Linear Algebra Review
Evaluation Metrics
Project Overview

The toolkit for the next twelve weeks

CSCI 567 — Machine Learning • Discussion 01 • Aug 28, 2026

Outline

1. Linear algebra review: vectors, matrices, the tools Lecture 02 needs
2. Evaluation metrics: precision, recall, and why a single number can mislead
3. Project overview

1. Linear algebra review

Vectors, matrices, the toolkit for Lecture 02

What is linear algebra?

Linear algebra is the study of **vector spaces** and **linear functions**.

Vector space V

scaling + addition, satisfying

$1 \cdot v = v$, $\alpha(u + v) = \alpha u + \alpha v$, $(\alpha + \beta)v = \alpha v + \beta v$

Linear function

$T : \mathbb{R}^d \to \mathbb{R}^k$

$T(u + v) = T(u) + T(v)$,

$T(\alpha v) = \alpha T(v)$

$u, v \in \mathbb{R}^d$, $v = (v_1, \ldots, v_d)$; $u + v = (u_1 + v_1, \ldots, u_d + v_d)$, $\alpha v = (\alpha v_1, \ldots, \alpha v_d)$.

A linear function is determined by where it sends basis vectors

key idea: a linear function is determined by where it maps $(1, 0, \ldots, 0)$, $(0, 1, \ldots, 0)$, $\ldots$, $(0, \ldots, 0, 1)$

For instance:

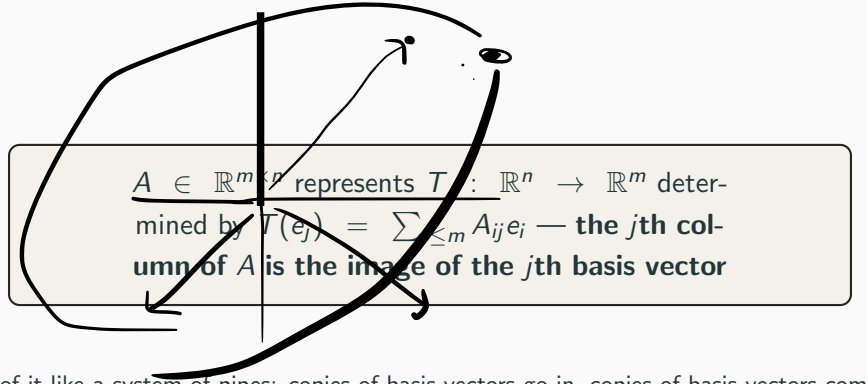
$$T((2,3)) = T((2,0)) + T((0,3)) = 2 \cdot T((1,0)) + 3 \cdot T((0,1))$$

Basic notation

$x \in \mathbb{R}^n$: a vector with n entries, $x = (x_1, \ldots, x_n)$.

e_i: the vector with a 1 in position i and 0 elsewhere, e.g. $e_2 = (0, 1, 0, \ldots, 0)$

Matrices: a concise way to write a linear function



$A \in \mathbb{R}^{m \times n}$ represents $T : \mathbb{R}^n \to \mathbb{R}^m$ determined by $T(e_j) = \sum_{i \leq m} A_{ij} e_i$ — **the jth column of A is the image of the jth basis vector**

Think of it like a system of pipes: copies of basis vectors go in, copies of basis vectors come out.

A 2×2 example

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

- Column 1 says: turn each copy of e_1 into a copies of e_1 and c copies of e_2
- Column 2 says: turn each copy of e_2 into b copies of e_1 and d copies of e_2

That's it — now you understand matrices.

Worth watching: 3Blue1Brown, "Essence of linear algebra" — the best visual intuition for what a matrix does, geometrically.

Matrix multiplication is function composition

AB represents $T_A \circ T_B$, whenever the composition is legal (dimension of B's output = dimension of A's input)

For $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{n \times p}$: $C = AB \in \mathbb{R}^{m \times p}$,

$$C_{i,j} := \sum_{k=1}^{n} A_{i,k} B_{k,j}.$$

Why matrix algebra looks so different from number algebra

It's composition of linear functions in disguise.

Not commutative: $AB \neq BA$ possible — $T_A \circ T_B$ can differ from $T_B \circ T_A$

Inverses may not exist: linear functions can destroy information, e.g. $T(x) = 0$

Not always defined: composition $f \circ g$ needs $\text{codomain}(g) = \text{domain}(f)$

Special matrices

Identity $I_n \in \mathbb{R}^{n \times n}$: 1s on the diagonal, 0 elsewhere. $AI_n = A = I_m A$

Diagonal $D = \operatorname{diag}(d_1, \ldots, d_n)$: d_i on the diagonal, 0 elsewhere

Vector-vector: inner (dot) product

$$x^\top y = \sum_{i=1}^{n} x_i y_i \in \mathbb{R}$$

geometric intuition: $x^\top y$ $=$
(length of x projected onto y) $\cdot$ (length of y)

Vector-vector: outer product

$$xy^\top \in \mathbb{R}^{m \times n}, \qquad (xy^\top)_{ij} = x_i y_j$$

geometric intuition: $xy^\top$ is the linear map that measures how much an input aligns with y, then outputs that amount in direction x

Applications: attention, covariance matrices, PCA.

Matrix-vector product: two views

By rows: $y = Ax$, $y_i = a_i^\top x$ — the set of inner products with each row. $a_i^\top x$ is how much of e_i gets produced by x

By columns: $y = Ax = \sum_j a_j x_j$ — a **linear combination of the columns** of A, weighted by x

Key corollary: Ax is restricted to the column space of A.

Vector-matrix product: two views

By rows: $y^\top = x^\top A = \sum_i x_i a_i^\top$
— a linear combination of A's **rows**

By columns: $y^\top = x^\top A = [x^\top a_1 \ \cdots \ x^\top a_n]$
— inner products with each column

$x^\top A$ combines rows of A; Ax combines columns of A — the same operation, read two ways.

Matrix-matrix multiplication: two more views

Inner products: $C = AB$, $C_{ij} = a_i^\top b_j$
— matrix of all row/column inner products

Matrix-vector products: $C = AB = [Ab_1 \cdots Ab_n]$
— b_i is B's output on e_i, so $Ab_i = (AB)e_i$

Matrix multiplication: properties

Associative: $(AB)C = A(BC)$

Distributive: $A(B + C) = AB + AC$

In general **not** commutative: $AB \neq BA$

Transpose

$$A_{ij} \longleftrightarrow (A^\top)_{ji}$$

$$(A^\top)^\top = A, \quad (AB)^\top = B^\top A^\top, \quad (A+B)^\top = A^\top + B^\top$$

$$A = A^\top \Rightarrow \textbf{symmetric}.$$
$$A = -A^\top \Rightarrow \textbf{anti-symmetric}$$

Exercise

$x_1, \ldots, x_N \in \mathbb{R}^D$, $X \in \mathbb{R}^{N \times D}$ with nth row $x_n^\top$. Which are correct?

A. $X^\top X = \sum_{n=1}^N x_n x_n^\top$
B. $X^\top X = \sum_{n=1}^N x_n^\top x_n$
C. $X X^\top = \sum_{n=1}^N x_n x_n^\top$
D. $X X^\top = \sum_{n=1}^N x_n^\top x_n$

(A is correct: $X^\top X \in \mathbb{R}^{D \times D}$ matches $\sum_n x_n x_n^\top \in \mathbb{R}^{D \times D}$.)

Trace

$$\operatorname{tr}(A) = \sum_{i=1}^{n} A_{ii}$$

$$\operatorname{tr}(A) = \operatorname{tr}(A^\top), \quad \operatorname{tr}(A + B) = \operatorname{tr}(A) + \operatorname{tr}(B), \quad \operatorname{tr}(tA) = t \operatorname{tr}(A)$$

Cyclic property: $\operatorname{tr}(ABC) = \operatorname{tr}(BCA) = \operatorname{tr}(CAB)$

Norms: what makes something a valid “length”

A function $f : \mathbb{R}^n \to \mathbb{R}$ is a norm if, for all $x, y \in \mathbb{R}^n$:

Non-negativity: $f(x) \geq 0$

Definiteness: $f(x) = 0 \iff x = 0$

Homogeneity: $f(tx) = |t| f(x)$

Triangle inequality: $f(x + y) \leq f(x) + f(y)$

Examples of norms

$$\ell_2 \text{ (Euclidean): } \|x\|_2 = \sqrt{\sum_i x_i^2} = \sqrt{x^\top x}$$

$$\ell_1\text{: } \|x\|_1 = \sum_i |x_i|$$

$$\ell_\infty\text{: } \|x\|_\infty = \max_i |x_i|$$

ℓ_2 and ℓ_1 return in Lecture 03 as ridge and lasso regularizers.

The ℓ_p family, and the Frobenius norm

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}, \qquad p \geq 1$$

Matrix norm (Frobenius): $\|A\|_F = \sqrt{\sum_{i,j} A_{ij}^2} = \sqrt{\mathrm{tr}(A^\top A)}$

Span, linear independence, rank

Span$(\{x_1, \ldots, x_n\}) = \{v : v = \sum_i \alpha_i x_i,\ \alpha_i \in \mathbb{R}\}$ — column span = column space

Linearly dependent: some x_n is a combination of the rest, $x_n = \sum_{i<n} \alpha_i x_i$

Rank: largest number of linearly independent columns (= rows); $\mathrm{rank}(A) \leq \min(m, n)$

Rank and inverse: properties

$\mathrm{rank}(A) = \mathrm{rank}(A^\top)$; for $A \in \mathbb{R}^{m \times p}, B \in \mathbb{R}^{p \times n}$: $\mathrm{rank}(AB) \leq \min(\mathrm{rank}A, \mathrm{rank}B)$

A^{-1} is the unique matrix with $A^{-1}A = I_n = AA^{-1}$; exists iff A is full rank ($\mathrm{rank}(A) = n$)

Determinant: geometric intuition

$|A|$ is the (signed) **volume** of the parallelepiped spanned by A's columns, $S = \{v : v = \sum_i \alpha_i a_i,\ 0 \leq \alpha_i \leq 1\}$

2×2: $\left|\begin{smallmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{smallmatrix}\right| = a_{11}a_{22} - a_{12}a_{21}$.

Determinant: the recursive formula

Let $A_{\setminus i,\setminus j}$ be A with row i and column j deleted (cofactor expansion along any row i, or any column j):

$$|A| = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} \, |A_{\setminus i,\setminus j}| \quad (\forall i)$$

$$|A| = \sum_{i=1}^{n} (-1)^{i+j} a_{ij} \, |A_{\setminus i,\setminus j}| \quad (\forall j)$$

Base case: $|a_{11}| = a_{11}$ for a 1×1 matrix.

Determinant: properties

$$|A| = |A^\top|, \quad |AB| = |A||B|$$

$$|A| = 0 \iff A \text{ is singular}$$

non-singular: $|A^{-1}| = 1/|A|$

Exercises

Which identities are **not** correct (inverses exist, multiplications legal)?

A. $(AB)^{-1} = B^{-1}A^{-1}$
B. $(I + A)^{-1} = I - A$
C. $\mathrm{tr}(AB) = \mathrm{tr}(BA)$
D. $(AB)^{\top} = A^{\top}B^{\top}$

Also: for $x \in \mathbb{R}^n$, what is $\mathrm{rank}(xx^{\top})$?

(B and D are false in general; $\mathrm{rank}(xx^{\top}) = 1$ *for* $x \neq 0$*.)*

Matrix calculus

The tools Lecture 02 needs directly

The gradient

For $f : \mathbb{R}^n \to \mathbb{R}$ (scalar output, vector input):

$$\nabla_x f(x) = \begin{pmatrix} \partial f/\partial x_1 \\ \vdots \\ \partial f/\partial x_n \end{pmatrix}$$

Linearity carries over: $\nabla_x(f+g) = \nabla_x f + \nabla_x g, \;\; \nabla_x(tf) = t\nabla_x f.$

Gradient, Hessian, Jacobian

Hessian ($f : \mathbb{R}^n \to \mathbb{R}$): $\nabla_x^2 f(x) \in \mathbb{R}^{n \times n}$, entries $\partial^2 f / \partial x_i \partial x_j$ — symmetric

Jacobian ($f : \mathbb{R}^n \to \mathbb{R}^m$, vector output): $\nabla_x f(x) \in \mathbb{R}^{m \times n}$, row i is $\nabla_x f_i(x)^\top$

Gradient of a linear function, derived

$f(x) = b^\top x = \sum_i b_i x_i$ for fixed $b \in \mathbb{R}^n$. Then

$$\frac{\partial f(x)}{\partial x_k} = \frac{\partial}{\partial x_k} \sum_i b_i x_i = b_k.$$

$$\nabla_x \, b^\top x = b$$

Same shape as single-variable calculus: $\frac{\partial (ax)}{\partial x} = a$.

Jacobian of a linear function, derived

$f(x) = Ax$ for fixed $A \in \mathbb{R}^{m \times n}$, so $f_i(x) = a_i^\top x$.

$$\nabla_x f_i(x) = a_i \quad \Longrightarrow \quad \nabla_x f(x) = \begin{pmatrix} a_1^\top \\ \vdots \\ a_m^\top \end{pmatrix} = A$$

Gradient of a quadratic function, step by step

$f(x) = x^\top Ax = \sum_{i,j} A_{ij} x_i x_j$ for fixed $A \in \mathbb{R}^{n \times n}$.

Product rule on $f(x) = g(x)^\top x$ with $g(x) = A^\top x$:

$$\nabla_x f(x) = \nabla_x g(x)^\top x + \nabla_x x^\top g(x) = (A^\top)^\top x + I\, A^\top x = (A + A^\top)x$$

$$\nabla_x\, x^\top Ax = (A + A^\top)x$$

Hessian: $\nabla_x^2 f(x) = A + A^\top$ (symmetric A: simplifies to $2Ax$ and $2A$).

Exercises

$f(x) = x^\top Ax + b^\top x$ for $A \in \mathbb{R}^{n \times n}$, $b \in \mathbb{R}^n$. What is $\nabla_x f(x)$?

(Sum rule on the two pieces already derived: $(A + A^\top)x + b$.)

$f(A) = x^\top Ax$ for fixed $x \in \mathbb{R}^n$. What is $\partial f / \partial A$?

($f(A) = \sum_{i,j} x_i x_j A_{ij}$, so $\partial f / \partial A_{ij} = x_i x_j$: $\nabla_A f(A) = xx^\top$.)

Why this toolkit, right now

Lecture 02 sets $\nabla_{\mathbf{w}} \hat{\mathcal{R}}(\mathbf{w}) = 0$ to solve linear regression in closed form — exactly the gradient-of-quadratic-form rule just derived

every gradient descent algorithm this course builds computes $\nabla g(\mathbf{w})$ and moves against it: $\mathbf{w} \leftarrow \mathbf{w} - \eta \nabla g(\mathbf{w})$

$X^\top X$'s positive semi-definiteness (next week) and PCA's eigendecomposition (Lecture 06) both build directly on today

References

- Content this section: Vatsal Sharan, CSCI 567 Discussion 1 – Linear Algebra Review I (Spring 2026), used with permission.
- Vatsal Sharan's own slides are adapted from Nandita Bhaskhar's discussion slides for CS229 at Stanford.

2. Evaluation metrics

Why a single number can mislead

Same theme, one level down

Lecture 01: a benchmark score can look like strong performance while hiding what actually matters

Today: the exact same failure, at the level of a single metric.

Setup: a rare-disease classifier

1,000 people

1% actually
have the disease
(10 people)

A classifier that always predicts "no disease" scores **99% accuracy**.

The trap

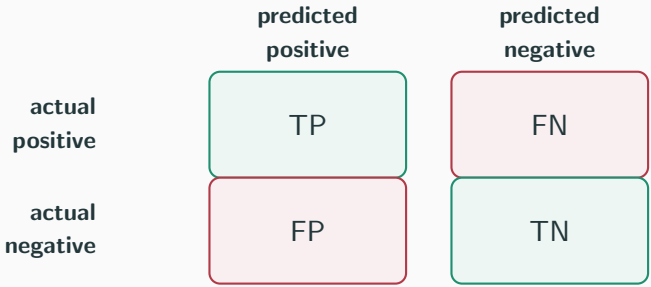
99%

accuracy

a classifier that never once looks at the input

Useless, and accuracy alone cannot tell you that.

The confusion matrix



Every metric on the next few slides is a ratio built from these four counts.

Precision: of what I flagged, how much was real?

$$\text{Precision} = \frac{TP}{TP + FP}$$

The cost of a **false positive**. Example: a spam filter that blocks real mail.

Recall: of what was real, how much did I catch?

$$\text{Recall} = \frac{TP}{TP + FN}$$

The cost of a **false negative**. Example: missing an actual disease case.

The disease example, worked

Always-negative classifier: $TP = 0$, $FN = 10$, $FP = 0$, $TN = 990$.

Accuracy $= \quad 990/1000 \quad = \quad \mathbf{99\%}$

Recall $= \quad 0/10 \quad = \quad \mathbf{0\%}$ — catches *none* of the real cases

Precision $= \quad 0/0$ — **undefined**, never once predicts positive

F1: one number, if you insist

$$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

The **harmonic** mean of precision and recall — not the arithmetic mean. Why does that choice matter?

Why harmonic, not arithmetic

Precision $= 1.0$, Recall $= 0.01$ (catches almost nothing):

arithmetic mean $= \frac{1.0+0.01}{2} = 0.505$
— **looks like half-decent performance**

$F_1 = \frac{2(1.0)(0.01)}{1.0+0.01} \approx 0.0198$ — **correctly reflects the near-total recall failure**

Harmonic mean is dominated by the *smaller* of the two numbers; arithmetic mean is not. Report the pair, not just F_1, when the two disagree this much.

These are estimates too

every metric here is computed on a **finite test set**
— Lecture 01's concentration bound applies directly

Accuracy is literally $\frac{1}{n}\sum_i Z_i$ for $Z_i = \mathbb{1}[\hat{y}_i = y_i] \in \{0, 1\}$ — exactly Hoeffding's setup.

A concrete confidence interval on accuracy

Same bound as Lecture 01: $\mathbb{P}[\,|\bar{Z} - \mu| \geq \epsilon\,] \leq 2\exp(-2n\epsilon^2)$.

$n = 1000$, 95% confidence:

$$\epsilon = \sqrt{\frac{\ln(2/0.05)}{2 \cdot 1000}} \approx 4.3\%$$

A reported accuracy of "91%" on 1,000 examples really means $91\% \pm 4.3$ **points**.

Recall's hidden trap: the effective n is tiny

Recall's denominator is $TP + FN$ — the number of *actual positives*, not the whole test set.

on our 10-positive disease set: $n = 10$, not 1000

$$\epsilon = \sqrt{\frac{\ln(2/0.05)}{2 \cdot 10}} \approx \mathbf{43\%}$$

A recall number on a rare class, from a small test set, is barely more than a guess — report the raw count of positives it was computed from.

Sweeping a threshold: ROC

A classifier outputs a score; thresholding it gives a label. Sweep the threshold, trace a curve.

TPR (recall)
$TP/(TP + FN)$

FPR
$FP/(FP + TN)$

A tiny worked example

example	a	b	c	d	e	f
true label	+	+	−	+	−	−
score	0.9	0.75	0.6	0.5	0.4	0.2

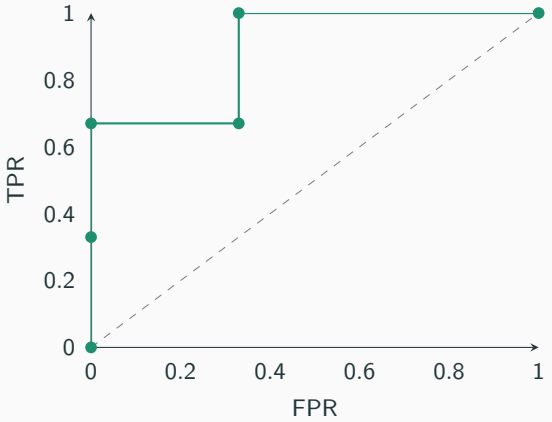
3 positives (a,b,d), 3 negatives (c,e,f). Sweep the threshold from high to low and recompute TPR/FPR at each point.

Sweeping the threshold, step by step

threshold τ	predicted +	*TP*	*FP*	TPR	FPR
> 0.9	{}	0	0	0.00	0.00
> 0.75	{a}	1	0	0.33	0.00
> 0.6	{a,b}	2	0	0.67	0.00
> 0.5	{a,b,c}	2	1	0.67	0.33
> 0.4	{a,b,c,d}	3	1	1.00	0.33
≥ 0.2	{a,b,c,d,e,f}	3	3	1.00	1.00

Every row: recompute TP/FP directly from the confusion matrix at that threshold.

The resulting ROC curve



The dashed diagonal is random guessing; our curve sits above it — better than chance at every threshold shown.

AUC: one number, precisely interpreted

area under the ROC curve = the probability a **randomly chosen positive** scores higher than a **randomly chosen negative**

AUC = 0.5: no better than a coin flip. AUC = 1: perfect separation.

When ROC is misleading: precision-recall curves

when negatives vastly outnumber positives, TN is huge,
so FPR can look tiny even while precision is terrible

On the rare-disease setting: prefer the **precision-recall curve** over ROC.

Multiclass: macro vs. micro averaging

macro-average: compute the metric per class, then average the K numbers — every class weighted equally

micro-average: pool $TP/FP/FN$ across all classes first, then compute one ratio — large classes dominate

Macro is the one that actually penalizes doing badly on a rare class.

Recap

report precision *and* recall, not accuracy alone — and know which false-error type you actually care about

every metric is an estimate with sampling uncertainty — ask what n it was actually computed from

sweep thresholds (ROC / PR) rather than committing to a single operating point too early

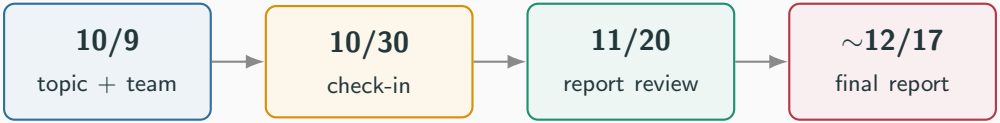
3. Project overview

What's due, and when

groups of **4**

% **20%** of course grade

Timeline



Topic and team are due at the end of Fall Recess — **six weeks from today**.

Why so early

picking a topic and a team early gives you the most runway — most of what you'll need for the project (models, evaluation, generalization) is built up across the whole semester

Start thinking about a direction now, even loosely.