Review of Probability and Estimators

Arun Das, Jason Rebello
16/05/2017

- State of robot (position, velocity) and state of its environment are unknown and only noisy sensors available (GPS, IMU)

- Probability helps to fuse sensory information

- Provides a distribution over possible states of the robot and environment

Probability for any event A in the set of all possible outcomes.

$$0 \le \Pr(A) \le 1$$

Probability over the set of all possible outcomes

$$\Pr(\Omega) = 1$$

Probability of the Union of events

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$$

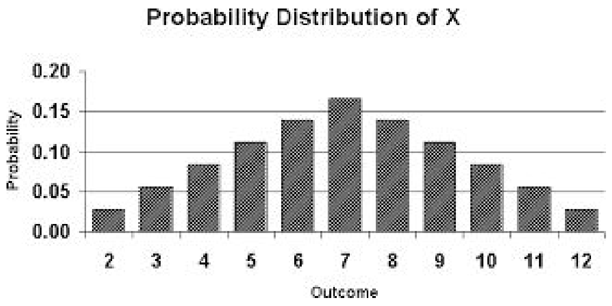
If the events are mutually exclusive

$$P(A_1 \cup A_2 \cup \ldots \cup A_n) = \sum_{i=1}^{n} P(A_i)$$

Random Variable assigns a value to each possible outcome of a probabilistic experiment. Example: Toss 2 dice: random Variable X is the sum of the numbers on the dice

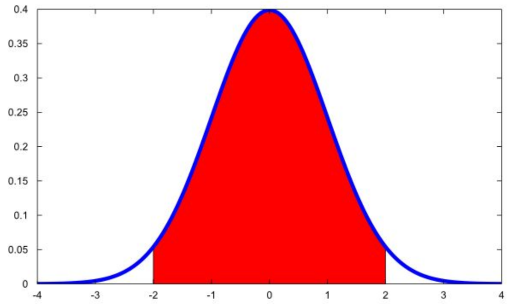
Discrete

- Distinct Values
- Probability Mass function
- Eg: X = (1) Heads, (0) Tails
  Y = Year a random student was born
  (2000,2001,2002,..)



Continuous

- Any value in some interval
- Probability Density function
- Eg: Weight of random animal



It is the long-run average value of repetitions of the experiment it represents. Expectation is the probability weighted average of all possible values.

$E[X] = \sum_x x * P(x)$

Eg. For a dice. $E(X) = (⅙)*1 + (⅙)*2 + (⅙)*3 + (⅙)*4 + (⅙)*5 + (⅙)*6 = 3.5$

Properties:

1) $E[c] = c$ … where 'c' is a constant

2) $E[X+Y] = E[X] + E[Y]$ and $E[aX] = aE[X]$ .. Expected value operator is linear

3) $E[X \mid Y=y] = \sum_x x * P(X=x \mid Y=y)$ … Conditional Expectation

4) $E[X] \leq E[Y]$ if $X \leq Y$ … Inequality condition

Variance measures the dispersion around the mean value.

$\mathrm{Var}[X] = \sigma^2 = E[X-\mu]^2$

$\mathrm{Var}[X] = E[X^2] - E[X]^2$

Eg: For a dice.

$E[X^2] = (1/6)*1^2 + (1/6)*2^2 + (1/6)*3^2 + (1/6)*4^2 + (1/6)*5^2 + (1/6)*6^2 = 91/6$

$E[X] = 7/2$

$\mathrm{Var}[X] = (91/6) - (7/2)^2 = 2.9166667$

Properties:

1) $\mathrm{Var}[aX+b] = a^2\,\mathrm{Var}[X]$ .. Variance is not linear

2) $\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]$ .. If X and Y are independent

Joint Probability	$P(X{=}x \text{ and } Y{=}y) = P(x,y)$
Joint Probability of Independent random variables	If X and Y are independent then $P(x,y) = P(x)\,P(y)$
Conditional Probability	$P(x \mid y)$ is the probability of x given y $P(x \mid y) = P(x,y) \,/\, P(y)$ $P(x,y) = P(x \mid y)\, P(y)$
Conditional Probability of Independent random variables	If X and Y are independent then $P(x \mid y) = P(x)$

What is the difference between Mutually exclusive events and Independent Events ?

- Events are mutually exclusive if the occurrence of one event excludes the occurrence of other events. Eg Tossing a coin. The result can either be heads or tails but not both

  $P(A \cup B) = P(A) + P(B)$

  $P(A,B) = 0$

- Events are independent if the occurrence of one event does not influence the occurrence of the other event. Eg Tossing two coins. The result of first flip does not affect the result of the second

  $P(A \cup B) = P(A) + P(B) - P(A)*P(B)$

  $P(A,B) = P(A)*P(B)$

Discrete case

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

$$P(x) = \sum_y P(x, y)$$

$$P(x) = \sum_y P(x \mid y) P(y)$$

Continuous case

$$\int p(x)\, dx = 1$$

$$p(x) = \int p(x, y)\, dy$$

$$p(x) = \int p(x \mid y) p(y)\, dy$$

Conditional distribution: $P(H|L)$

L / H	Red	Yellow	Green
Not Hit	0.99	0.9	0.2
Hit	0.01	0.1	0.8

Joint distribution: $P(H,L)$

L / H	Red	Yellow	Green	Marginal probability P(H)
Not Hit	0.198	0.09	0.14	0.428
Hit	0.002	0.01	0.56	0.572
Total	0.2	0.1	0.7	1

- Calculate Marginal Probability of person being hit by car without paying attention to traffic light ?
- Assume P(L=red) = 0.2, P(L=yellow)=0.1, P(L=green)=0.7
- P(hit | colour)+P(not hit | colour) = 1
- P(hit, L=red) = P(hit | L=red) * P(L=red) = 0.01*0.2 = 0.002
- $\sum_{colour}$P(hit) = $\sum_{colour}$P(hit,all colour)

  = $\sum_{colour}$P(hit | any colour) * P(any colour)

  = P(hit | red) * P(red) + P(hit | yellow) * P(yellow)+ P(hit | green) * P(green)

  = 0.01*0.2 + 0.1*0.1 + 0.8*0.7 = 0.572

Experiment: We flip a coin 10 times and have the following outcome.

What is the Probability that the next coin flip is T ?

≅ 0.3 ≅ 0.38 ≅ 0.5 ≅ 0.76

Every flip is random. So every sequence of flips is random. We have a parameter that tells us if the next flip is going to be tails.

$$p_i(F_i = \text{T}) = \theta_i.$$

The sequence is modeled by the parameters θ_1, ….. θ_{10}

$$p(\text{H}\,\text{T}\,\text{H}\,\text{H}\,\text{T}\,\text{H}\,\text{H}\,\text{H}\,\text{T}\,\text{H} \mid \theta_1, \theta_2, \ldots, \theta_{10})$$

Find θ_i's such that the above probability is as high as possible.

Maximize the likelihood of our observation (Maximum Likelihood)

Assumption 1 (Independence): The coin flips do not affect each other

$$
\begin{aligned}
& p(\text{H T H H T H H H T H} \mid \theta_1, \theta_2, \ldots, \theta_{10}) \\
&= p_1(F_1 = \text{H} \mid \theta_1) \cdot p_2(F_2 = \text{T} \mid \theta_2) \cdot \ldots \cdot p_{10}(F_{10} = \text{H} \mid \theta_{10}) \\
&= \prod_{i=1}^{10} p_i(F_i = f_i \mid \theta_i)
\end{aligned}
$$

Assumption 2 (Identically Distributed): The coin flips are qualitatively the same

$$
\prod_{i=1}^{10} p_i(F_i = f_i \mid \theta_i) = \prod_{i=1}^{10} p(F_i = f_i \mid \theta)
$$

Independent and **I**dentically **D**istributed : Each random variable has the same probability distribution as others and all are mutually independent

$$\prod_{i=1}^{10} p(F_i = f_i \mid \theta) = (1-\theta)\theta(1-\theta)(1-\theta)\theta(1-\theta)(1-\theta)(1-\theta)\theta(1-\theta)$$
$$= \theta^3(1-\theta)^7$$

Find critical point of the above function.

Monotonic functions preserve critical points. Use log to make things simpler

argmax$_\theta$ ln [θ^3 $(1-\theta)^7$]

= argmax$_\theta$ |T| ln θ + |H| ln (1-θ)

Taking the derivative and equating to 0.

$$\theta_{\text{MLE}} = \frac{|T|}{|T|+|H|} \quad = 3/10 = 0.3$$

This justifies the answer of 0.3 to the original question.

Suppose there is a sample x_1 .. x_n of n independent and identically distributed observations coming from a distribution with an unknown probability density function.

Joint Density Function: $$f(x_1, x_2, \ldots, x_n \mid \theta) = f(x_1 \mid \theta) \times f(x_2 \mid \theta) \times \cdots \times f(x_n \mid \theta).$$

Consider the observed values to be fixed parameters and allow θ to vary freely

Likelihood: $$\mathcal{L}(\theta\,;x_1,\ldots,x_n) = f(x_1, x_2, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta).$$

More convenient to work with natural log of the likelihood function

Log-Likelihood: $$\ln \mathcal{L}(\theta\,;x_1,\ldots,x_n) = \sum_{i=1}^{n} \ln f(x_i \mid \theta),$$

Average Log-Likelihood: $$\hat{\ell} = \frac{1}{n} \ln \mathcal{L}.$$

Maximum Likelihood Estimator: $$\{\hat{\theta}_{\mathrm{mle}}\} \subseteq \{\underset{\theta\in\Theta}{\arg\max}\ \hat{\ell}\,(\theta\,;x_1,\ldots,x_n)\},$$

Let's assume we tossed the coin twice and got the following sequence

H H

The probability of seeing a tails in the next toss is $\theta_{\text{MLE}} = \frac{|T|}{|T|+|H|}$

Since no tails observed $\theta_{\text{MLE}} = 0$

MLE is a point estimator and is prone to Overfitting.

How do we solve this ? Assume a prior on θ

Bayes Rule

$$p(\theta = x \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta = x) \cdot p(\theta = x)}{p(\mathcal{D})}$$

$p(\mathcal{D} \mid \theta = x)$: Likelihood

$p(\theta = x)$: Prior

$p(\mathcal{D})$: Evidence

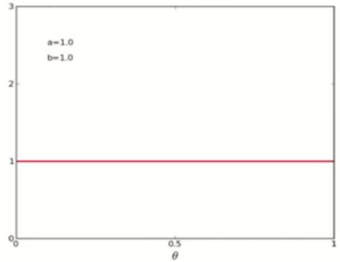
$p(\theta = x \mid \mathcal{D})$: Posterior

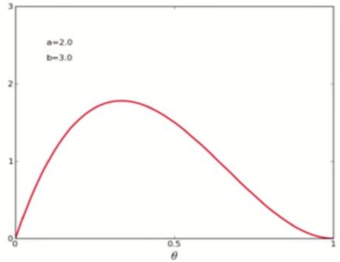
Choosing the Prior

$$\mathrm{Beta}(\theta \mid a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1}, \quad \theta \in [0,1]$$

(a-1) and (b-1) are the number of T and H we think we would see, if we made (a+b-2) many coin flips

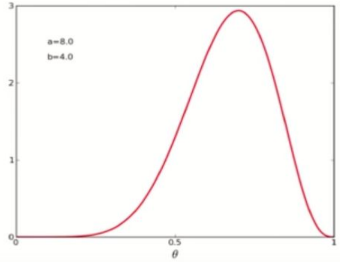
a = 1.0
b = 1.0

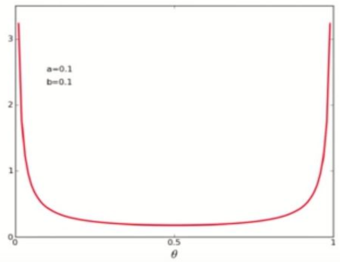




a = 2.0
b = 3.0

a = 8.0
b = 4.0





a = 0.1
b = 0.1

Evidence Likelihood Prior

$$p(\theta = x \mid \mathcal{D}) = \boxed{\frac{1}{p(\mathcal{D})}} \cdot \boxed{x^{|T|}(1-x)^{|H|}} \cdot \boxed{\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}}$$
$$\propto x^{|T|+a-1}(1-x)^{|H|+b-1},$$

Maximum A Posteriori Estimation

$$\theta_{\text{MAP}} = \frac{|T| + a - 1}{|H| + |T| + a + b - 2}$$

Determine the Maximum Likelihood Estimator for the mean and variance of a Gaussian Distribution