Data Mining & Machine Learning

CS37300
Purdue University

Aug 28, 2017

Announcement

- Homework due date changed to Sept 6, 11:59pm
    - Many new students just enrolled
- There is much setting up to do. Downloading files, getting a python editor, getting python to work, etc.. START NOW!
- There will be no further extensions!

Probability and statistics basics

Modeling uncertainty

- Necessary component of almost all data analysis
- Approaches to modeling uncertainty:
  - Fuzzy logic
  - Possibility theory
  - Rough sets
  - **Probability *(focus in this course)***

Probability

- Probability theory *(some disagreement)*
  - Concerned with interpretation of probability
  - 17th century: Pascal and Fermat develop probability theory to analyze games of chance
- Probability calculus *(universal agreement)*
  - Concerned with manipulation of mathematical representations
  - 1933: Kolmogorov states axioms of modern probability

Probability basics

- Basic element: **Random variable**
  - Mapping from a property of objects to a variable that can take one of a set of possible values
  - *X* refers to random variable; *x* refers to a value of that random variable
- Types of random variables
  - Discrete RV has a finite set of possible values;
    Continuous RV can take any value within an interval
  - Boolean: e.g., Warning (is there a storm warning? = <yes, no>)
  - Discrete: e.g., Weather is one of <sunny,rainy,cloudy,snow>
  - Continuous: e.g., Temperature

Probability basics

- **Sample space (S)**
  - Set of all possible outcomes of an experiment
- **Event**
  - Any subset of *outcomes* contained in the sample space S
  - When events ***A*** and ***B*** have no outcomes in common they are said to be *mutually exclusive*

Examples

Random variable(s)	**Sample space**
One coin toss	H, T
Two coin tosses	HH, HT, TH, TT
Select one card	2♥, 2♦, ..., A♣ (52)
Play a chess game	Win, Lose, Draw
Inspect a part	Defective, OK
Cavity and toothache	TT, TF, FT, FF

Axioms of probability

- For a sample space S with possible events $\mathbf{A}_\mathbf{S}$, a function that associates real values with each event A is called a ***probability function*** if the following properties are satisfied:

  1. $0 \leq P(A) \leq 1$ for every A

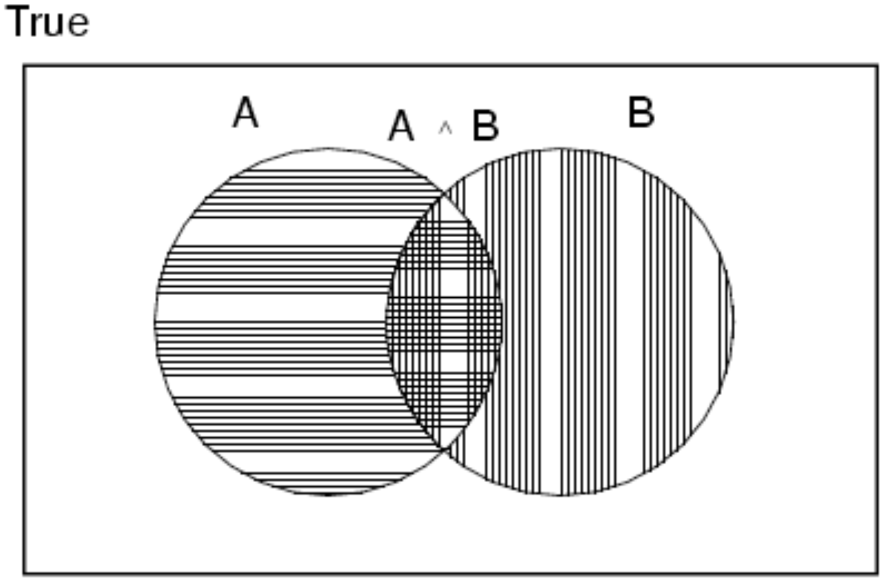
  2. $P(S) = 1$

  3. $P(A_1 \text{ U } A_2 \ \ldots \text{ U } A_{n \in S}) = P(A_1) + P(A_2) + \ldots + P(A_n)$

     if A_1, A_2, ..., A_n *are pairwise mutually exclusive events*

Implications of axioms

- For any events ***A, B*** in universe S
  - P(*A*) = 1 - P(S\A)
  - P(true) = 1   and   P(false) = 0
  - If A and B are mutually exclusive then P(*A* ∩ *B*) = 0
  - P(*A* U *B*) = P(*A*) + P(*B*) - P(*A* ∩ *B*)



Interpreting probabilities

- Meaning of probability is focus of debate and controversy
- Two main views: Frequentist and Bayesian

Frequentist view

- Dominant perspective for last century
- Probability is an **objective** concept
  - Defined as the frequency of an event occurring under repeated trials in "same" situation
  - E.g., number of heads in repeated coin tosses
- Restricts application of probability to repeatable events

Bayesian vs. frequentist

- Bayesian central tenet:
  - Explicitly model all forms of uncertainty
  - E.g., Parameters, model structure, predictions
- Frequentist often model same uncertainty but in less-principled manner, e.g.,:
  - Parameters set by cross-validation
  - Model structure averaged in ensembles
  - Smoothing of predicted probabilities
- Although interpretation of probability is different, underlying calculus is the same

Calculating probabilities (frequentist)

- Frequentist view
  - Let n be the number of times an experiment is performed
  - Let n(A) be the number of outcomes in which A occurs
  - Then as $n \to \infty$ $P(A) = n(A) / n$
- When the various outcomes of an experiment are equally likely, the task of computing probability reduces to counting
  - Let N be size of sample space (i.e., number of simple outcomes)
  - Let N(A) be the number of outcomes contained in A
  - Then: $P(A) = N(A) / N$

Example

- Roll two 6-sided dice. What is the probability that the result sums to 8?
  - P = num ways event can occur / possible outcomes
- What is the size of the sample space?
  - 6*6=36
- How many events involve the two dice summing to 8?
  - {2,6},{3,5},{4,4},{5,3},{6,2} = 5
- Overall probability?
  - 5/36 = 0.139

Permutations and combinations

- An **ordered** sequence of k objects taken from a set of n distinct objects without replacement, is called a **permutation** of size k
    - The number of permutations of size k that can be constructed from the n objects is:

$$P_{k,n} = \frac{n!}{(n-k)!}$$

- An **unordered** sequence of k objects taken from a set of n distinct objects without replacement, is called a **combination** of size k
    - The number of combinations of size k that can be constructed from the n objects is:

$$C_{k,n} = \frac{P_{k,n}}{k!} = \frac{n!}{k!(n-k)!}$$

Example

- An urn contains ten balls, six of which are red and four of which are white. Five balls are drawn at random. What is the probability of drawing three red and two white balls?

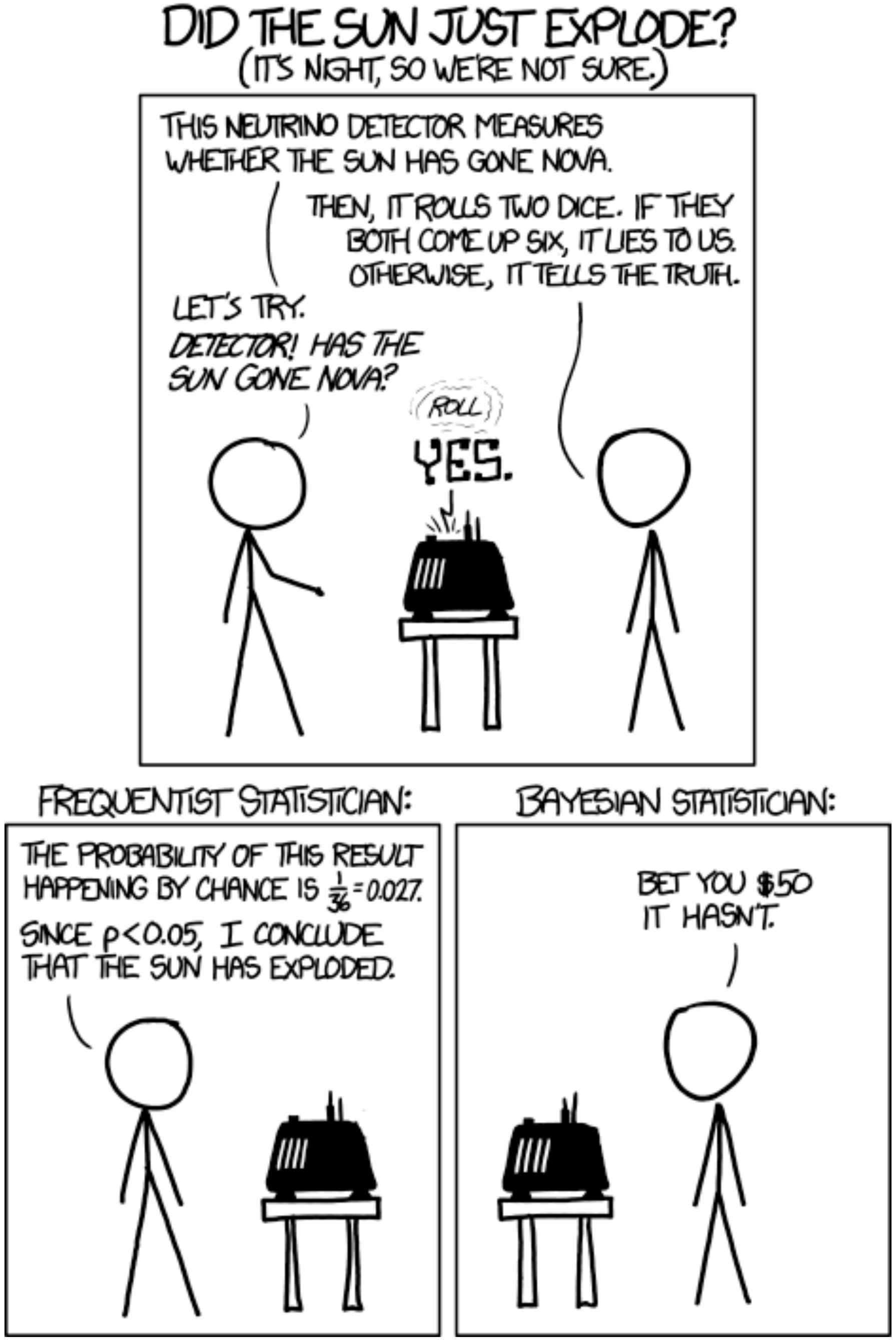
$$\frac{C_{3,6} \cdot C_{2,4}}{C_{5,10}} = \frac{\binom{6}{3}\binom{4}{2}}{\binom{10}{5}} = \frac{6!}{3!3!}\frac{4!}{2!2!}\frac{5!5!}{10!}$$

- An urn contains five balls, numbered from 1 to 5. Three balls are drawn at random. What is the probability that we draw the sequence 3, 4, 1?

$$\frac{1}{P_{3,5}} = \frac{(5-3)!}{5!}$$

Bayesian view

- Increasing importance over last decades
    - Due to increase in computational power that facilitates previously intractable calculations
- Probability is a **subjective** concept
    - Defined as individual degree-of-belief that event will occur
    - E.g., belief that we will have another snow storm tomorrow
- Begin with prior belief estimates and update those by conditioning on observed data


DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)
THIS NEUTRINO DETECTOR MEASURES WHETHER THE SUN HAS GONE NOVA.
THEN, IT ROLLS TWO DICE. IF THEY BOTH COME UP SIX, IT LIES TO US. OTHERWISE, IT TELLS THE TRUTH.
LET'S TRY.
DETECTOR! HAS THE SUN GONE NOVA?
(ROLL)
YES.
FREQUENTIST STATISTICIAN:
THE PROBABILITY OF THIS RESULT HAPPENING BY CHANCE IS $\frac{1}{36}$=0.027.
SINCE $p<0.05$, I CONCLUDE THAT THE SUN HAS EXPLODED.
BAYESIAN STATISTICIAN:
BET YOU $50 IT HASN'T.

http://xkcd.com/1132/

Calculating probabilities: Bayesian

- *Begin with prior belief estimates*: P(A)
  - E.g., After the Seahawks won their conference, Vegas casinos believed the Seahawks were likely to win the Superbowl over the Patriots:
    P(S wins)=0.525, P(P wins)=0.475
- Observe data
  - But then Vegas observed a heavy majority of the betters (80%) chose the Patriots, which is unlikely given their current belief
- *Update belief by conditioning on observed data*
  P(A|data) = P(data|A) P(A) / P(data)
  - So they updated their belief to increase the the Patriots's chance of a win:
    P(S wins | betting) = P(betting | S wins) P(S wins) / P(betting) = 0.50
- Even when the same data is observed, if people have different priors, they can end up with different posterior probability estimates P(A|data)

Bayesian vs. frequentist

- Bayesian central tenet:
  - Explicitly model all forms of uncertainty
  - E.g., Parameters, model structure, predictions
- Frequentist often model same uncertainty but in less-principled manner, e.g.,:
  - Parameters set by cross-validation
  - Model structure averaged in ensembles
  - Smoothing of predicted probabilities
- Although interpretation of probability is different, underlying calculus is the same

Probability distribution

- **Probability distribution** *(i.e., probability mass function or probability density function)* specifies the probability of observing every possible value of a random variable

- Discrete

  - Denotes probability that *X* will take on a particular value:

$$P(X = x)$$

- Continuous

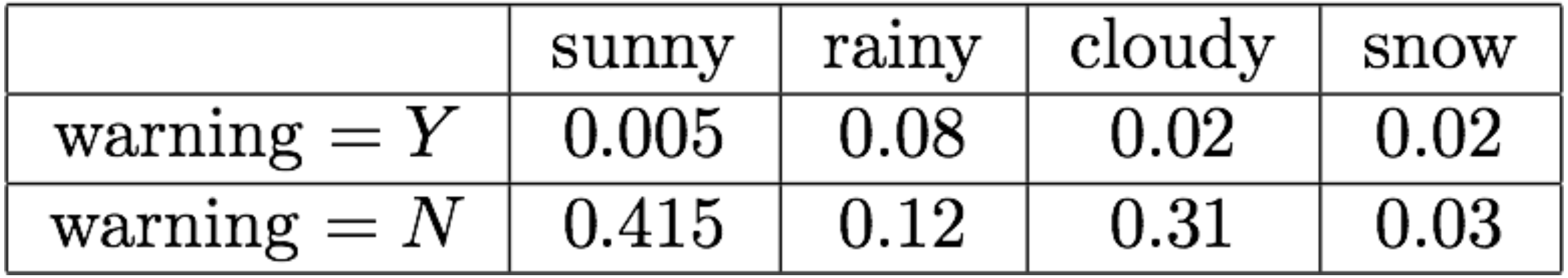
  - Probability of any particular point is 0, have to consider probability within an interval:

$$P(a < X < b) = \int_a^b p(x)dx$$

Joint probability

- **Joint probability distribution** for a set of random variables gives the probability of every atomic event on those random variables

  E.g., P(Weather, Warning) = a 4 × 2 matrix of values:



	sunny	rainy	cloudy	snow
warning $= Y$	0.005	0.08	0.02	0.02
warning $= N$	0.415	0.12	0.31	0.03

- Every question about events can be answered by the joint distribution

Conditional probability

- **Conditional** (or posterior) probability:
  - e.g., P(warning=Y | snow=T) = 0.4
  - Complete conditional distributions specify conditional probability for all possible combinations of a set of RVs:
    P(warning | snow) =
    {P(warning = Y | snow = T), P(warning = N | snow = T)},
    {P(warning = Y | snow = F), P(warning = N | snow = F)}
- If we know more, then we can update the probability by conditioning on more evidence
  - e.g., if Windy is also given then P(warning | snow, windy) = 0.5

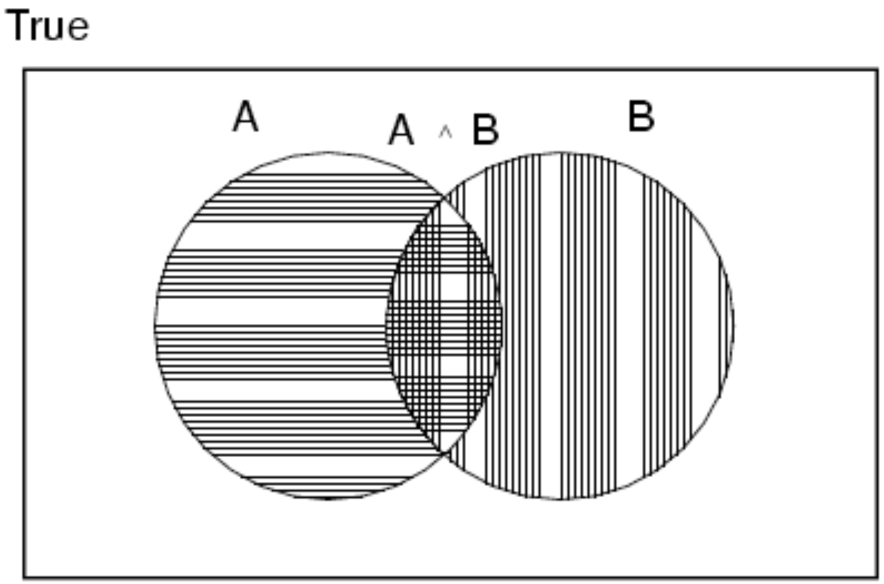
Conditional probability

- Definition of conditional probability:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \qquad \text{if } P(B) > 0$$

- **Product rule** gives an alternative formulation:

$$\begin{aligned} P(A \cap B) &= P(A|B)P(B) \\ &= P(B|A)P(A) \end{aligned}$$



- **Bayes rule** uses the product rule:

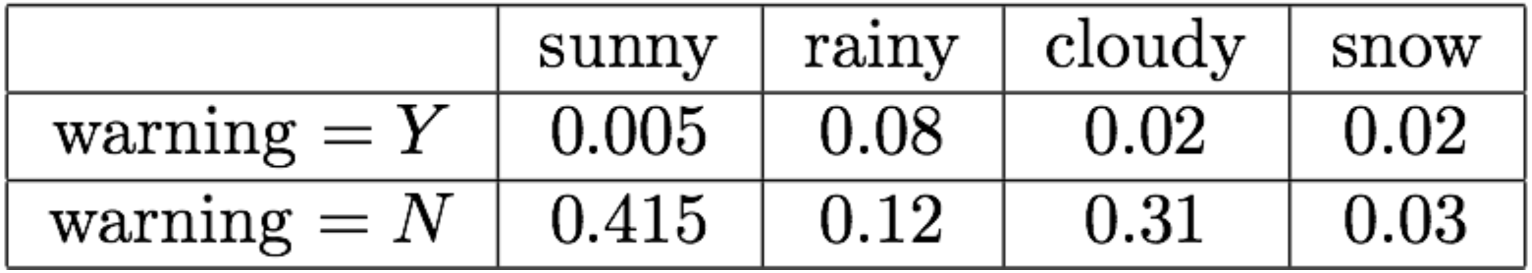
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Example

- Conditional probability:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \qquad \text{if } P(B) > 0$$

- Example: What is P(sunny | warning = Y)?



	sunny	rainy	cloudy	snow
warning $= Y$	0.005	0.08	0.02	0.02
warning $= N$	0.415	0.12	0.31	0.03

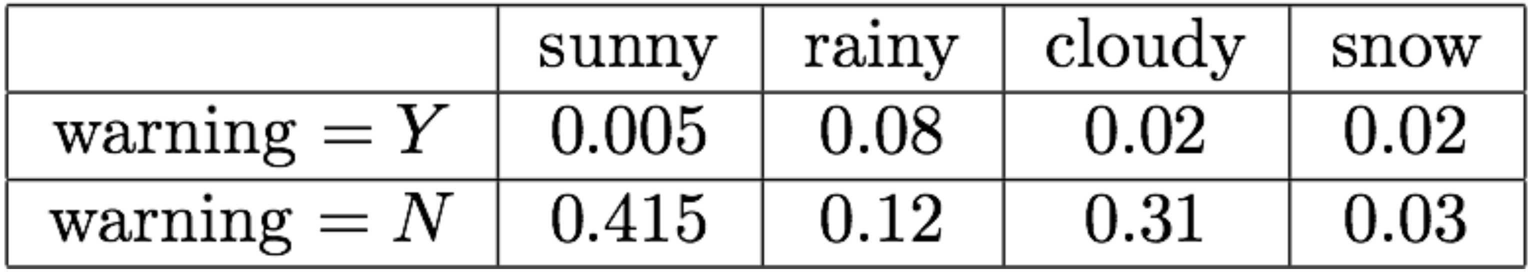
Conditional probability

- **Chain rule** is derived by successive application of product rule:

$$\begin{aligned}
P(X_1,\ldots,X_n) &= P(X_n|X_1,\ldots,X_{n-1})P(X_1,\ldots,X_{n-1}) \\
&= P(X_n|X_1,\ldots,X_{n-1})P(X_{n-1}|X_1,\ldots,X_{n-2})P(X_1,\ldots,X_{n-2}) \\
&= \ldots \\
&= \prod_{i=1}^{n} P(X_i|X_1,\ldots,X_{i-1})
\end{aligned}$$

Marginal probability

- **Marginal** (or unconditional) probability corresponds to belief that event will occur regardless of conditioning events
- Marginalization: $$P(A) = \sum_{b \in B} P(A, b)$$ $$= \sum_{b \in B} P(A|b)P(b)$$
- Example: What is P(cloudy)?



	sunny	rainy	cloudy	snow
warning $= Y$	0.005	0.08	0.02	0.02
warning $= N$	0.415	0.12	0.31	0.03

Independence

- A and B are independent iff:
    - P(A|B) = P(A) or P(B|A) = P(B) or P(A, B) = P(A) P(B)
    - *Knowing B tells you nothing about A*
- Examples
    - Coin flip 1 and coin flip 2?
    - Weather and storm warning?
    - Weather and coin flip=H?
    - Weather and election?

Conditional independence

- Two variables A and B are **conditionally** independent given Z iff for all values of A, B, Z:

$$P(A, B \mid Z) = P(A \mid Z)\, P(B \mid Z)$$

- *Note: independence does not imply conditional independence or vice versa*

Example 1

- **Conditional independence does not imply independence**
- Gender and lung cancer are not independent

  $P(C \mid G) \neq P(C)$

- Gender and lung cancer are conditionally independent given smoking

  $P(C \mid G, S) = P(C \mid S)$

- Why? Because gender indicates likelihood of smoking, and smoking causes cancer

Example 2

- **Independence does not imply conditional independence**

- Sprinkler-on and raining are independent
  $P(S \mid R) = P(S)$

- Sprinkler-on and raining are not conditionally independent given grass is wet
  $P(S \mid R, W) \neq P(S \mid R)$

- Why? Because once we know the grass is wet, if it’s not raining, then the explanation for the grass being wet has to be the sprinkler

Expectation

- Denotes the expected value or mean value of a random variable X
- Discrete

$$E[X] = \sum_x x \cdot p(x)$$

- Continuous

$$E[X] = \int_x x \cdot p(x) dx$$

- Expectation of a function

$$E[h(X)] = \sum_x h(x) \cdot p(x)$$

$$E[aX + b] = a \cdot E[X] + b$$

Example

- Let X be a random variable that represents the number of heads which appear when a fair coin is tossed three times.

- X = {0, 1, 2, 3}

- P(X=0) = 1/8; P(X=1) = 3/8; P(X=2) = 3/8; P(X=3) = 1/8

- What is the expected value of X, E[X]?

$$\begin{aligned} E[X] &= (0 \cdot \frac{1}{8}) + (1 \cdot \frac{3}{8}) + (2 \cdot \frac{3}{8}) + (3 \cdot \frac{1}{8}) \\ &= \frac{3}{2} \end{aligned}$$

Variance

- Denotes the squared deviation of X from its mean

- Variance

$$\begin{aligned} Var(X) &= E[(x - E[X])^2] \\ &= E[X^2] - (E[X])^2 \end{aligned}$$

$$\sigma = \sqrt{Var(X)}$$

- Standard deviation

$$Var(aX + b) = a^2 \cdot Var(X)$$

- Variance of a function

$$Var(h(X)) = \sum_x (h(x) - E[h(x)])^2 \cdot p(x)$$

Example

- Let X be a random variable that represents the number of heads which appear when a fair coin is tossed three times.

- X = {0, 1, 2, 3}

$$\begin{aligned} E[X] &= (0 \cdot \frac{1}{8}) + (1 \cdot \frac{3}{8}) + (2 \cdot \frac{3}{8}) + (3 \cdot \frac{1}{8}) \\ &= \frac{3}{2} \end{aligned}$$

- What is the variance of X, Var(X)?

$$\begin{aligned} Var(X) &= \left(\left[0 - \frac{3}{2}\right]^2 \cdot \frac{1}{8}\right) + \left(\left[1 - \frac{3}{2}\right]^2 \cdot \frac{3}{8}\right) + \left(\left[2 - \frac{3}{2}\right]^2 \cdot \frac{3}{8}\right) + \left(\left[3 - \frac{3}{2}\right]^2 \cdot \frac{1}{8}\right) \\ &= \left(\frac{9}{4} \cdot \frac{1}{8}\right) + \left(\frac{1}{4} \cdot \frac{3}{8}\right) + \left(\frac{1}{4} \cdot \frac{3}{8}\right) + \left(\frac{9}{4} \cdot \frac{1}{8}\right) \\ &= \frac{3}{4} \end{aligned}$$