BLG 527E Machine Learning

FALL 2021-2022

Assoc. Prof. Yusuf Yaslan & Assist. Prof. Ayse Tosun

Probability and Statistics for Machine Learning

Random variables

- Assign numerical values to **random events**
- Random: (like tossing a coin) we do not know what value the variable will take until the event happens
- Variable: takes a number of numerical values
- **X** is a discrete random variable whose values can be systematically listed through all possible outcomes (**sample space**)
- **X** is a continuous random variable whose values cannot be systematically listed with a list of all possible outcomes

- Toss a coin, Roll of a die, Bug proneness of a software class, Failure of a node in a computer network → Discrete
- Height of a human, Score of a Olympics race, Temperature of the next week → Continuous

Probability

- Y tossing a coin. Heads = 1, Tails = 0
- $P(Y=1)$ and $P(Y=0)$
- How to calculate $P(Y=1)$?
- If we knew that $P(Y=0) = 0.3$, can we compute $P(Y=1)$?
- $\sum P(Y=y) = 1$
- $P(Y=y) \rightarrow P(y)$ **Probability distribution**

We roll a die, what is the probability that the result is 4? $P(Y=4)$

What is the probability that the result is less than 4? $P(Y<4)$

What is the probability that the result is not 4? $P(Y\neq4)$

Conditional probability

- When the outcome of one event affects the outcome of another.
- X for tossing a coin, and Y for me telling you the result of the tossing
- P(Y=y|X=x) or P(y|x)
- If I am honest, P(Y=1|X=1) = 1 and P(Y=0|X=0) = 1
- What are P(Y=0|X=1) and P(Y=1|X=0)?
-
- What if I sometimes lie? The probability of me telling ‘heads’ when the coin is ‘heads’ is 0.8, whereas telling ‘tails’ when it is ‘heads’ is 0.2
  - P(Y=heads|X=heads) = 0.8
  - P(Y=tails|X=heads) = 0.2
  - ∑P(Y=y|X=1) = 1 and ∑P(Y=y|X=0) = 1

Joint probability

- $P(Y=y, X=x)$
- Depends on whether the random variables are *dependent*
- If there is no dependence,
  - $P(Y=y, X=x) = P(Y=y) * P(X=x)$
- If there is dependence
  - $P(Y=y, X=x) = P(Y=y|X=x) * P(X=x)$
- The probability that coin is ‘heads’ AND I say ‘heads’
  - $P(Y=heads, X=heads) = P(Y=heads|X=heads) * P(X=heads)$

$$= 0.8 \times 0.5$$

- $\sum_{x,y} P(Y=y, X=x) = 1$

Marginalisation

- $P(Y=y)$ is calculated by **marginalising** out X from the joint distribution $P(Y=y, X=x)$
  - $P(Y=y) = \Sigma_{x} P(Y=y, X=x)$
- In the coin example:
  - $P(Y=y, X=0) + P(Y=y, X=1)$

- For joint distribution of J random variables, to get $P(Y_J=y_j)$ the marginal distribution of one of them is given by
  - $$P(y_j) = \sum_{y_1,\ldots y_{j-1},\, y_{j+1},\ldots,y_J} p(y_1, \ldots, y_J)$$

Example

- We have 3r and 1b balls.
- Probability of drawing 2r balls:
  - $P(B_1=r, B_2=r)= P(B_2=r|B_1=r)P(B_1=r)$
- Probability of drawing the second ball red?
  - $P(B_2=r) = P(B_2=r,B_1=r) + P(B_2=r,B_1=b)$

Example

- We have a disease whose test is 99% accurate.
  - Given that you have the disease, the probability that the test is positive $P(T=1|D=1)$
- This is a rare disease hitting 1 out of 10,000 people
- What is the chance that we actually have the disease?
  - $P(D=1|T=1) = P(T=1|D=1) * P(D=1) / P(T=1)$
  - $P(T=1) = P(T=1|D=1) * P(D=1) + P(T=1|D=0) * P(D=0)$

Monty Hall

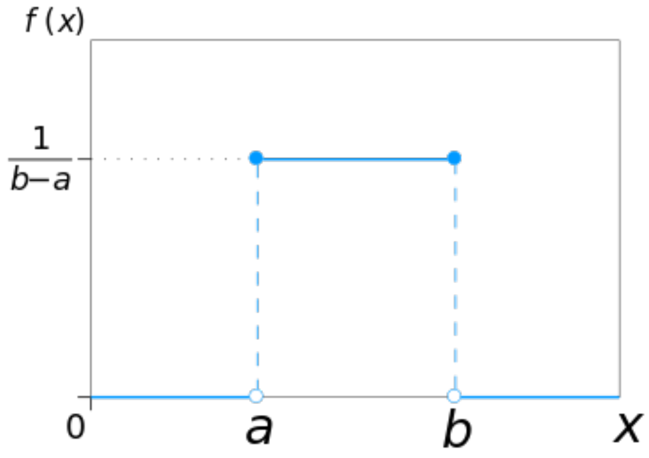
Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice?

Popular discrete distributions

- Bernoulli (toss a coin)
  - $P(X = x) = q^x (1 - q)^{1-x}$
- Binomial (observing certain number of heads in a total of N tosses)
  - $P(Y = y) = P(y) = \binom{N}{y} q^y (1 - q)^{N-y}$
- Multinomial
  - $P(y) = \frac{N!}{\Pi_j y_j!} \Pi_j q_j^{y_j}$

Popular continuous distributions

- Uniform
  - all events in a partition are equally likely.



- Gaussian
  - $p(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma^2}(x-\mu)^2)$

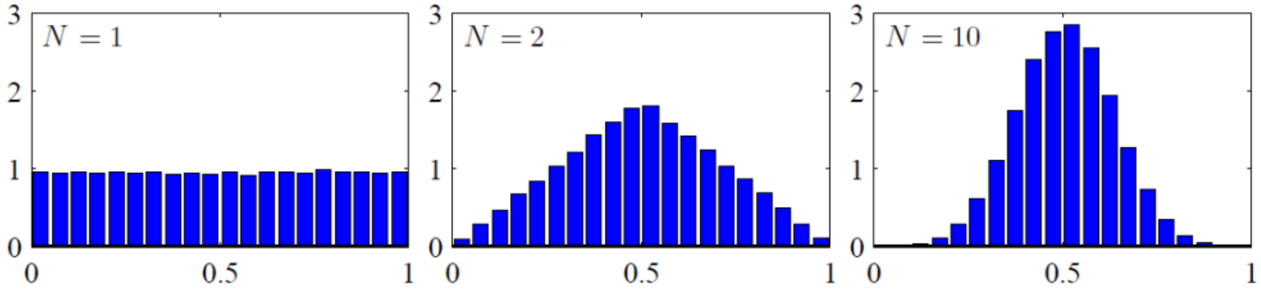


Figure 6: Histogram plots of the mean of N uniformly distributed numbers for various values of N. The effect of the Central Limit Theorem is seen: as N increases, the distribution becomes more Gaussian. (Figure from *Pattern Recognition and Machine Learning* by Chris Bishop.)