Data Mining

Lecture 5

Ananya Jana
CS360

Fall 2024

MARSHALL

Important Characteristics of Data

- Dimensionality (number of attributes)
  - ◆ High dimensional data brings a number of challenges
- Sparsity
  - ◆ Only presence counts
- Resolution
  - ◆ Patterns depend on the scale
- Size
  - ◆ Type of analysis may depend on size of data

Random Variables and Expected Value

A **random variable** is a numerical quantity that is determined by the outcome of a random experiment. It is a function that assigns a real number to each possible outcome in the sample space of a probabilistic event.

$$X:S\rightarrow R$$

where R is the set of real numbers.

Discrete Example: In a roll of a fair die, the outcome can be any integer from 1 to 6. If we define X as the number on the die, X is a discrete random variable that can take values {1, 2, 3, 4, 5, 6}.

Continuous Example: If we measure the height of a randomly chosen person, the random variable X representing height could take any real value within a range (e.g., 150 cm to 200 cm). This makes X a continuous random variable.

Random Variables and Expected Value

Expected Value of a random variable :

For a **discrete random variable** X, with possible values $x_1,x_2,\ldots,x_n$ and corresponding probabilities $P(X=x_i)$, the expected value E(X) is calculated as:

$$E(X)=\sum_{i=1}^{n} x_i \cdot P(X=x_i)$$

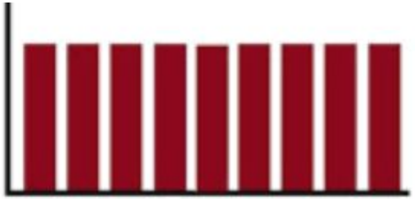
Consider a fair die. Let us roll the die and denote the outcome to be the random variable X. Then the expected value of the random variable X is

$$E(X) = (⅙).1 + (⅙).2 + (⅙).3 + (⅙).4 + (⅙).5 + (⅙).6 = 3.5$$

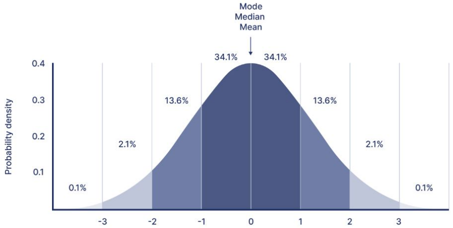
If the probabilities of all the outcomes are equal, then E(x) is the average value of the random variable, otherwise it is the weighted average.

Probability Distribution

A probability distribution is a mathematical description of the probabilities of events of the sample space S.



Uniform distribution



Normal Distribution

Probability distribution can be described for both discrete and continuous variables but in different ways. For discrete variables we generally use probability mass functions.

Probability distribution can give us a sense of how likely an outcome is relative to another outcome.

Information Based Measures

- Information theory is a well-developed and fundamental disciple with broad applications

- Some similarity measures are based on information theory
  - Mutual information in various versions
  - Maximal Information Coefficient (MIC) and related measures
  - General and can handle non-linear relationships
  - Can be complicated and time intensive to compute

Information and Probability

- Information relates to possible outcomes of an event
  - transmission of a message, flip of a coin, or measurement of a piece of data

- The more certain an outcome, the less information that it contains and vice-versa
  - For example, if a coin has two heads, then an outcome of heads provides no information
  - More quantitatively, the information is related the probability of an outcome
    - The smaller the probability of an outcome, the more information it provides and vice-versa
  - Entropy is the commonly used measure

Entropy

- For
  - a variable (event), X,
  - with n possible values (outcomes), $x_1, x_2 \ldots, x_n$
  - each outcome having probability, $p_1, p_2 \ldots, p_n$
  - the entropy of X , $H(X)$, is given by

$$H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

- Entropy is between 0 and $\log_2 n$ and is measured in bits
  - Thus, entropy is a measure of how many bits it takes to represent an observation of X on average

Entropy Examples

- For a coin with probability p of heads and probability $q = 1 - p$ of tails

$$H = -p \log_2 p - q \log_2 q$$

  - For p= 0.5, $q = 0.5$ (fair coin) $H = 1$
  - For $p = 1$ or $q = 1$, $H = 0$

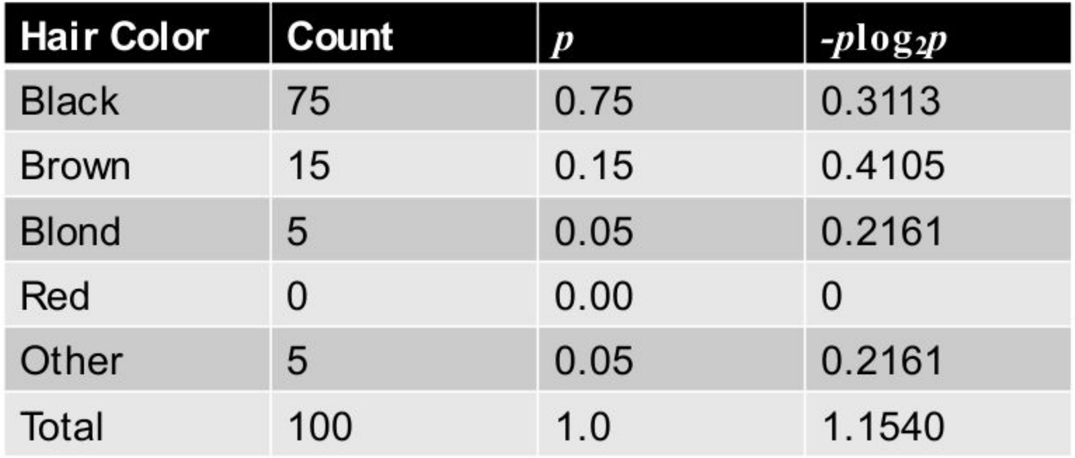
Entropy for Sample Data

- Suppose we have
  - a number of observations (m) of some attribute, X, e.g., the hair color of students in the class,
  - where there are n different possible values
  - And the number of observation in the i^{th} category is m_i
  - Then, for this sample

$$H(X) = -\sum_{i=1}^{n} \frac{m_i}{m} \log_2 \frac{m_i}{m}$$

- For continuous data, the calculation is harder

Entropy for Sample Data: Example



Hair Color	Count	p	$-p\log_2 p$
Black	75	0.75	0.3113
Brown	15	0.15	0.4105
Blond	5	0.05	0.2161
Red	0	0.00	0
Other	5	0.05	0.2161
Total	100	1.0	1.1540

Maximum entropy is $\log_2 5 = 2.3219$

Remember the method of defining probability using relative frequency

$$P(A) = \lim_{N \to \infty} \frac{\text{\# of occurrences of event } A}{N \text{ (total \# of trials)}}$$

Mutual Information

- Information one variable provides about another

Formally, $I(X,Y) = H(X) + H(Y) - H(X,Y)$, where

$H(X,Y)$ is the joint entropy of X and Y,

$$H(X,Y) = -\sum_{i}\sum_{j} p_{ij} \log_2 p_{ij}$$

Where p_{ij} is the probability that the ith value of X and the jth value of Y occur together

- For discrete variables, this is easy to compute

- Maximum mutual information for discrete variables is $\log_2(\min(n_X, n_Y))$, where n_X (n_Y) is the number of values of X (Y)

Mutual Information Example

Student Status	Count	p	$-p\log_2 p$
Undergrad	45	0.45	0.5184
Grad	55	0.55	0.4744
Total	100	1.00	0.9928

Grade	Count	p	$-p\log_2 p$
A	35	0.35	0.5301
B	50	0.50	0.5000
C	15	0.15	0.4105
Total	100	1.00	1.4406

Student Status	Grade	Count	p	$-p\log_2 p$
Undergrad	A	5	0.05	0.2161
Undergrad	B	30	0.30	0.5211
Undergrad	C	10	0.10	0.3322
Grad	A	30	0.30	0.5211
Grad	B	20	0.20	0.4644
Grad	C	5	0.05	0.2161
Total		100	1.00	2.2710

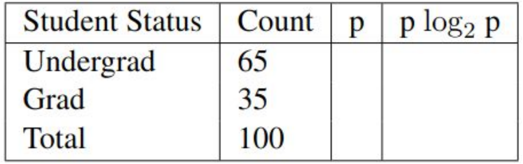
Mutual information of Student Status and Grade = 0.9928 + 1.4406 - 2.2710 = 0.1624

Tasks

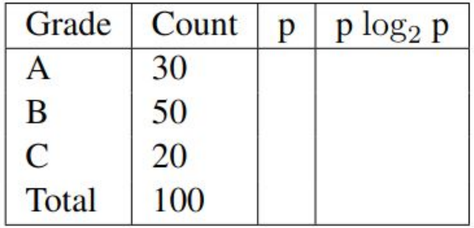
1. What is the entropy of a fair four sided die?

2. Can you show that the maximum amount value of entropy is $\log_2 n$ where n is the number of outcomes (Hint: Entropy is maximum when all the outcomes are equally likely)

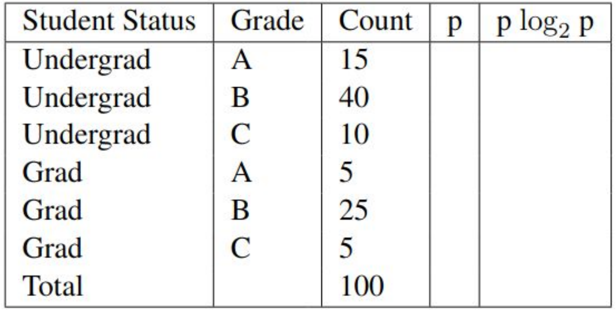
3. Calculate the mutual information of the student status and the grade from the tables below.



Student Status	Count	p	p $\log_2$ p
Undergrad	65		
Grad	35		
Total	100		



Grade	Count	p	p $\log_2$ p
A	30		
B	50		
C	20		
Total	100		



Student Status	Grade	Count	p	p $\log_2$ p
Undergrad	A	15		
Undergrad	B	40		
Undergrad	C	10		
Grad	A	5		
Grad	B	25		
Grad	C	5		
Total		100		