CS229 - Machine Learning
Linear Algebra & Probability

Tim Reinhart - rtim@stanford.edu

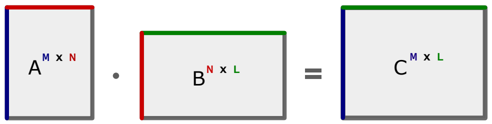
Version: October 31, 2023

Matrices

Matrix Multiplication

Matrices can be multiplied with each other in the following manner:

$$\boxed{A \cdot B = C \quad \Rightarrow \quad c_{ik} = \sum_{j=1}^{n} a_{ij} \cdot b_{jk}}$$



Associative & Distributive Laws:

$$(A \cdot B) \cdot C = A \cdot (B \cdot C)$$
$$(A + B) \cdot C = A \cdot C + B \cdot C$$
$$A \cdot (C + D) = A \cdot C + A \cdot D$$

Warning! The commutative law does not apply! Generally, $A \cdot B \neq B \cdot A$.

Transpose

The transpose of a matrix is obtained by "mirroring" it along its diagonal.

Example: $\begin{pmatrix} a & b \\ c & d \\ e & f \end{pmatrix}^T = \begin{pmatrix} a & c & e \\ b & d & f \end{pmatrix}$

Calculation Rules:

$$(A + B)^T = A^T + B^T \qquad (A^T)^{-1} = (A^{-1})^T$$
$$(A \cdot B)^T = B^T \cdot A^T \qquad rank(A^T) = rank(A)$$
$$(c \cdot A)^T = c \cdot A^T \qquad det(A^T) = det(A)$$
$$(A^T)^T = A \qquad eig(A^T) = eig(A)$$

Inverse

The inverse A^{-1} of A reverses a multiplication with A. When you multiply A with A^{-1}, you get the identity matrix.

Properties:

- Only square matrices can be invertible.
- An invertible matrix is called **regular**, a non-invertible one **singular**.
- The inverse is unique.
- A is invertible if and only if A has full rank.
- A is invertible if and only if A^T is invertible.
- A is symmetric if and only if A^{-1} is symmetric.
- A is a triangular matrix if and only if A^{-1} is a triangular matrix.
- A is invertible if and only if $\det(A) \neq 0$.
- A is invertible if and only if no eigenvalue $\lambda = 0$.
- A and B are invertible implies AB is invertible.

Calculation rules:

$$I^{-1} = I \qquad (A^T)^{-1} = (A^{-1})^T$$
$$(A^{-1})^{-1} = A \qquad rang(A^{-1}) = rang(A)$$
$$(A^k)^{-1} = (A^{-1})^k \qquad det(A^{-1}) = det(A)^{-1}$$
$$(c \cdot A)^{-1} = c^{-1} \cdot A^{-1} \qquad eig(A^{-1}) = eig(A)^{-1}$$
$$(A \cdot B)^{-1} = B^{-1} \cdot A^{-1}$$

Matrix Tricks

Probability Rules for Matrices:

- Pull Matrix Multiply out of Variance:

$$Var[Mx] = MVar[x]M^T$$

Eigenvalues and Eigenvectors

$$\boxed{\text{Eigenvalues of A: } \det(A - \lambda \cdot I) = 0}$$

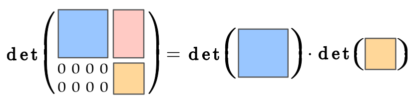
Verify Computation

- Trace(A) $= a_{11} + a_{22} + \cdots + a_{nn} = \sum \lambda_i$
- det(A) = product of λ_i

Eigenvectors: Kernel of the matrix $A - \lambda_i \cdot I$, where λ_i is the eigenvalue corresponding to the eigenvector.

Determinant

Block Sentence for Determinant Computation



Positive (Semi-)Definite Matrices

Definitions

A symmetric matrix $A \in \mathbb{R}^{n \times n}$ is called:

- **Positive Semi-Definite (PSD)** if for any non-zero vector $\mathbf{x} \in \mathbb{R}^n$, we have $\mathbf{x}^T A \mathbf{x} \geqslant 0$.
- **Positive Definite (PD)** if for any non-zero vector $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{x}^T A \mathbf{x} > 0$.

Properties

- All eigenvalues of a PSD matrix are non-negative, and those of a PD matrix are positive.
- A matrix is PSD if and only if it can be written as $B^T B$, where B is any matrix.
- If A is PD (or PSD), then so is A^{-1} (if A is invertible).
- For any matrix A, the matrices $A^T A$ and AA^T are PSD.
- The sum of two PSD matrices is also PSD.

Checking for Positive (Semi-) Definiteness

Determining if a matrix is PSD or PD can be done in several ways:

- **Eigenvalue Criterion:** A symmetric matrix is PSD if and only if all its eigenvalues are non-negative. It is PD if all eigenvalues are positive.
- **Principal Minors:** A symmetric matrix A is PD if all its leading principal minors (determinants of the top-left $k \times k$ submatrix, $1 \leqslant k \leqslant n$) are positive. For PSD, all leading principal minors should be non-negative.
- **Cholesky Decomposition:** A matrix is PD if and only if it has a Cholesky decomposition. For numerical algorithms, attempting a Cholesky decomposition and checking for failure can be an effective way to test for positive definiteness.

Matrix Calculus

Gradient

The gradient of a scalar function $f : \mathbb{R}^n \to \mathbb{R}$ with respect to a vector $\mathbf{x} \in \mathbb{R}^n$ is a vector of partial derivatives:

$$\nabla f(\mathbf{x}) = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix}$$

Hessian

The Hessian matrix of a scalar-valued function $f : \mathbb{R}^n \to \mathbb{R}$ is a square matrix of second-order partial derivatives:

$$H(f)(\mathbf{x}) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix}$$

Examples

$f(\mathbf{x}) = \mathbf{A}\mathbf{x}$ $\qquad \mathbf{A} \in \mathbb{R}^{m \times n}$

Gradient:

$$\nabla f = \mathbf{A}^T$$

$f(\mathbf{x}) = \mathbf{x}^T \mathbf{A}\mathbf{x}$ $\qquad \mathbf{A} \in \mathbb{R}^{m \times n}$

Gradient:

$$\nabla f = (\mathbf{A} + \mathbf{A}^T)\mathbf{x}$$

Hessian:

$$H(f) = \mathbf{A} + \mathbf{A}^T$$

Linear Regression Loss (ℓ_2 norm)

For the loss function $L(\mathbf{w}) = ||\mathbf{y} - \mathbf{X}\mathbf{w}||^2$:
Gradient:

$$\nabla L = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\mathbf{w})$$

Hessian:

$$H(L) = 2\mathbf{X}^T\mathbf{X}$$

Logistic Regression Loss

- Binary classification with labels $y_i \in \{0, 1\}$
- Predicted probabilities $p_i = \frac{1}{1+e^{-\mathbf{x}_i^T \mathbf{w}}}$
- $L(\mathbf{w}) = -\sum_i [y_i \log(p_i) + (1 - y_i)\log(1 - p_i)]$

Gradient:

$$\nabla L = \mathbf{X}^T(\mathbf{p} - \mathbf{y})$$

Hessian:

$$H(L) = \mathbf{X}^T\mathbf{S}\mathbf{X}$$

where $\mathbf{S}$ is a diagonal matrix with $S_{ii} = p_i(1 - p_i)$.

Basic Probability

Bayes Theorem

$$P(X{=}x|Y{=}y) = \frac{P(Y{=}y|X{=}x)P(X{=}x)}{P(Y{=}y)}$$

Where:

- $P(X{=}x|Y{=}y)$ is the posterior probability: the probability of event $X{=}x$ given that $Y{=}y$ has occurred.
- $P(Y{=}y|X{=}x)$ is the likelihood: the probability of observing $Y{=}y$ given $X{=}x$.

- $P(X{=}x)$ is the prior probability: the initial belief about $X{=}x$.
- $P(Y{=}y)$ is the marginal probability: the total probability of observing $Y{=}y$ under all possible outcomes of X.

Law of Total Probability

A key concept related to Bayes' Theorem is the Law of Total Probability. It is useful for calculating $P(Y{=}y)$, the marginal probability in Bayes' formula, especially when dealing with compound events. The law states:

$$P(Y{=}y) = \sum_i P(Y{=}y|X{=}x_i)P(X{=}x_i)$$

Where $X{=}x_i$ represents all disjoint outcomes that cover the sample space. In the context of Bayes' Theorem, it's used to marginalize over the different possible states of knowledge or evidence.

Bayes' Rule for Multiple Events

In cases involving more than two events, Bayes' Theorem can be generalized as:

$$P(X_1{=}x_1, \dots, X_n{=}x_n|Y{=}y) = \frac{P(Y{=}y|X_1{=}x_1, \dots, X_n{=}x_n)\prod_{i=1}^n P(X_n{=}x_n)}{P(Y{=}y)}$$

Bayes' Theorem with Continuous Variables

When dealing with continuous variables, Bayes' Theorem takes the form of probability densities:

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)f_X(x)}{f_Y(y)}$$

Where $f_{X|Y}(x|y)$ is the conditional density of X given Y, and so on.

Prior and Posterior Probabilities

In Bayesian analysis, the prior probability $P(X{=}x)$ represents our belief about X before observing the evidence Y, while the posterior probability $P(X{=}x|Y{=}y)$ is our updated belief after observing Y. The transformation from the prior to the posterior, via the likelihood and marginal likelihood, is the essence of Bayesian inference.

Expectation Value

$$\mathrm{E}[X] = \sum_i x_i p_i \quad \text{(for discrete var.)} \quad \text{or}$$

$$\mathrm{E}[X] \equiv \int_\Omega X \, d\,\mathrm{P} = \int_\mathbb{R} x f(x)\, dx \quad \text{(for cont. var.)}$$

Properties of Expectation

Linearity The expectation operator is linear:

$$\mathrm{E}[aX + bY] = a\,\mathrm{E}[X] + b\,\mathrm{E}[Y]$$

where a and b are constants, and X and Y are random variables.

Monotonicity If $X \leqslant Y$ (i.e., X is always less than or equal to Y), then:

$$\mathrm{E}[X] \leqslant \mathrm{E}[Y]$$

Law of the Unconscious Statistician This law states that if $Y = g(X)$ for some function g, then:

$$\mathrm{E}[Y] = \mathrm{E}[g(X)] = \sum_x g(x)P(X = x) \quad \text{(discrete case)}$$

or

$$\mathrm{E}[g(X)] = \int_{-\infty}^{+\infty} g(x) f_X(x)\, dx \quad \text{(continuous case)}$$

where $f_X(x)$ is the probability density function of X.

Independence If two random variables X and Y are independent, then:

$$\mathrm{E}[XY] = \mathrm{E}[X] \cdot \mathrm{E}[Y]$$

Conditional Expectation

$$\mathrm{E}(X|Y = y) = \sum_x xP(X = x|Y = y) = \sum_x x \frac{P(X = x, Y = y)}{P(Y = y)}$$

Variance

Variance quantifies the spread or dispersion of a set of data points or the spread of a probability distribution. It is defined as the expected value of the squared deviation from the mean (denoted by μ):

$$Var(X) = \mathrm{E}[(X - \mu)^2] = \mathrm{E}[X^2] - (\mathrm{E}[X])^2$$

Properties of Variance

Non-negativity The variance is always non-negative:

$$Var(X) \geqslant 0$$

Variance of a Constant

$$Var(a) = 0$$

where $a \in \mathbb{R}$ is a constant.

Factor Out Constants

$$Var(aX) = a^2 Var(X)$$

where $a \in \mathbb{R}$ is a constant.

Variance of a Sum For any random variables X, Y:

$$\begin{aligned}\mathrm{Var}(aX + bY) &= a^2\,\mathrm{Var}(X) + b^2\,\mathrm{Var}(Y) \\ &\quad + 2ab\,\mathrm{Cov}(X, Y) \\ \mathrm{Var}(aX - bY) &= a^2\,\mathrm{Var}(X) + b^2\,\mathrm{Var}(Y) \\ &\quad - 2ab\,\mathrm{Cov}(X, Y)\end{aligned}$$

If X and Y are independent, then $Cov(X, Y) = 0$, and this simplifies to:

$$\mathrm{Var}(aX \pm bY) = a^2\,\mathrm{Var}(X) + b^2\,\mathrm{Var}(Y)$$

Sum of uncorrelated variables

$$\mathrm{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \mathrm{Var}(X_i)$$

Sum of correlated variables

$$\begin{aligned}\mathrm{Var}\left(\sum_{i=1}^n X_i\right) &= \sum_{i=1}^n \sum_{j=1}^n \mathrm{Cov}\left(X_i, X_j\right) \\ &= \sum_{i=1}^n \mathrm{Var}\left(X_i\right) + 2\sum_{1 \leqslant i<j \leqslant n} \mathrm{Cov}\left(X_i, X_j\right)\end{aligned}$$

Exponential Family

A single-parameter exponential family is a set of probability distributions whose probability density function (or probability mass function, for the case of a discrete distribution) can be expressed in the form

$$p(y; \eta) = b(\eta)\, \exp\left[\, \eta^T\, T(y) - a(\eta)\, \right]$$

- η: natural parameter
- $T(y)$: sufficient statistic
- $a(\eta)$: log partition function

Canonical Response Funtion

$$g(\eta) = \mathrm{E}[T(y); \eta]$$

- For the Gaussian family: identify function
- For the Bernoulli family: logistic function

Probability Distributions

Discrete Distributions

Bernoulli Distribution

PMF:

$$P(X = x) = p^x (1-p)^{1-x} \quad \text{for } x \in \{0, 1\}$$

Mean and Variance:

$$\mu = p, \quad \sigma^2 = p(1 - p)$$

Binomial Distribution

PMF:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Mean and Variance:

$$\mu = np, \quad \sigma^2 = np(1 - p)$$

Poisson Distribution

PMF:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Mean and Variance:

$$\mu = \sigma^2 = \lambda$$

Geometric Distribution

PMF:

$$P(X = k) = (1-p)^{k-1}p$$

Mean and Variance:

$$\mu = \frac{1}{p}, \quad \sigma^2 = \frac{1-p}{p^2}$$

Continuous Distributions

Exponential Distribution

PDF:

$$f(x) = \lambda e^{-\lambda x}, \quad x \geqslant 0$$

Mean and Variance:

$$\mu = \frac{1}{\lambda}, \quad \sigma^2 = \frac{1}{\lambda^2}$$

Uniform Distribution

PDF:

$$f(x) = \frac{1}{b-a} \quad \text{for } a \leqslant x \leqslant b$$

Mean and Variance:

$$\mu = \frac{a+b}{2}, \quad \sigma^2 = \frac{(b-a)^2}{12}$$

Beta Distribution

PDF:

$$f(x; \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha, \beta)}$$

Mean and Variance:

$$\mu = \frac{\alpha}{\alpha+\beta}, \quad \sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Gaussian Distributions

Univariate Gaussian

The probability density of a univariate Gaussian distribution is given by:

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Where μ is the mean and σ^2 is the variance.

Multivariate Gaussian

The probability density of a multivariate Gaussian distribution is:

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\Sigma) = \frac{1}{\sqrt{(2\pi)^k|\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

Where $\boldsymbol{\mu}$ is the mean vector, Σ is the covariance matrix, and k is the number of dimensions.

Mean Vector and Covariance Matrix

- The mean vector $\boldsymbol{\mu}$ represents the mean of each dimension. If $\mathbf{x}$ is an n-dimensional random vector, then $\boldsymbol{\mu}$ is given by $\boldsymbol{\mu} = \mathrm{E}[\mathbf{x}]$.
- The covariance matrix Σ represents how each pair of dimensions of the random vector $\mathbf{x}$ co-varies. If $\mathbf{x}$ has dimensions $x_1, x_2, ..., x_n$, then the element Σ_{ij} of the matrix Σ is the covariance between x_i and x_j: $\Sigma_{ij} = \mathrm{Cov}(x_i, x_j)$.
- The determinant of Σ (denoted as $|\Sigma|$) and its inverse Σ^{-1} play a key role in defining the shape and orientation of the multivariate Gaussian distribution in its multi-dimensional space.

Multinomial Distribution

The multinomial distribution is a generalization of the binomial distribution. It models the probabilities of the various outcomes of a categorical variable over n trials.