02 · Math for ML

Linear algebra

Idea	Formula	NumPy
dot product	$a \cdot b = \Sigma a_i b_i = \|a\|\|b\|\cos\theta$	`a @ b`
linear model	$\hat{y} = x \cdot w + b$	`X @ w + b`
matrix product	(n,d) @ (d,k) → (n,k)	`A @ B`
transpose	(n,d) → (d,n)	`X.T`
length (L2 norm)	$\|x\| = \sqrt{\Sigma x_i^2}$	`np.linalg.norm(x)`
distance	$\|a - b\|$	`np.linalg.norm(a - b)`
cosine similarity	$a \cdot b / (\|a\|\|b\|)$	`a @ b / (norm(a) * norm(b))`

`*` is element-wise; `@` is the matrix product.

Calculus

f(x)	f′(x)
c	0
x^n	$n \cdot x^{n-1}$
e^x	e^x
ln x	1/x
$\sigma(x) = 1/(1+e^{-x})$	$\sigma(x)(1 - \sigma(x))$
ReLU max(0,x)	1 if x > 0 else 0

- **Chain rule:** d/dx f(g(x)) = f′(g(x))·g′(x) — backprop multiplies local slopes backwards.
- **Gradient descent:** $w \leftarrow w - \eta \cdot \nabla L(w)$. Repeat.
- **Linear regression, MSE:** $\nabla w = (2/n) \cdot X^T(\hat{y} - y)$, $\partial b = (2/n) \cdot \Sigma(\hat{y} - y)$.
- **Numerical check:** $(L(w+\varepsilon) - L(w-\varepsilon)) / 2\varepsilon \approx \partial L/\partial w$ with $\varepsilon \approx$ 1e-6.

Learning rate	Symptom
too small	loss falls very slowly
good	loss falls fast, then flattens
too large	loss oscillates or becomes inf / NaN

Probability

Idea	Formula
conditional	P(A|B) = P(A and B) / P(B)
Bayes	P(A|B) = P(B|A)·P(A) / P(B)
expected value	$E[X] = \Sigma x \cdot P(x)$
sigmoid	$p = 1 / (1 + e^{-z})$
softmax	p_i = e^{z_i} / Σ_j e^{z_j} (subtract max z first for stability)
cross-entropy	−log p(true class)

Distribution	Use	NumPy
Bernoulli / binomial	yes/no, counts	`rng.binomial(n, p)`
categorical	one of k	`rng.choice(k, p=probs)`
normal	noise, sums	`rng.normal(mu, sigma)`
uniform	random search	`rng.uniform(a, b)`
Poisson	rare counts	`rng.poisson(lam)`

Normal: 68% within 1σ, 95% within 2σ, 99.7% within 3σ.

Statistics

Idea	NumPy
mean / median	`x.mean()` · `np.median(x)`
std (pandas uses ddof=1)	`x.std(ddof=0)`
z-score	`(x - x.mean()) / x.std()`
correlation	`np.corrcoef(a, b)[0, 1]`
bootstrap CI	resample with replacement 2,000×, take 2.5th and 97.5th percentiles

- Standard error of a mean ≈ σ/√n: 100× more data → 10× less wobble.
- Correlation ≠ causation; only an experiment (A/B test) shows cause.