Mathematical Background for ML

FHL Workshop

Girum Demisse
girumdemisse@microsoft.com

March 3, 2025

References

- Linear algebra: A great starting point is Gilbert starng's lectures and video Gilber Strang's web site
- Probablity basics: I recommend Chapter-1, and Chpter-2 of Pattern Recognition and Machine learning.
- Overview of ML: If you've an engineering backgourn, this is a great book Informtion Theory, Inference, and Learning Algorithms
- Statistics: good reference book Mathematical statistics and Data Analysis

1 Vector space and transformations

Vector space

- In most cases, a mathematical structure called vector space is used to model the space of collected data points – in a rare and advanced scenarios a manifold (non-linear smooth space) is used as a data representation model.
- Vector space is a set of things that satisfy certain closure axioms.
- A typical vector space is one defined over $\mathbb{R}$ (real number).
- A vector space can be described by a small set of selected vectors called **basis**. Any set of vectors can be basis vectors as long as they are linearly independent.
- The dimensionality of a vector space is determined by the number of basis.
- Apart from solving algebraic problems, you can solve geometric questions by associating a vector space with a coordinate system (Cartesian coordinate system is used as a default coordinate system in most cases).

Linear transformations

- Linear Transformation is the mapping of a vector space from one set of basis to another or can represent a linear coordinate system transformation; transformations are represented by what are called **matrices**.

Change of coordinate system

- Change of coordinate system rarely happens in machine learning problems.

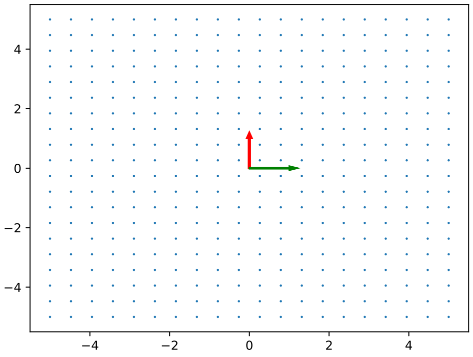


Figure 1: Vector space with selected basis in Cartisian coordinate system.

Change of basis

- In most cases you want to find a set of basis vectors that satisfy your requirement for representing the data. Typical machine learning algorithms of these kinds are linear dimension reduction algorithms, sparsity based model, signal compression methods and so on.
- Vectors change inversely to change of basis, hence are called **contravariant vectors**. Computing transformation between two basis is done as follows, see Figure1.

$$A = B_2 \cdot B_1^{-1} \tag{1}$$

  while coefficent of a contravariant vector change as

$$(A^{-1})^T c_{B_1} = c_{B_2} \tag{2}$$

- Dual vectors (e.g. gradient) change similarly with change of basis, hence are called **covariant vectors**.
- Matrix decomposition: You can have a better understanding of what a transformation matrix is doing by decomposing it into simpler transformations (example methods; SVD, eigen decomposition, ...)

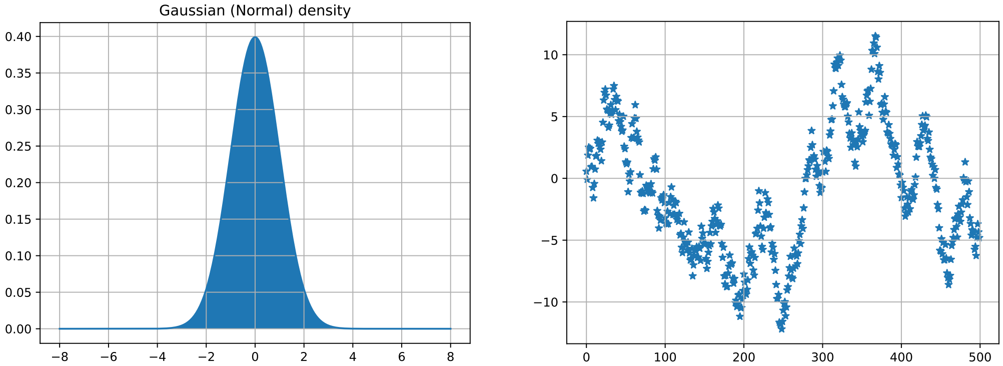
2 Probablity

Random variables

- are functions that map value from some measurable event space to measurable sample space. To ground this intuition, consider the formalism from foundations of probability theory

$$x : (\Omega, \mathcal{F}, \mu) \to (\mathbb{R}, S, \mu_i) \tag{3}$$
$$x : \mathcal{F} \mapsto S \tag{4}$$

- *Vocabulary*
  - Ω : the world of all possible events that can happen in the system under observation.
  - $\mathcal{F}$: a closed subset of Ω (informally, these represent the outcomes you are interested in measuring. Formally, are know as sigma algebra)
  - μ: is a measure defined over $\mathcal{F}$, if it sums up to one it is called probability measure



(a) Normal distribution

(b) $x_{t+1} = x_t + \mathcal{N}(0,1)$

Figure 2: (a)Illustartion of a normal gaussian distribution, (b) a gaussian random walk as a stochastic process.

- x: random variable.

Probablity distributions

- Given a probability measure, the chance of an event w happening is

$$\mu_i(x(w)) \qquad w \subset \mathcal{F} \tag{5}$$

- The probability distribution for x is mathematically modelled by what is known as pdf (probability density function). A function $h(\cdot)$ is a pdf if

$$\mu_i(x(w)) = \int_{e \subset x(w)} h(e)de. \tag{6}$$

- pdf's are not the only option for modelling distributions, you can model distributions with other objects like cdf (cumulative distribution function), mass density function, ...
- Typical distributions:
  - Gaussian (Normal) distribution:

$$f(x) =\sim \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)2}{2\sigma 2}} \tag{7}$$

  - Histograms:

$$\frac{n_i}{N\Delta_i} \tag{8}$$

    where n_i is the number of observed elements in bin i, N the total observed elements, Δ_i is width of the bin.

Working with distributions

- *Bayes rule*: follows from chain rule which itself follows from one of the measure axioms of probability theory. Bayes rule is given as

$$p(x_1|x_2) = \frac{p(x_2|x_1)p(x_1)}{p(x_2)} \tag{9}$$

- *Marginalization*: isolating the distribution for one variable from a joint.

$$
\begin{aligned}
p(x_1) &= \int p(x_1, x_2) dx_2 && (10) \\
&= \int p(x_1|x_2) p(x_2) dx_2 && (11)
\end{aligned}
$$

- *Expectations*: We can compute value of a function that is defined over a random variable. Lets say f is such a function, then we can compute what we should expect f's value to be as

$$E[f(x)] = \int f(x) p(x) dx \tag{12}$$

  - Expectation of a random variables

$$E[x] = \int x p(x) dx \approx \frac{1}{n} \sum_i x_i \tag{13}$$

  - Variance of a random variable's expectation

$$E[(x - E[x])^2] = \int (x - E[x])^2 p(x) dx \approx \frac{1}{n} \sum_i (x - \mu)^2 \tag{14}$$

Stochastic process

- a joint distribution of random variables indexed in time/space.
- Examples, see Figure 2
  - Brownian motion
  - Markov process (typically used in ML for language models, speech recognition, ...)
  - Poission process