Math for Machine Learning

Machine Learning Course, A.Y. 2022/23, Padova

Fabio Aiolli

October 5th, 2022

Probability

- A random experiment is one whose outcome is not predictable with certainty in advance.
- We will mostly consider discrete domains (because they are simpler)
- One interpretation of probability is as a frequency: When an experiment is continuously repeated under the same conditions, for any event E, the proportion of time that the outcome is in E approaches some constant value. This constant limiting frequency is the probability of the event and we denote it as $P(E)$.
- An alternative interpretation is as a degree of belief. What we mean in such a case is a subjective degree of belief in the occurrence of the event.

Axioms of Probability

- $0 \leq P(E) \leq 1$. If E_1 is an event that cannot possibly occur, then $P(E_1) = 0$. If E_2 is certain to occur, $P(E_2) = 1$.
- S is the sample space containing all possible outcomes, $P(S) = 1$
- If $E_i, i = 1, \ldots, n$ are mutually exclusive (i.e., if they cannot occur at the same time, as in $E_i \cap E_j = \emptyset, i \neq j$), we have $P(\cup_{i=1}^n E_i) = \sum_{i=1}^n P(E_i)$.
  In particular, $P(E^c) = 1 - P(E)$ holds if E^c denotes the complement of E.
- If the intersection of E and F is not empty, we have:
  $P(E \cup F) = P(E) + P(F) - P(E \cap F)$

Conditional Probability

- $P(E|F)$ (or posterior probability of E given F) is the probability of the occurrence of event E given that F occurred and is given as $P(E|F) = \frac{P(E \cap F)}{P(F)}$.
- Since $P(F)P(E|F) = P(E \cap F) = P(E)P(F|E)$ (the $\cap$ operator is commutative), we obtain the *Bayes formula*: $P(F|E) = \frac{P(E|F)P(F)}{P(E)}$
- Two events E, F are said independent when $P(E|F) = P(E)$. That is, knowledge of whether F has occurred does not change the probability that E occurs. When events are independent, then $P(E \cap F) = P(E)P(F)$.

Mean and Variance.

The mean (a.k.a. expected value or expectation) of a random variable X, $E[X]$, is the average value of X in a large number of experiments: $E[X] = \sum_i x_i P(x_i)$ (discrete). It has the following properties:

- $E[aX + b] = aE[X] + b$
- $E[X + Y] = E[X] + E[Y]$
- $E[g(X)] = \sum_i g(x_i)P(x_i)$
- $E[X^n] = \sum_i x_i^n P(x_i)$ nth moment

The variance measures how much X varies around the expected value. If $\mu = E[X]$, the variance is defined as:

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

Typically, the symbol σ^2 is used to denote the variance. The standard deviation is defined as $\sigma(X) = \sqrt{Var(X)}$. Same unit as X, easier to interpret.

Popular Distributions

Discrete Case:

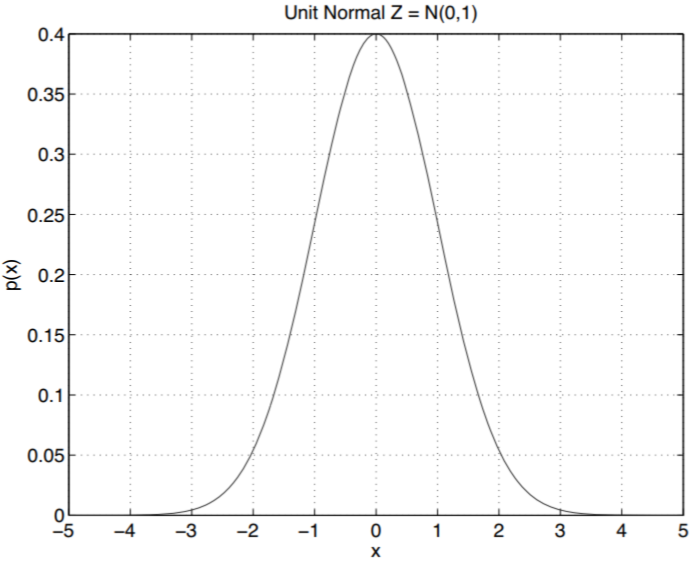
- Bernoulli - Output 1 (success), 0 (failure). p is the probability of success. Then, $P(X=1)=p$ and $P(X=0)=1-p$. $E[X]=p$, $Var(X)=p(1-p)$.
- Binomial - If N identical independent Bernoulli trials are made, the random variable X that represents the number of successes that occurs in N trials is binomial distributed:
  $P(X=i)=\binom{N}{i}p^i(1-p)^{N-i}$, $i=0,\dots N$,
  $E[X]=Np$, $Var(X)=Np(1-p)$.

Continuous Case:

- Uniform in the interval $[a,b]$. Then, $p(x)=\frac{1}{b-a}$ if $a\leq x\leq b$ and $p(x)=0$ otherwise. $E[X]=\frac{a+b}{2}$, $Var(X)=\frac{(b-a)^2}{12}$.
- Normal (Gaussian) of mean μ and variance σ^2, $N(\mu,\sigma^2)$:

$$p(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), -\infty < x < +\infty$$

Normal Distribution $N(0,1)$



Vectors

A n-dimensional vector $\mathbf{x} \in \mathbb{R}^n$, is a collection of n scalar values arranged in a column:

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

Given two n-dimensional vectors $\mathbf{x}, \mathbf{z}$, their sum is still an n-dimensional vector where the elements are summed entry by entry:

$$\begin{pmatrix} x_1 + z_1 \\ x_2 + z_2 \\ \vdots \\ x_n + z_n \end{pmatrix}$$

Multiplication by scalars is trivially the vector obtained by multiplying each entry of the vector by that scalar.

Dot product

Given two n-dimensional vectors $\mathbf{x}, \mathbf{z}$, their dot product is a scalar:

$$\mathbf{x} \cdot \mathbf{z} = \sum_i x_i z_i$$

The length of a vector $\mathbf{x}$ is denoted $|\mathbf{x}|$. The square value of the length is:

$$|\mathbf{x}|^2 = \sum_i x_i^2$$

The dot product has a natural geometrical interpretation:

$$\mathbf{x} \cdot \mathbf{z} = |\mathbf{x}||\mathbf{z}| \cos(\theta)$$

where θ is the angle formed between the two vectors. Note that this quantity is maximized when $\theta = 0$ and is equal to 0 when the vectors are orthogonal.

Binary Vectors and Sets

Consider now binary valued vectors, that is, $\mathbf{x} \in \{0,1\}^n$. Then, a vector can be interpreted as a set by considering a universe of n elements and $\mathbf{x}$ the set containing the elements corresponding to the ones in the vector.

The squared length (a.k.a. norm) of a vector will indicate the number of elements in the set (cardinality of the set).

The dot product between two vectors will indicate the number of shared elements between the two sets (cardinality of the intersection)

We can also compute the projection of a vector $\mathbf{x}$ along the direction of another vector $\mathbf{z}$, as

$$\mathbf{x_z} = \frac{(\mathbf{x} \cdot \mathbf{z})}{|\mathbf{z}|} \frac{\mathbf{z}}{|\mathbf{z}|} = \left(\frac{\mathbf{x} \cdot \mathbf{z}}{\mathbf{z} \cdot \mathbf{z}} \right) \mathbf{z} = \alpha \mathbf{z}$$

The coefficient α has an interesting probabilistic interpretation, i.e. $P(\mathbf{x}|\mathbf{z})$.

Matrices

A matrix $m \times n$, $\mathbf{A} \in \mathbb{R}^{m \times n}$, is a collection of scalar values arranged in a rectangle of m rows and n columns:

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}$$

The i, j entry of the matrix $\mathbf{A}$ can be written $a_{ij} = [\mathbf{A}]_{ij}$

Note that a vector $\mathbf{x} \in \mathbb{R}^n$ can also be seen as a matrix $\mathbf{x} \in \mathbb{R}^{n \times 1}$

Matrix sum and multiplication

Given two matrices, **A** and **B** of the same dimension,

$$[\mathbf{A}+\mathbf{B}]_{ij} = [\mathbf{A}]_{ij} + [\mathbf{B}]_{ij} = a_{ij} + b_{ij}$$

Given two matrices, $\mathbf{A} \in \mathbb{R}^{m \times k}$ and $\mathbf{B} \in \mathbb{R}^{k \times n}$, their product **AB** is the matrix having elements:

$$[\mathbf{AB}]_{ij} = \sum_{q=1}^{k} [\mathbf{A}]_{iq} [\mathbf{B}]_{qj} = \sum_{q=1}^{k} a_{iq} b_{qj}$$

Note: In general $\mathbf{AB} \neq \mathbf{BA}$

Transpose

The transpose $\mathbf{A}^\top \in \mathbb{R}^{m \times n}$ of a matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$ is defined by:

$$[\mathbf{A}^\top]_{ij} = \mathbf{A}_{ji}$$

Properties:

- $(\mathbf{A}^\top)^\top = \mathbf{A}$
- $(\mathbf{AB})^\top = \mathbf{B}^\top \mathbf{A}^\top$

If $\mathbf{A} = \mathbf{A}^\top$ then the matrix $\mathbf{A}$ is said to be symmetric.

Inverse

The identity matrix is a diagonal matrix (necessarily square) $\mathbf{I} \in \mathbb{R}^{n \times n}$ having values equals to 1 in the diagonal and 0 out of the diagonal.

The inverse matrix of a square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is a matrix $\mathbf{A}^{-1}$ such that

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{I} = \mathbf{A}^{-1}\mathbf{A}$$

Note that it is not always possible to find such a matrix (only if the rank is maximal, i.e., the determinant is not equal to 0).

If the inverse matrix exists, then

$$(\mathbf{A}\mathbf{B})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$$

For rectangular matrices, if the square matrix $\mathbf{A}\mathbf{A}^\top$ is invertible, then the matrix $\mathbf{A}^\dagger = \mathbf{A}^\top(\mathbf{A}\mathbf{A}^\top)^{-1}$ (a.k.a. pseudo-inverse) satisfies $\mathbf{A}\mathbf{A}^\dagger = \mathbf{I}$.

Solving linear problems

Problem: Given a square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, and a vector $\mathbf{b}$, find the vector $\mathbf{x}$ such that:

$$\mathbf{Ax} = \mathbf{b},$$

that is we are looking for the linear combination of the columns of $\mathbf{A}$ ($\mathbf{a}_i$) giving $\mathbf{b}$ as result ($\mathbf{b} = \sum_i x_i \mathbf{a}_i$).

Solution:

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$$

Complexity: Solving a linear problem of this type requires $O(n^3)$ operations. There exist more efficient methods which approximate the solution (e.g., the conjugate gradient method).

Positive Definite Matrices

A symmetric matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ with the property that $\mathbf{x}^\top \mathbf{A} \mathbf{x} \geq 0$ for any vector $\mathbf{x} \in \mathbb{R}^n$ is said to be positive semi-definite (eigenvalues ≥ 0).

A symmetric matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ with the property that $\mathbf{x}^\top \mathbf{A} \mathbf{x} > 0$ for any vector $\mathbf{x} \in \mathbb{R}^n$ is said to be positive definite (eigenvalues > 0).

Positive definite matrices are always invertible. Easy to see as the determinant is also the product of the eigenvalues.