, this is the subspace of the domain of the transformation which is mapped to the zero vector.

Column Space/Range

The column space of a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ is the span of its column vectors. In the view of linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$, this equals the *image* of the transformation.

Rank

The rank of $\mathbf{A} \in \mathbb{R}^{m\times n}$ is the dimension of its column space. It is equal to the number of linearly independent rows or columns.

Rank-Nullity Theorem

$\text{rank}(A) + \text{nullity}(A) = n$.

Full Rank

A matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ is full rank if $\mathrm{rank}(A) = \min(m, n)$.

Invertibility

A matrix must be *square* and *full rank* to be invertible. A non-invertible square matrix is called singular.

Symmetric Matrix

A symmetric matrix is a square matrix that is equal to its transpose, $\mathbf{A} = \mathbf{A}^T$.

Orthogonal Matrix

A orthogonal matrix is a square matrix whose columns and rows are orthonormal vectors, $\mathbf{A}^T\mathbf{A} = \mathbf{A}\mathbf{A}^T = \mathbf{I}$, or equivalently $\mathbf{A}^T = \mathbf{A}^{-1}$.

Eigenvalue and Eigenvectors

Eigenvectors

An eigenvector of a square matrix $\mathbf{A}$ is a nonzero vector $\mathbf{v}$ such that

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v},$$

where λ is a scalar called the eigenvalue corresponding to the eigenvector $\mathbf{v}$.

Eigendecomposition

Let $\mathbf{A} \in \mathbb{R}^{n \times n}$ with n linearly independent eigenvectors $\mathbf{q}_i$. Then $\mathbf{A}$ can be factorized as

$$\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{-1},$$

where $\mathbf{Q}$ is the square matrix whose i-th column is the eigen vector $\mathbf{q}_i$, and $\mathbf{\Lambda}$ is the diagonal matrix whose diagonals are the corresponding eigenvalues, $(\mathbf{\Lambda})_i i = \lambda_i$. If $\mathbf{Q}$ is an orthogonal matrix, it can be written as:

$$\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T,$$

Singular Value Decomposition

The SVD is a factorization that generalizes the eigendecomposition of a square $n \times n$ matrix to any $m \times n$ matrix.

Singular Value Decomposition

Let $\mathbf{A} \in \mathbb{R}^{m \times n}$. Then $\mathbf{A}$ can be factorized as

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T,$$

where $\mathbf{U} \in \mathbb{R}^{m \times m}, \mathbf{V} \in \mathbb{R}^{n \times n}$ are orthogonal matrices, and $\mathbf{\Sigma} \in \mathbb{R}^{m \times n}$ is a diagonal matrix.

Diagonals in $\mathbf{\Sigma}$ are known as singular values, and the columns in $\mathbf{U}$ and $\mathbf{V}$ are known as left- and right-singular vectors, respectively.

Table of contents

Probabilistic Experiments

Sample Space

The sample space is a set Ω of all possible outcomes. E.g. $\Omega = \{1, 2, 3, 4, 5, 6\}$ for a die.

Events

An event is a subset of Ω. E.g. $E = \{1, 3, 5\}$ is the event of an odd outcome.

Event Space

A collection of all events. E.g. $\mathcal{F} = \{\varnothing, \{1\}, \{2\}, \dots, \{1, 2\}, \dots, \{1, 2, 3, 4, 5, 6\}\}$ for rolling a die.

$\mathcal{F}$ must satisfy 3 properties:

- Contains empty event $\varnothing$ and trivial event Ω.
- Closed under complementation: If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$.
- Closed under union: If $A, B \in \mathcal{F}$, then $A \cup B \in \mathcal{F}$.

Probability Measure

Probability Measure

A probability measure is a function $\mathbb{P} : \mathcal{F} \to [0, 1]$, which assigns a nonnegative real number $\mathbb{P}(A)$ to every set A in $\mathcal{F}$. We call this the probability of the event A.

$(\Omega, \mathcal{F}, \mathbb{P})$ must satisfy three properties:

1. $\mathbb{P}(A) \geq 0 \; \forall \, A \in \mathcal{F}$,
2. $\mathbb{P}(\Omega) = 1$,
3. If $A, B \in \mathcal{F}$ and $A \cap B = \varnothing$ then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.

Conditional Probability,Chain Rule & Bayes' Rule

The probability of event B given event A:

$\mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}$, provided $\mathbb{P}(A) > 0$.

From definition above, we can see that $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B|A)$. More generally:

Chain rule of conditional probabilities

For events $A_1, \ldots, A_k$,
$\mathbb{P}(A_1 \cap \cdots \cap A_k) = \mathbb{P}(A_1)\mathbb{P}(A_2|A_1) \ldots \mathbb{P}(A_k|A_1 \cap \cdots \cap A_{k-1})$.

Also from definition we have $\mathbb{P}(B)\mathbb{P}(A|B) = \mathbb{P}(A)\mathbb{P}(B|A)$, which gives us Bayes' rule:

Bayes' rule

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}.$$

Independence

Independence

Two events A and B are independent if and only if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.

Alternatively, we have

$$\mathbb{P}(A) = \mathbb{P}(A|B) \text{ and } \mathbb{P}(B) = \mathbb{P}(B|A).$$

Conditional Independence

Events A and B are conditionally independent given event C if and only if $\mathbb{P}(A \cap B|C) = \mathbb{P}(A|C)\mathbb{P}(B|C)$.

Random Variables

Random Variable

A random variable is defined by a function $X : \Omega \to \mathbb{R}$ that associates with each outcome in Ω a real number.

Discrete Random Variable

Discrete random variables map to finite or countable values.

The event that X takes on a value x, denoted as $X = x$, refers to the event $\{\omega \in \Omega |\ X(\omega) = x\}$.

Probability Mass Function (PMF)

The probability mass function defined by $p_X(x) = \mathbb{P}(X = x)$ maps from a state of a r.v. X to the probability of the event that the X takes on that value x.

Joint & Marginals Distributions

PMFs can also act on multiple random variables *simultaneously*:

Joint PMF

$p_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \mathbb{P}(X_1 = x_1,\ldots,X_n = x_n).$

The set of r.v.s $\{X_1,\ldots,X_n\}$ is sometimes denoted as a *random vector* $\mathbf{x} = [X_1 \; \ldots \; X_n]$, with its joint PMF written as $p_{\mathbf{x}}(x)$.

A marginal distribution describes a distribution of a *subset* of a collection of r.v.s. Given the joint PMF, we can obtain the marginal PMFs by marginalizing, or summing out discarded variables.

Marginal PMF

Given a known joint distribution $p_{X,Y}(x,y)$ of discrete r.v.s X and Y,

$$p_X(x) = \sum_y p_{X,Y}(x,y) \; \forall\, x,$$

$$p_Y(y) = \sum_x p_{X,Y}(x,y) \; \forall\, y.$$

Conditional Probability & Independence

We can also extend notions of conditional probability and independence from events to r.v.s.

Conditional PMF

$$p_{Y|X=x}(y|x) = \mathbb{P}(Y = y|X = x) = \frac{\mathbb{P}(Y = y, X = x)}{\mathbb{P}(X = x)}.$$

Independence

Two r.v.s X, Y are independent if and only if $p_{X,Y}(x, y) = p_X(x)p_Y(y)$ *for any* $x, y \in \mathbb{R}$.

Conditional Independence

R.v.s X and Y are conditionally independent given r.v. Z if and only if $p_{X,Y|Z=z}(x, y|z) = p_{X|Z=z}(x|z)p_{Y|Z=z}(y|z)$ *for any* $x, y, z \in \mathbb{R}$.

Expectation

Expectation

The expectation (or expected value, mean) of a discrete r.v. X with PMF p_X is

$$\mathbb{E}[X] = \sum_x x p_X(x),$$

when the sum is well defined.

Properties of Expectation

- For X and $a, b \in \mathbb{R}$, $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$.
- For X and Y, $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$.
- If X and Y are independent, $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.
- $\mathbb{E}[g(X)] = \sum_x g(x) p_X(x)$.

Variance & Covariance

Variance

$\mathrm{var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$

Standard Deviation

$\sigma_X = \sqrt{\mathrm{var}(X)}.$

Covariance

$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$

Covariance Matrix

For a random vector $\mathbf{x} \in \mathbb{R}^n$, we can obtain an $n \times n$ covariance matrix $\mathrm{Cov}(\mathbf{x})$ such that

$$\left(\mathrm{Cov}(\mathbf{x})\right)_{i,j} = \mathrm{Cov}(x_i, x_j).$$

Note that the diagonal elements are the variances:

$$\left(\mathrm{Cov}(\mathbf{x})\right)_{i,i} = \mathrm{var}(x_i).$$

End