Probability Concepts

Sai Bharadwaj Sirigadi

1 Introduction

Given some probability space (Ω, P), a random variable $X : \Omega \to R$ is a function that maps the sample space to the reals.
When we say $P(X = a)$, we actually mean probability of the inverse image $X^{-1}(a)$.
That is, $P(X = a) = P(X^{-1}(a)) = P(\{\omega \in \Omega | X(\omega) = a\})$

$$0 \leq P(X = x) \leq 1$$

$$\sum P(X = x) = 1$$

Write $p(X = x)$ as simply $p(x)$: read probability of x.

1.1 Density Functions

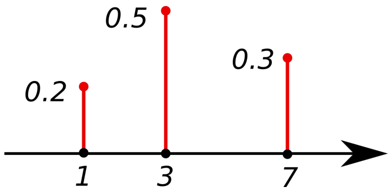
A Probability Function describing the density of continuous/discrete random variable lying between a specific range of values.
The density function describes the dependence between a random variable and its probabilities.

$$f : R \to [0, 1]$$

$$\forall a \in R, f_x(a) = p(X = a)$$

1.1.1 Probability Mass Function (PMF) for discrete RVs.

A function that gives the probability of a discrete random variable is exactly equal to some value.
The value of the random variable having the largest probability mass is called mode.
It is the function $p : R \to [0, 1]$
It is defined as $p_X(x) = P(X = x)$ and satisfies the following:



- The probabilities associated with all values must be non-negative and sum upto 1.

$$\sum_x p_X(x) = 1$$

$$p_X(x) \geq 0$$

1.1.2 Probability Density Function (PDF) for continuous RVs

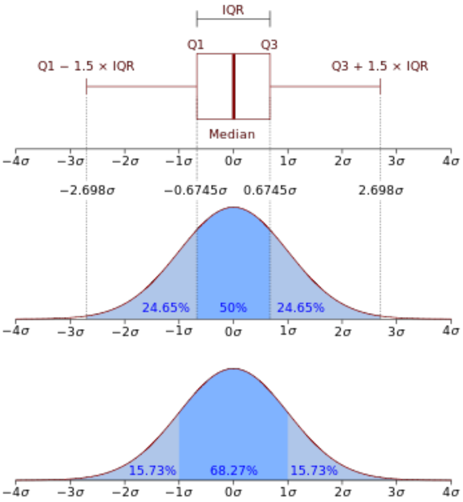
The probability density function(pdf) of a continuous random variable X with support S is an integrable function $f(x)$ which satisfies the following:

- $f(x)$ is positive everywhere in the support S, that is, $f(x) > 0$ for all x in S.
- The area under the curve $f(x)$ in the support S is 1, that is:

$$\int_S f(x)dx = 1$$

- If $f(x)$ is the p.d.f of x, then the probability that x belongs to A, where A is some interval, is given by the integral of $f(x)$ over that interval, that is:

$$P(X \in A) = \int_A f(x)dx$$



1.2 Cumulative Distribution Function

The Cumulative distribution function of a real-valued random variable X is the function given by

$$\forall a \in R, F_x(a) = P(X \leq a)$$

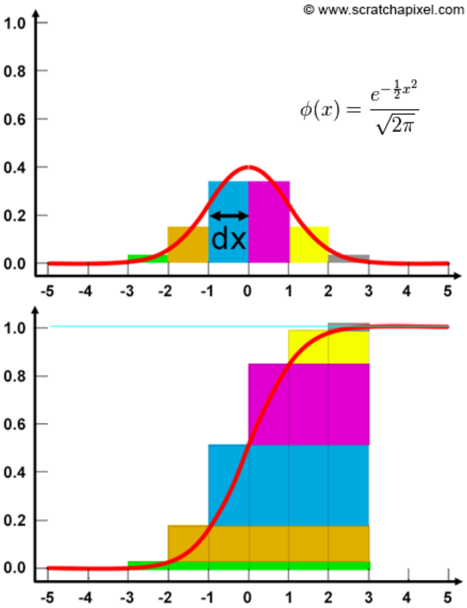
where $F : R \to [0, 1]$
A CDF is monotonically increasing function, i.e.,$\forall x \leq y, F(x) \leq F(y)$
The probability that X lies in the semi-closed interval $(a, b]$, where $a < b$, is therefore,

$$P(a < X \leq b) = F_X(b) - F_X(a)$$

The CDF of a continuous random variable X can be expressed as the integral of its probability density function f_X as follows:

$$F_X(t) = \int_{-\infty}^{a} f_X(t)dt$$

Probability density function(p.d.f): $p(x) = \frac{dF(x)}{dx}$



1.3 Joint Density Function

Joint distributions are high-dimensional PDF(or PMF or CDF).
$f_X(x) \longrightarrow One\ Variable$
$f_{X_1,X_2}(x_1,x_2) \longrightarrow Two\ Variables$
$f_{X_1,X_2,X_3}(x_1,x_2,x_3) \longrightarrow Three\ Variables$
$f_{X_1,...,X_N}(x_1,...,x_N) \longrightarrow N\ Variables$
where $f : R^2 \rightarrow [0,1]$

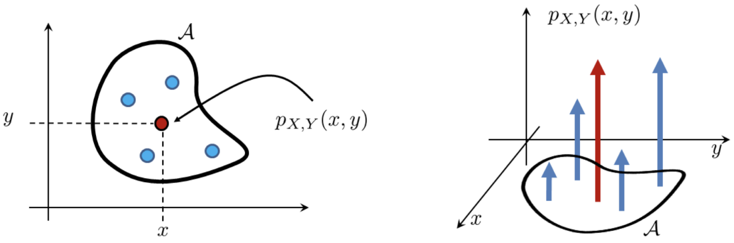
$$\forall(a,b) \in R^2, f_{XY}(a,b) = P(X = a, Y = b)$$

It is a simple multiplication of Marginal densities when random variables are Independent.

1.3.1 Joint PMF

Let X and Y be two discrete random variables. The joint PMF of X and Y is defined as,

$$p_{X,Y}(x,y) = P[X = x\ and\ Y = y]$$

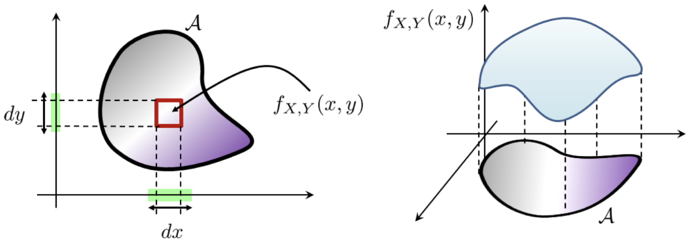


1.3.2 Joint PDF

Let X and Y be two continuous random variables. The joint PDF of X and Y is a function $f_{X,Y}(x,y)$ that can be integrated to yield a probability.

$$P[A] = \int_A f_{X,Y}(x,y)dxdy$$

for any event $A \subseteq \Omega_X \times \Omega_Y$



1.4 Marginal Density Function

Let $X_1, ..., X_K$ be K continuous random variables forming $K \times 1$ continuous random vector. Then, for each $i = 1, ..., K$, the pdf of the random variable X_i, denoted by $f_X(x)$, is called marginal probability density function.

1.4.1 Marginal PMF

The marginal PMF is defined as:

$$p_X(x) = \sum_{y\in\Omega_Y} p_{X,Y}(x,y)\ \ and\ \ p_Y(y) = \sum_{x\in\Omega_X} p_{X,Y}(x,y)$$

1.4.2 Marginal PDF

The marginal PDF is defined as:

$$f_X(x) = \int_{\Omega_Y} f_{X,Y}(x,y)dy\ \ and\ \ f_Y(y) = \int_{\Omega_X} f_{X,Y}(x,y)dx$$

1.5 Conditional Density Function

$$P_{X|Y}(x_i|y_j) = \frac{P_{XY}(x_i, y_j)}{P_Y(y_j)}$$
$$P_{Y|X}(y|x_i) = \frac{P_{XY}(x_i, y)}{P_X(x_i)}$$

1.6 Relationship between Densities

$$conditional\ \ density = \frac{joint\ \ density}{marginal\ \ density}$$
$$p(x|y) = \frac{p(x, y)}{p(y)}$$

1.7 Expectation & Variance

Mean - average value of distribution
For Discrete Random Variable.

$$E(x) = \sum_x xP(X = x) = \mu$$

For Continuous Random Variable.

$$\mu_x = EX = \int_{-\infty}^{\infty} xf_X(x)dx$$

Variance - spread of distribution
For Discrete Random Variable.

$$V(x) = \sum_x x^2P(X = x) = \mu^2$$

For Continuous Random Variable.

$$Var(X) = E[(X - \mu_X)^2] = \int_{-\infty}^{\infty} x^2 f_X(x)dx - \mu_X^2$$

1.8 Independent Random Variables

If two random variables X and Y are independent, then

$$p_{X,Y}(x, y) = p_X(x)p_Y(y),\ \ and\ \ f_{X,Y}(x, y) = f_X(x)f_Y(y)$$

If a sequence of random variables $X_1, ..., X_N$ are independent, then their joint PDF (or joint PMF) can be factorized.

$$f_{X_1,...,X_N}(x_1, ..., x_N) = \prod_{n=1}^{N} f_{X_n}(x_n)$$

1.9 Independent & Identically Distributed

A collection of random variables $X_1, ..., X_N$ are called independent and identically distributed(i.i.d) if

- All $X_1, ..., X_N$ are independent.
- All $X_1, ..., X_N$ have the same distribution, i.e., $f_{X_1}(x) = ... = f_{X_N}(x)$

why is i.i.d. so important?

- If a set of random variables are i.i.d., then the joint PDF can be written as a product of PDFs.
- Integrating a joint PDF is not easy.Integrating a product of PDFs is a lot easier.

1.10 Theorem of Total Probability

The rule states that if the probability of an event is unknown, it can be calculated using the known probabilities of several distinct events.
Used to recover probability of random variable conditioned on another random variable.
Marginalization is an application of Theorem of Probability.

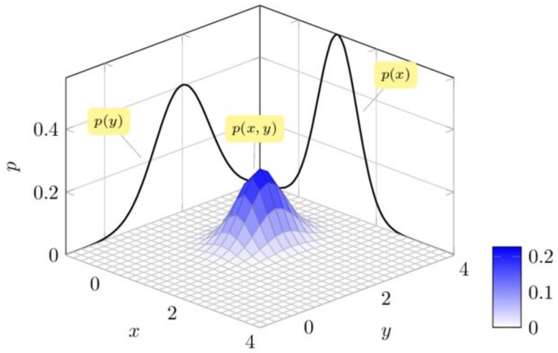
1.11 Bayes Rule

Let X and Y be two random variables.

- $p(x|y)$ read Probability of x given y.
- Conditional Probability: $p(x|y)p(y) = p(x,y)$.

$$p(x|y) = \frac{p(x,y)}{p(y)} \;\; if \;\; p(y) > 0.$$

- $p(x,y) = p(x|y)p(y) = p(y|x)p(x)$
- **Bayes Rules:** $p(x|y) = \frac{p(y|x)p(x)}{p(y) \to \int p(y|x)p(x)dx}$
  This means finding event $p(y|x)p(x)$ considering all events that y distribution can take, which is marginal probability of y i.e., $p(y) = \int p(y|x)p(x)dx$



1.12 Bayes rule multiple variables

We can have arbitrary "Conditioning Variables".

$$p(x|y,z) = \frac{p(y|x,z)p(x|z)}{p(y|z)}$$

1.13 Conditional Independence

It plays an important role. It applies whenever a variable y carries no information about a variable, x is another variables's value, and z is known.

$$p(x,y|z) = p(x|z)p(y|z)$$

1.14 Random Vectors

It is simply, stacking of all random variables in a matrix to form a vector.
Useful when dealing with multiple random vectors.
Makes equations more compact and lets us use useful properties from Linear Algebra.

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$$

1.15 Random Vectors E & Cov

Expectation of Random Vector:

$$EX = \begin{bmatrix} EX_1 \\ EX_2 \\ \vdots \\ EX_n \end{bmatrix}$$

Covariance of Random Vector:

$$C_X = E[(X - EX)(X - EX)^T]$$

which gives

$$\begin{bmatrix} Var(X_1) & Cov(X_1,X_2) & \dots & Cov(X_1,X_n) \\ Cov(X_2,X_1) & Var(X_2) & \dots & Cov(X_2,X_n) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(X_n,X_1) & Cov(X_n,X_2) & \dots & Var(X_n) \end{bmatrix}$$

Notations

- Random Variables: $X, Y, ..$
- Probability of a variable: $p(x), p(X = x)$
- Comma means "and": $p(X = x, Y = y)\ \ or\ \ p(X = x\ \ and\ \ Y = y)$
- Probability for a set of Variables: $P(X, Y, Z, ...)$
- Probability Mass Function: P_X, P_{XY}
- Density Functions: $f_X, f_{XY}\ \ or\ \ P_X, P_{XY}$
- Cumulative Density Functions: F_X, F_{XY}
- Random Vectors: **X,Y**

2 State Estimation

It is a filtering problem.
The goal is to find an estimate of the current state given all available measurements and control inputs.
We model the error between the true state and state obtained from model as a random vector(collection of random variables)

$$y_t = h(x_t) + v_t \Leftrightarrow p(y_t|x_t) = p_{v_t}(y_t - h(x_t))$$

X(state), Z(Measurement), U(Control Input) are the main random Variables.
These variables progressing over time are each treated as random Variables.
Recovering underlying state of Robotics system under the face of uncertainty.
Hope is to capture the max bound of uncertainty and acknowledge the fact that we are making a best guess.

3 Gaussian

By considering distributions, state estimation algorithms can capture and represent the uncertainty associated with the estimated state variables. Instead of providing a single point estimate, the algorithms provide probability distributions that describe the likelihood of different values for the state variables.
Standard Random Variable
A continuous random variable Z is said to be a standard normal random variable, if its PDF is given by

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{(-\frac{z^2}{2})}$$

$Z \sim N(0, 1), then\ \ \ EZ = 0\ \ \ and\ \ \ Var(Z) = 1$
Normal Random Variable
Normal Random Variables can be obtained by shifting and scaling a standard normal random variable.

$$X = \sigma Z + \mu, \quad where \quad \sigma > 0$$

$$EX = \sigma EZ + \mu = \mu$$
$$Var(X) = \sigma^2 Var(Z) = \sigma^2$$

$X \sim N(\mu, \sigma^2)$
Getting Standard Random Variable from Normal.

$$X = \sigma Z + \mu$$
$$Z = \frac{X - \mu}{\sigma}$$

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

When working with probability distributions, it is common to transform them to a standard form or to a normal distribution for various reasons.
Linear transformation of a normal random variable is itself as normal random variable
Linear transformation of a random variables is nothing but a "Error Propagation".

4 Covariance & Correlation

Covariance
Measure of how two random variables are related to each other.
It represents uncertainty of two random variables.

$$Cov(X, Y) = E[(X - E(X))(Y - E(Y))]$$

$$Cov(X, Y) = E[X, Y] - E[Y]E[X]$$

- $Cov(X, X) = Var(X)$
- if X and Y are independent then $Cov(X, Y) = 0$.
- $Cov(X, Y) = Cov(Y, X)$
- $Cov(aX, Y) = aCov(X, Y)$
- $Cov(X + c, Y) = Cov(X, Y)$
- $Cov(X + Y, Z) = Cov(X, Z) + Cov(Y, Z)$

Correlation coefficient correlation between two variables.

$$p(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}$$

- $p(X, Y) = 0$, X and Y are uncorrelated.
- $p(X, Y) > 0$, X and Y are positively correlated.
- $p(X, Y) < 0$, X and Y are negatively correlated.

5 Maximum Likelihood Estimation

The goal of maximum likelihood estimation is to find the mean and covariance of the Gaussian distribution that best explains the measurements.
In statistics, maximum likelihood estimation (MLE) is a method of estimating the parameters of an assumed probability distribution, given some observed data. This is achieved by maximizing a likelihood function so that, under the assumed statistical model, the observed data is most probable.
Lets get our original model.
$y = f(x) + \epsilon$
$Consider \;\; x = f(x)$
$where,$
$random \;\; noise \;\; is \;\; \epsilon \sim N(0, \sigma^2)$
$p(y|x) = N(x, \sigma^2)$
We make measurements y_i from an unknown distribution.

Assumption that the distribution is Gaussian
MLE-which x makes our measurement more likely.

$$\begin{aligned}\hat{x} &= argmax_x p(y|x)\\ p(y|x) &= N(y;x,\sigma^2)\\ &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\frac{-(y-x)^2}{2\sigma^2}\end{aligned}$$

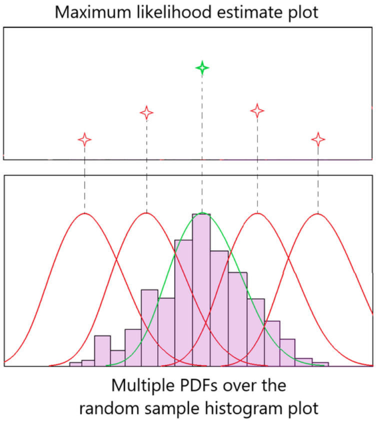
$$p(y|x) = N(y_1;x,\sigma^2)N(y_2;x,\sigma^2)\times\cdots\times N(y_n;x,\sigma^2) = \frac{1}{\sqrt{(2\pi)^n\sigma^2 n}}\exp\frac{-\sum_{i=1}^{n}(y_i-x)^2}{2\sigma^2}$$

$$\hat{x}_{MLE} = argmax_x p(y|x)$$

As $p(y|x)$ is positive,logarithm is used to optimize.

$$\begin{aligned}\hat{x}_{MLE} &= argmax_x log(p(y|x))\\ \log(p(y|x)) &= \frac{-1}{2R}((y_1-x)^2+\cdots+(y_n-x)^2)+c\\ \hat{x}_{MLE} &= argmax_x\frac{-1}{2R}((y_1-x)^2+\cdots+(y_n-x)^2)+c\\ &= argmin_x\frac{1}{2R}((y_1-x)^2+\cdots+(y_n-x)^2)+c\\ argmax_z f(z) &= argmin_z(-f(z))\end{aligned}$$

Least Square is equivalent to maximum Likelihood under Gaussian Noise.



6 Central Limit Theorem

The central limit theorem (CLT) states that the distribution of sample means approximates a normal distribution as the sample size gets larger, regardless of the population's distribution.

6.0.1 Example:

- Let $Z_1, Z_2, ..., Z_n$ be a sequence of n identically distributed random variables.
- Let x be mean of the distribution and $q > 0$ be its variance.
- Then CLT states that, as the sample size n increase the distribution of the sample average of these random variables approaches the normal distribution with a mean x and variance q/n irrespective of the shape of the common distribution.

7 State Estimation

7.1 Recall(State Estimation)

- State of Robotics system is a collection of Random variables or a Random vector(X).
- Measurements are taken by Robot using equipped sensors.These are random variables(Z).
- Control inputs are random variables that bring change in State of the Robot(U).
- These can be Time Varying(X_t, Z_t, U_t).
- The measurements themselves can be from different sources.

7.2 State

State is a collection of all aspects of robot and its environment that can impact the future.
State can be static state or dynamic state.
From Localization perspective, state includes variables regarding the robot itself, such as its pose, velocity etc.

7.3 Belief

Robot's internal knowledge about the state of the environment.
Belief's are represented through conditional probability distributions.
Belief distribution assign a probability(or a density value) to each possible hypothesis with regards to its true state.
Belief distributions are posterior probabilities over state variables conditioned on the available data.

7.4 Bayesian Vs Frequentist

Frequentist: Computation is based on event occurance frequency.
Bayesian treats probability as belief about single event.
Incorporates ”Prior” information.
We can generate ”prior” information for Robotics problems.
Hence, whenever we say probability of ”something” we are talking about belief of that ”something”, taking into consideration the past events.

7.4.1 State Estimation Problem

We want to estimate the world state X from

- Sensor Measurements Z.
- Controls (or Odometry readings) u.

We need to model relationship between these random variables i.e.,

- $p(x|z)$ & $p(x^{'}|x, u)$

7.5 Bayes rule on measurement

$p(x|z)$(posterior) is diagnostic.
$p(z|x)$(Likelihood) is casual.
Diagnostic models are often hard to find, except when it is static state problem.
Casual models are easy to obtain.
Bayes rule allows us to use casual knowledge in diagnostic reasoning.

$$p(x, z) = p(x|z)p(z) = p(z|x)p(x)$$
$$p(x|z) = \frac{p(z|x)p(x)}{p(z)}$$

where,
$p(x|z)$ is Posterior.
$p(z|x)$ is Observation Likelihood.
$p(x)$ is Prior on World state.

Direct computation of $p(z)$ is hard, as it is a marginal distribution.

$$p(z) = \sum_x p(z, x) = \sum_x p(z|x)p(x)$$

$$p(x|z) = \frac{p(z|x)p(x)}{\sum_x p(z|x)p(x)} \propto \eta p(z|x)p(x)$$

where,
η is a Normalization constant, that makes it as PDF.

References

[1] Skill-Lync, *Localization, Mapping & SLAM.*

[2] Sebastian Thrun, Wolfram Burgard and Dieter Fox. *Probabilistic Robotics* (2005).