Bayesian Probabilities

There ~~are~~ are two school of thoughts in probability theory —

① Frequentist or Classical → random repeatable events

② Bayesian View → quantification of uncertainity.

There are events which doesn't get repeated numerous times in order to suit the classical or frequentist probability.

Example - how quickly ice cap is melting. In many circumstances, we would like to be able to quantify our expression of uncertainity and make ~~precision~~ precise revisions of uncertainity in presence of new evidence. This can be done using Bayesian interpretation of probability.

Cox in 1946, showed that if numerical values are used to represent degrees of belief, then a simple set of axioms ~~~~ encoding common sense properties of such beliefs leads uniquely to a set of rules for manipulating degrees of belief that are equivalent to the sum and product rule of probability.

— x —

In field of pattern recognition. a general notion of probability helps a lot.

Bayes Theorem –

$$p(w|D) = \frac{p(D|w) \cdot p(w)}{p(D)}$$

w = parameters

D = observed data $= \{t_1, \ldots t_N\}$

This allows us to evaluate the uncertainity in w after we have observed D in the form of the posterior probability $p(w|D)$.

The quantity $p(D|w)$ can be viewed as a function of the parameter vector w, it is called likelihood function. It expresses how probable the observed data set is for different settings of the parameter vector w.

posterior $\propto$ likelihood $\times$ prior.

The denominator is a normalization constant

$$p(D) = \int p(D|w) \cdot p(w)\, dw.$$

In both Bayesian and frequentist paradigms, the likelihood function $p(D|w)$ plays a central role.

In frequentist setting, w is considered fixed whose value is determined by some estimator and error bars. considering some distribution of data.

In Bayesian there is only a single dataset D and uncertainity in the parameters is expressed thro' a p.d over w.

Maximum likelihood, in which 'w' is set to the value which maximizes $p(D|w)$. (negative log likelihood error function, minimization).

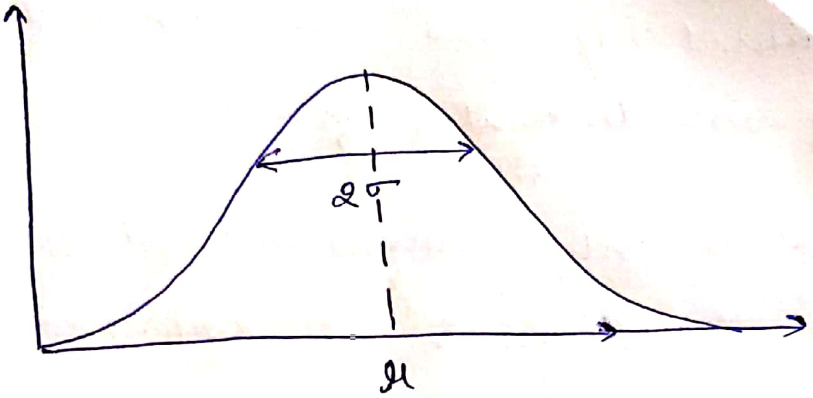
One advantage with Bayesian viewpoint, is inclusion of prior knowledge arises naturally. (talk about a fair coin toss).

"Monte carlo":

Gaussian Distribution

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

μ = mean , σ = std. deviation.



$$\int_{-\infty}^{\infty} N(x|\mu,\sigma^2)\,dx = 1.$$

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} N(x|\mu,\sigma^2)\, x\, dx = \mu$$

$$\mathbb{E}[x^2] = \int_{-\infty}^{\infty} N(x|\mu,\sigma^2) \cdot x^2\, dx = \mu^2 + \sigma^2$$

$$\mathrm{Var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x] = \sigma^2.$$

The max. of a distribution is known as mode.
For Gaussian, it coincides with mean.

over D- dimensions -

$$N(x|\mu,\Sigma) = \frac{1}{(2\pi)^{D/2}} \times \frac{1}{|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right\}$$

D- dimensional vector μ → mean

Σ = D X D dimensional vector → covariance.

$|\Sigma|$ determinant of Σ.

Data points that are drawn independently from the same distribution are said to be i.i.d. independent and identical distribution.

~~iden~~ independent
