

18-661 Introduction to Machine Learning

Review of Mathematics for ML

Fall 2018

ECE – Carnegie Mellon University

Outline

1. Linear Algebra
2. Calculus and Optimization
3. Probability
4. Review on Statistics

Linear Algebra

Linear Algebra

Calculus and Optimization

Probability

Review on Statistics

Vector spaces – definition

Vector Space $(V, +, \cdot)$ over a field \mathbb{F}

Set of elements (vectors) with two operations:

- sum of elements: $\mathbf{u} + \mathbf{v}$, where $\mathbf{u}, \mathbf{v} \in V$
- and multiplication by a scalar: $\alpha \cdot \mathbf{u}$, $\alpha \in \mathbb{F}$ ($\mathbb{F} = \mathbb{R}, \mathbb{C}, \dots$).

Vector spaces – definition

Vector Space $(V, +, \cdot)$ over a field \mathbb{F}

Set of elements (vectors) with two operations:

- sum of elements: $\mathbf{u} + \mathbf{v}$, where $\mathbf{u}, \mathbf{v} \in V$
- and multiplication by a scalar: $\alpha \cdot \mathbf{u}$, $\alpha \in \mathbb{F}$ ($\mathbb{F} = \mathbb{R}, \mathbb{C}, \dots$).

Satisfying:

1. $\exists \mathbf{0} \in V : \mathbf{x} + \mathbf{0} = \mathbf{x}$,
2. $\forall \mathbf{x} \in V : \exists -\mathbf{x} : \mathbf{x} + (-\mathbf{x}) = \mathbf{0}$,
3. $\exists \zeta \in \mathbb{F} : \zeta \mathbf{x} = \mathbf{x}$ we denote $\zeta = 1$,
4. *Commutativity*: $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$.
5. *Associativity*: $(\mathbf{x} + \mathbf{y}) + \mathbf{z} = \mathbf{x} + (\mathbf{y} + \mathbf{z})$ and $\alpha(\beta \mathbf{x}) = (\alpha\beta)\mathbf{x}$,
6. *Distributivity*: $\alpha(\mathbf{x} + \mathbf{y}) = \alpha\mathbf{x} + \alpha\mathbf{y}$ and $(\alpha + \beta)\mathbf{x} = \alpha\mathbf{x} + \beta\mathbf{x}$,

for all $\mathbf{x}, \mathbf{y}, \mathbf{z} \in V$ and $\alpha, \beta \in \mathbb{F}$.

Linear Independence

$\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n \in V$ are *linearly independent* if

$$\sum_{i=1}^n \alpha_i \mathbf{x}_i = 0 \implies \alpha_1, \dots, \alpha_n = 0.$$

Linear Independence

$\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n \in V$ are *linearly independent* if

$$\sum_{i=1}^n \alpha_i \mathbf{x}_i = \mathbf{0} \implies \alpha_1, \dots, \alpha_n = 0.$$

Span

The *span* of $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ in V is

$$\mathcal{L}\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\} = \{\mathbf{x} \in V : \exists_{\alpha_1, \dots, \alpha_n \in \mathbb{F}} : \sum_{i=1}^n \alpha_i \mathbf{x}_i = \mathbf{x}\}.$$

Basis

$\mathcal{B} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ is a *basis* of a vector space V if

$$\forall \mathbf{x} \in V \exists \alpha_1, \dots, \alpha_n \in \mathbb{F} : \sum_{i=1}^n \alpha_i \mathbf{x}_i = \mathbf{x},$$

and $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ are linearly independent.

Norm

Let V be a real vector space. A *Norm* is a function, denoted by $\|\cdot\| : V \rightarrow \mathbb{R}$, that satisfies:

1. $\|\mathbf{x}\| \geq 0$, and $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = 0$,
2. $\|\alpha\mathbf{x}\| = |\alpha|\|\mathbf{x}\|$,
3. $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ (*triangular inequality*).

Normed spaces

Norm

Let V be a real vector space. A *Norm* is a function, denoted by $\|\cdot\| : V \rightarrow \mathbb{R}$, that satisfies:

1. $\|\mathbf{x}\| \geq 0$, and $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = 0$,
2. $\|\alpha\mathbf{x}\| = |\alpha|\|\mathbf{x}\|$,
3. $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ (*triangular inequality*).

Examples (Norms in \mathbb{R}^n):

- $\|\mathbf{x}\|_1 = \sum_{i=1}^n |x_i|$,
- $\|\mathbf{x}\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{\frac{1}{p}}$, $p \geq 1$,
- $\|\mathbf{x}\|_\infty = \max_{1 \leq i \leq n} |x_i|$.

Inner product spaces

Inner product

An *inner product* on a real vector space V is a function $\langle \cdot \rangle : V \times V \rightarrow \mathbb{R}$ satisfying:

1. $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0$ and $\langle \mathbf{x}, \mathbf{x} \rangle = 0$ iff $\mathbf{x} = \mathbf{0}$,
2. $\langle \mathbf{x} + \mathbf{y}, \mathbf{z} \rangle = \langle \mathbf{x}, \mathbf{z} \rangle + \langle \mathbf{y}, \mathbf{z} \rangle$ and $\langle \alpha \mathbf{x}, \mathbf{y} \rangle = \alpha \langle \mathbf{x}, \mathbf{y} \rangle$,
3. $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$,

$\forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in V$ and $\forall \alpha \in \mathbb{R}$.

Example

Inner product in \mathbb{R}^n

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^n x_i y_i = \mathbf{x}^\top \mathbf{y}.$$

Remark

Any inner product in V induces a norm on V : $\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$.

Inner product spaces

Remark

Any inner product in V induces a norm on V : $\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$.

Orthogonality

Two vectors $\mathbf{x}, \mathbf{y} \in V$ are *orthogonal*, $\mathbf{x} \perp \mathbf{y}$ if $\langle \mathbf{x}, \mathbf{y} \rangle = 0$.

Inner product spaces

Remark

Any inner product in V induces a norm on V : $\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$.

Orthogonality

Two vectors $\mathbf{x}, \mathbf{y} \in V$ are *orthogonal*, $\mathbf{x} \perp \mathbf{y}$ if $\langle \mathbf{x}, \mathbf{y} \rangle = 0$.

Pythagorean Theorem

If $\mathbf{x} \perp \mathbf{y}$, then

$$\|\mathbf{x} + \mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2.$$

Inner product spaces

Remark

Any inner product in V induces a norm on V : $\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$.

Orthogonality

Two vectors $\mathbf{x}, \mathbf{y} \in V$ are *orthogonal*, $\mathbf{x} \perp \mathbf{y}$ if $\langle \mathbf{x}, \mathbf{y} \rangle = 0$.

Pythagorean Theorem

If $\mathbf{x} \perp \mathbf{y}$, then

$$\|\mathbf{x} + \mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2.$$

Cauchy-Schwarz Inequality

$$\|\langle \mathbf{x}, \mathbf{y} \rangle\| \leq \|\mathbf{x}\| \|\mathbf{y}\|, \quad \forall \mathbf{x}, \mathbf{y} \in V.$$

Singular value decomposition (SVD) i

Every matrix has the following decomposition

SVD

Let $A \in \mathbb{R}^{m \times n}$ then

$$A = U\Sigma V^T,$$

where $U \in \mathbb{R}^{m \times m}$, $V \in \mathbb{R}^{n \times n}$ are orthogonal matrices (i.e. $UU^T = U^T U = I$) and $\Sigma \in \mathbb{R}^{m \times n}$ is a diagonal matrix with *singular values* of A denoted by σ_i appearing by non-increasing order:

$$\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > \sigma_{r+1} = \dots = \sigma_{\min(m,n)} = 0.$$

Calculus and Optimization

Linear Algebra

Calculus and Optimization

Probability

Review on Statistics

Gradient

Consider a multivariate function $f : \mathbb{R}^d \rightarrow \mathbb{R}$, the *gradient* of f is:

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix} \quad [\nabla f]_i = \frac{\partial f}{\partial x_i} \quad \forall i \in \{1, 2, \dots, d\}$$

$\nabla f(\mathbf{x})$ points in the direction of the steepest ascent from \mathbf{x} .

Jacobian

The *Jacobian* of a vector field $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is:

$$\mathbf{J}_f = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix} \quad [\mathbf{J}_f]_{ij} = \frac{\partial f_i}{\partial x_j}$$

Hessian

The *Hessian* of a vector field $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is:

$$\mathbf{H}_f = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix} \quad [\mathbf{H}_f]_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}$$

Note that: $\mathbf{H}_f(\mathbf{x}) = \mathbf{J}_{\nabla f^\top}(\mathbf{x})$.

Clairaut's Theorem

If the second order partial derivatives of $f : \mathbb{R}^d \rightarrow \mathbb{R}$ are continuous, at a point \mathbf{x} , then

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x}) = \frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{x}), \quad \forall_{i,j \in \{1, \dots, d\}},$$

in this case the Hessian is symmetric $[H_f]_{ij}(\mathbf{x}) = [H_f]_{ji}(\mathbf{x})$.

A lot of the computations in Optimization amounts to finding stationary points (gradient vanishes) and optimal points (stationary plus condition on the Hessian).

Differentiation rules for vectors and matrices

The most important rules for ML are

$$\nabla_{\mathbf{x}}(\mathbf{a}^{\top} \mathbf{x}) = \mathbf{a}$$

$$\nabla_{\mathbf{x}}(\mathbf{x}^{\top} A \mathbf{x}) = \begin{cases} (A + A^{\top})\mathbf{x}, \\ 2A\mathbf{x}, \end{cases} \quad \text{if } A \text{ is symmetric.}$$

Chain rule

For single-variable function

$$(f \circ g)'(x) = f'(g(x))g'(x).$$

Chain rule

For single-variable function

$$(f \circ g)'(x) = f'(g(x))g'(x).$$

Chain rule for multivariate functions

Let $f : \mathbb{R}^m \rightarrow \mathbb{R}^k$ and $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$, then

$$\mathbf{J}_{f \circ g}(\mathbf{x}) = \mathbf{J}_f(g(\mathbf{x}))\mathbf{J}_g(\mathbf{x}).$$

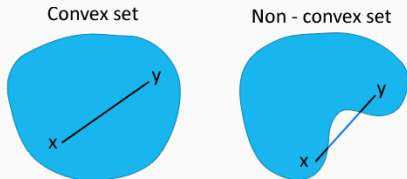
If $k = 1$, we have $f \circ g : \mathbb{R}^n \rightarrow \mathbb{R}$ and

$$\nabla(f \circ g)(\mathbf{x}) = \mathbf{J}_g(\mathbf{x})^\top \nabla f(g(\mathbf{x})).$$

Convexity

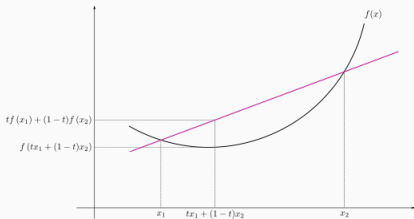
- **Convex set:** A set $\mathcal{X} \subseteq \mathbb{R}^d$ is **convex** if

$$tx + (1 - t)y \in \mathcal{X}, \text{ for all } \mathbf{x}, \mathbf{y} \in \mathcal{X}, \text{ and } t \in [0, 1].$$



- **Convex function:** A function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is **convex** if

$$f(tx + (1 - t)y) \leq tf(x) + (1 - t)f(y) \text{ for all } \mathbf{x}, \mathbf{y} \in \text{dom } f, \text{ and } t \in [0, 1].$$



Probability

Linear Algebra

Calculus and Optimization

Probability

Review on Statistics

Setup

Sample Space: a set of all possible outcomes or realizations of some random trial.

Example: Toss a coin twice; the sample space is $\Omega = \{HH, HT, TH, TT\}$.

Event: A subset of sample space

Example: the event that at least one toss is a head is $A = \{HH, HT, TH\}$.

Probability: We assign a real number $P(A)$ to each event A , called the probability of A .

Probability Axioms: The probability P must satisfy three axioms:

1. $P(A) \geq 0$ for every A ;
2. $P(\Omega) = 1$;
3. If A_1, A_2, \dots are disjoint, then $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

Definition: A random variable is a function that maps from the sample space to the reals ($X : \Omega \rightarrow R$), i.e., it assigns a real number $X(\omega)$ to each outcome ω .

Example: X returns 1 if a coin is heads and 0 if a coin is tails. Y returns the number of heads after 3 flips of a fair coin.

Random variables can take on many values, and we are often interested in the distribution over the values of a random variable, e.g., $P(Y = 0)$

Distribution function

Definition: Suppose X is a random variable, x is a specific value that it can take,

Cumulative distribution function (CDF) is the function $F : R \rightarrow [0, 1]$, where $F(x) = P(X \leq x)$.

If X is discrete \Rightarrow *probability mass function*: $f(x) = P(X = x)$.

If X is continuous \Rightarrow *probability density function* for X if there exists a function f such that $f(x) \geq 0$ for all x , $\int_{-\infty}^{\infty} f(x)dx = 1$ and for every $a \leq b$,

$$P(a \leq X \leq b) = \int_a^b f(x)dx.$$

If $F(x)$ is differentiable everywhere, $f(x) = F'(x)$.

Example of distributions

Discrete variable	Probability function	Mean	Variance
Uniform $X \sim U[1, \dots, N]$	$1/N$	$\frac{N+1}{2}$	
Binomial $X \sim Bin(n, p)$	$\binom{n}{x} p^x (1-p)^{(n-x)}$	np	
Geometric $X \sim Geom(p)$	$(1-p)^{x-1} p$	$1/p$	
Poisson $X \sim Poisson(\lambda)$	$\frac{e^{-\lambda} \lambda^x}{x!}$	λ	
Continuous variable	Probability density function	Mean	Variance
Uniform $X \sim U(a, b)$	$1/(b-a)$	$(a+b)/2$	
Gaussian $X \sim N(\mu, \sigma^2)$	$\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$	μ	
Gamma $X \sim \Gamma(\alpha, \beta) (x \geq 0)$	$\frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}$	$\alpha\beta$	
Exponential $X \sim exponen(\beta)$	$\frac{1}{\beta} e^{-x/\beta}$	β	

Expected Values

- Discrete random variable X , $E[g(X)] = \sum_{x \in \mathcal{X}} g(x)f(x)$;
- Continuous random variable X , $E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)$

Mean and Variance $\mu = E[X]$ is the mean; $\text{var}[X] = E[(X - \mu)^2]$ is the variance.

We also have $\text{var}[X] = E[X^2] - \mu^2$.

Definition:

$$F_{X,Y}(x, y) := P(X \leq x, Y \leq y),$$

and

$$f_{X,Y}(x, y) := \frac{\partial^2 F_{X,Y}(x, y)}{\partial x \partial y},$$

Marginal Distribution of X (Discrete case):

$$f_X(x) = P(X = x) = \sum_y P(X = x, Y = y) = \sum_y f_{X,Y}(x, y)$$

or $f_X(x) = \int_y f_{X,Y}(x, y) dy$ for continuous variable.

Conditional Probability and Bayes Rule

Conditional probability of X given $Y = y$ is

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

Bayes Rule:

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

Independent Variables X and Y are *independent* if and only if:

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

or $f_{X,Y}(x, y) = f_X(x)f_Y(y)$ for all values x and y .

IID variables: *Independent and identically distributed* (IID) random variables are drawn from the same distribution and are all mutually independent.

Linearity of Expectation: Even if X_1, \dots, X_n are not independent,

$$E\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n E[X_i].$$

Review on Statistics

Suppose X_1, \dots, X_n are random variables:

Sample Mean:

$$\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i$$

Sample Variance:

$$S_{N-1}^2 = \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})^2.$$

If X_i are iid:

$$E[\bar{X}] = E[X_i] = \mu,$$

$$\text{Var}(\bar{X}) = \sigma^2/N,$$

$$E[S_{N-1}^2] = \sigma^2$$

Definition The *point estimator* $\hat{\theta}_N$ is a function of samples X_1, \dots, X_N that approximates a parameter θ of the distribution of X_i .

Sample Bias: The bias of an estimator is

$$\text{bias}(\hat{\theta}_N) = E_{\theta}[\hat{\theta}_N] - \theta$$

An estimator is *unbiased estimator* if $E_{\theta}[\hat{\theta}_N] = \theta$

Example

Suppose we have observed N realizations of the random variable X :

$$x_1, x_2, \dots, x_N$$

Then,

- Sample mean $\bar{X} = \frac{1}{N} \sum_n x_n$ is an unbiased estimator of X 's mean.
- Sample variance $S_{N-1}^2 = \frac{1}{N-1} \sum_n (x_n - \bar{X})^2$ is an unbiased estimator of X 's variance
- Sample variance $S_N^2 = \frac{1}{N} \sum_n (x_n - \bar{X})^2$ is *not* an unbiased estimator of X 's variance