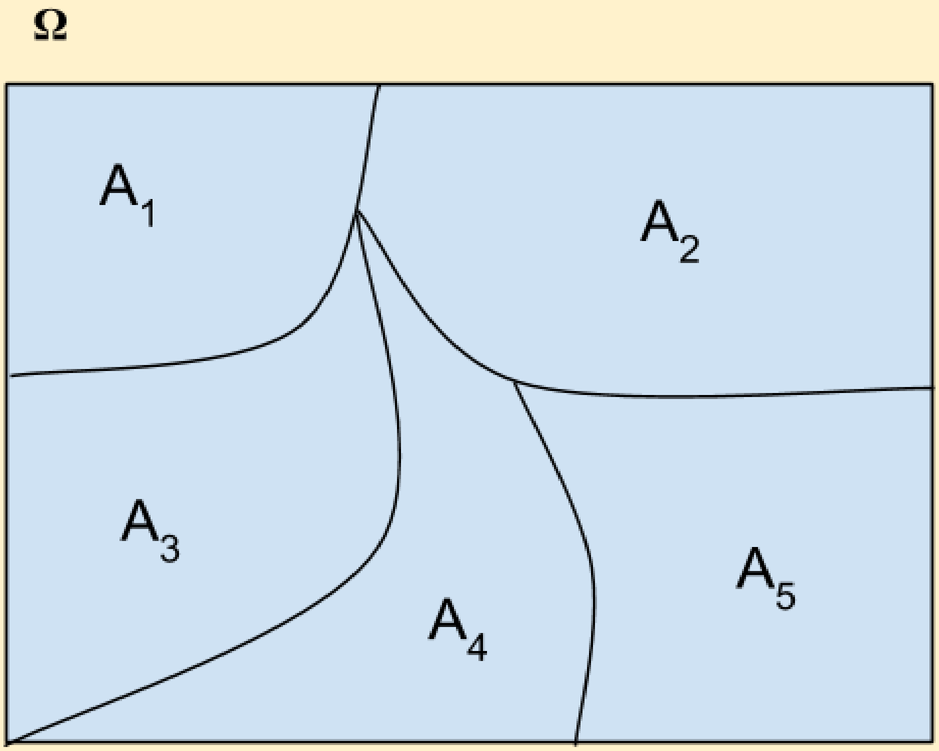
PadhAI: Probability Theory

One Fourth Labs

Basics of Probability Theory

What are the axioms of Probability

1. Consider the following sample space

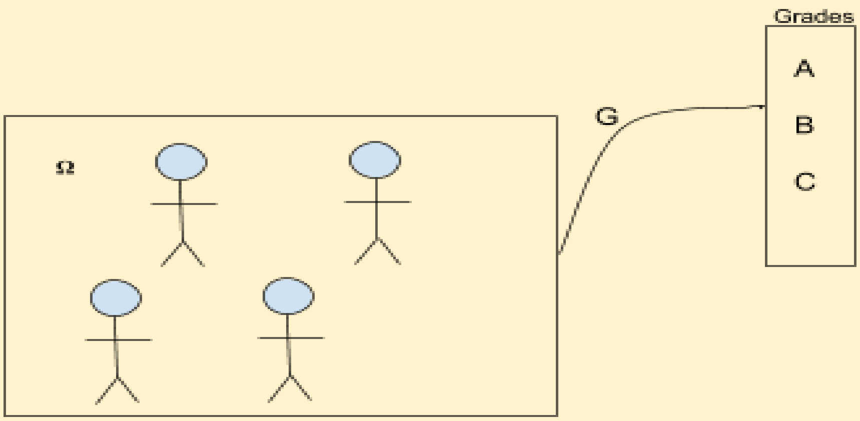


2. For any event A,
   a. 0 <= P(A) <= 1
3. If A_1, A_2,...A_n are disjoint events, ie $A_i \cap A_j = \phi \quad \forall (!i) = j$
   a. $P(\cup A_i) = \Sigma_i P(A_i)$
   b. The probability of the union of all the events is equal to the sum of the individual probabilities of those events
   c. $P(\cup A_i) = P(A_1) + P(A_2) + P(A_3) + P(A_4) + P(A_5)$
4. If $\mathbf{\Omega}$ is the universal set containing all the events, then
   a. $P(\Omega) = 1$

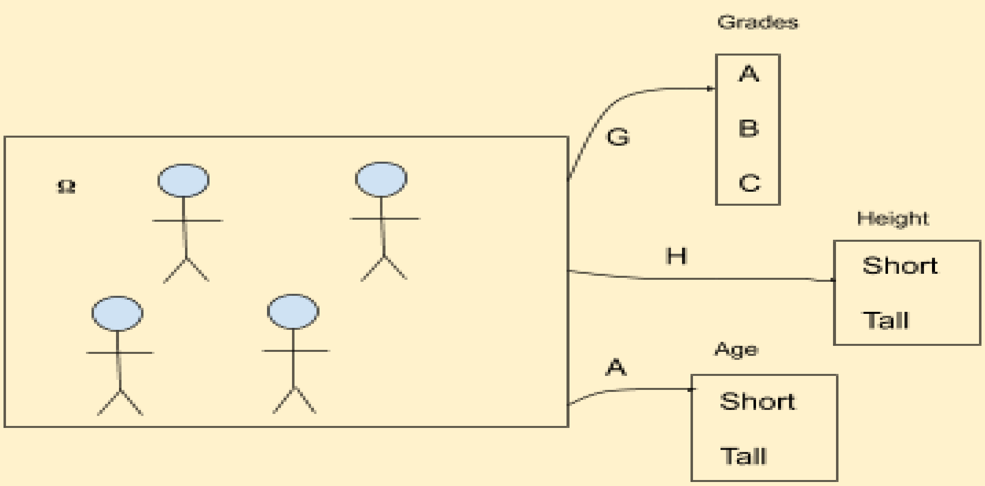
Random Variable Intuition

What is a Random Variable (intuition)

1. Suppose a student gets one of 3 possible grades in a course: A, B, C
2. One way of interpreting this is that there are 3 possible events here.
    a. For eg, to find P(A) we take $\frac{No.\ of\ students\ with\ A\ grade}{Total\ No.\ of\ students}$
3. Another way of looking at this is that there is a random variable G which maps each student to one of the 3 possible values



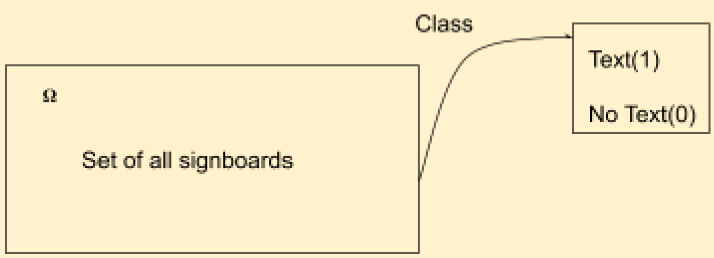
4. Here, the random variable G is treated more like a function that serves to map a student to a grade
5. And we are interested in $P(G = g)$ where $g \in \{A, B, C\}$
6. The benefit of this is that we can use multiple random variables on the same set to map to different outcomes



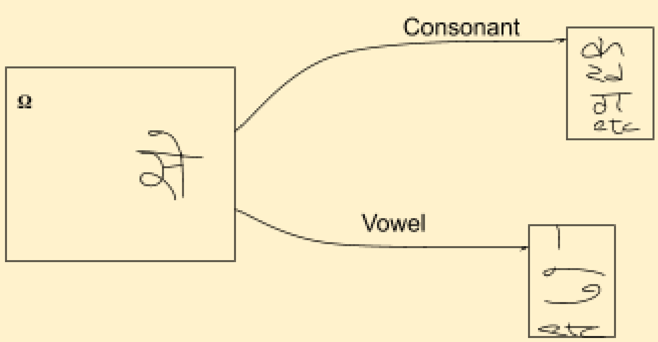
Random Variable Formal Definition

What is a random variable (formal definition)

1. A random variable is a function which maps each outcome in Ω to a value
2. In the previous example, G (or f_{grade}) maps each student in Ω to a value: A, B or C
3. The event Grade=A is a shorthand for the event
    a. $\{\omega \in \Omega : f_{grade} = A\}$
    b. In other words, All the elements such that when you apply f_{grade} the answer is A
    c. Grade is a random variable
    d. P(grade = A) = $\frac{\{\omega \in \Omega: f_{grade}=A\}}{Total\ number\ of\ students}$
    e. In the context of our example



4. This also applies to multiclass classification
    a. Mapping one Letter to its respecting vowel, and consonant.



5. Here, it would be P(Consonant=स) and P(Vowel = ौ)

Random Variable Continuous and Discrete

What are continuous and discrete random variables

1. A random variable can either take a continuous values/Real values (ie, weight, height)
2. Or discrete values(ie, Grade, Nationality)
3. For the scope of this course, we will mostly be dealing with discrete random variables. Ie, P(Vowels), P(Consonants) which all draw from a fixed set of discrete values

Probability Distribution

What is a marginal distribution?

1. Consider a random variable G for grades

G	P(G=g)
A	0.1
B	0.2
C	0.7

2. The above table represents the marginal distribution over G
   a. $(G = g) \quad \forall g \in A, B, C$
3. i.e. The probability of every possible value that the random variable can take (sums to 1)
4. We denote this marginal distribution compactly by P(G)

True and Predicted Distribution

What are true and predicted distributions

1. Consider the above example

G	P(G=g) (y)	(ŷ)
A	0.1	0.2
B	0.2	0.3
C	0.7	0.5

2. Here, y refers to the true distribution, or the actual probabilities for each value of G
3. And ŷ is the predicted distribution, or what we estimate the probabilities to be based on our observations
4. To measure the degree of correctness of our predictions, we can use a loss function.
5. However, Squared-error function might not be appropriate as it doesn't factor in some of the basic assumption of probability theory, ie P(G) >= 0 and <= 0, etc
6. So, we must select a different loss function that is more rooted in probability theory (Cross Entropy)

Certain Events

Events with 100% probability

1. We need something better than the squared error loss
2. Consider the scenario of a random variable X that maps to the winner in a tournament of 4 teams: A, B, C, D
3. We stop watching after the semi-finals, so we are unaware of the outcome, but in truth, team A has won, thus it is a certain event, with probabilities (P(A) = 1, P(B) = 0, P(C) = 0, P(D) = 0).

X	P(X=x) True distribution, unknown to us.	$\hat{Y}$ Predicted by us
A	1 (Certain event)	0.6
B	0	0.2
C	0	0.15
D	0	0.15

4. Before the tournament's completion, based on the point we have watched till(Semi-finals), we can predict the probabilities of each team's chance at victory (P(A) = 0.6, P(B) = 0.2, P(C) = 0.15, P(D) = 0.15)

PadhAI: Probability Theory

One Fourth Labs

Why do we care about Distributions

Let us put it into the context of our final project

1. Consider the signboard with the text '**Mumbai**'. Now our classifier is analysing the text character by character, and a random variable <u>char</u> maps the character to one of the 26 possible characters in the english language
2. For the first character **M,** we know the True distribution intuitively.

char	**Y = P(char=c)** **The certain event/True distribution**	**Ŷ** **Obtained from model**
a	0	0.01
b	0	0.01
...	0...	0.01...
m	1	0.7
...	...0...	...0.01...
z	...0	...0.01

3. We compute the difference between the True and Predicted distributions using squared-error loss or some other loss function. From this, it is clear why we use distributions in the scope of our learning.