PadhAI Week 3: Probability Theory & Information Theory

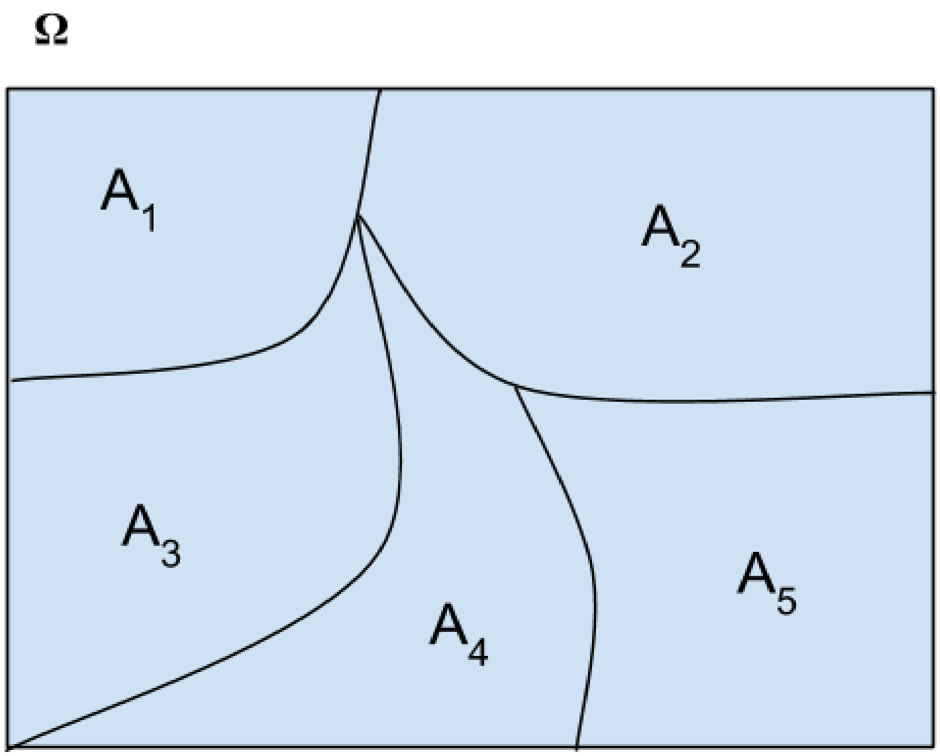
by Manick Vennimalai

3.4: Probability Theory

3.4.1: Basics of Probability Theory

What are the axioms of Probability

1. Consider the following sample space

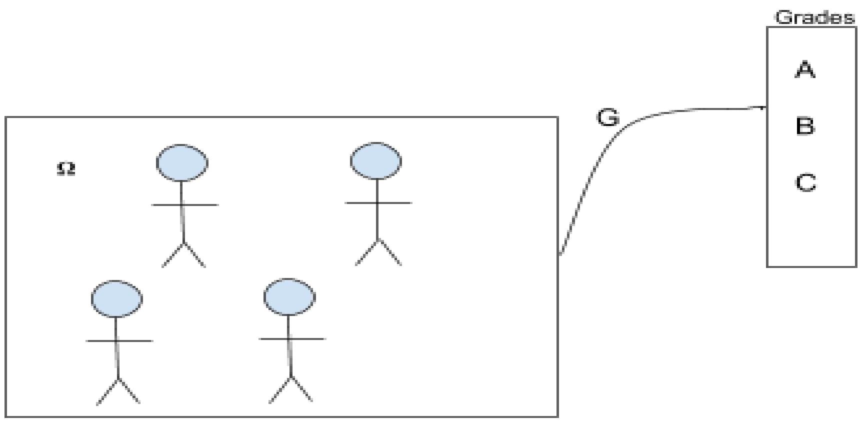


2. For any event A,
   a. 0 <= P(A) <= 1
3. If A_1, A_2,...A_n are disjoint events, ie $A_i \cap A_j = \phi \quad \forall (!i) = j$
   a. $P(\cup A_i) = \Sigma_i P(A_i)$
   b. The probability of the union of all the events is equal to the sum of the individual probabilities of those events
   c. $P(\cup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

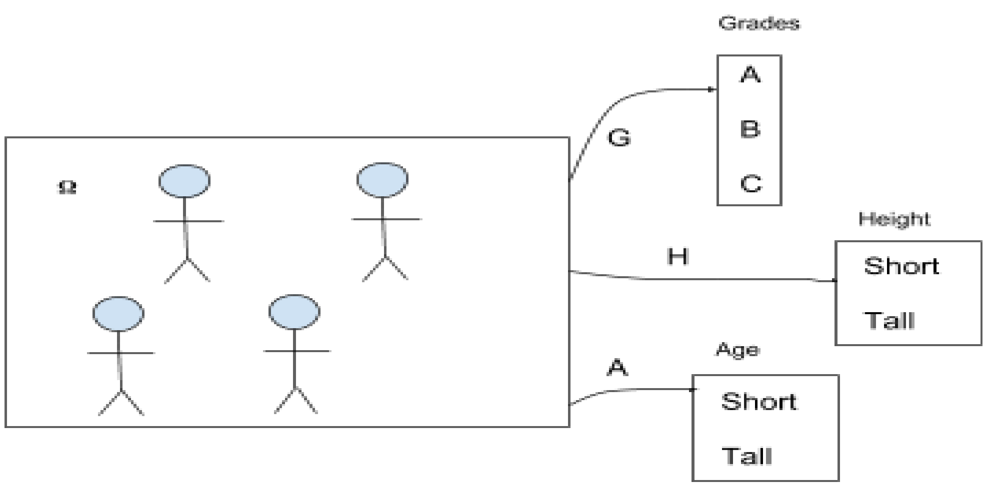
3.4.2: Random Variable Intuition

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



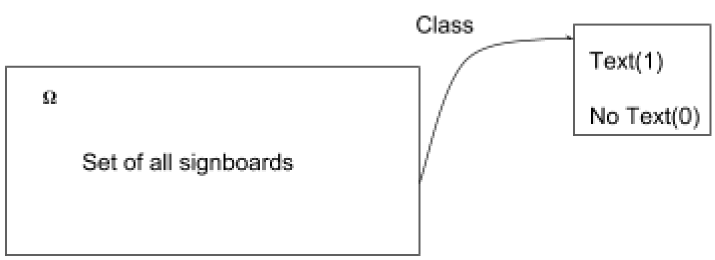
4. Here, the random variable G is treated more like a function that serves to map a student to a grade
5. And we are interested in $P(G = g)$ where $g \in \{A, B, C\}$
6. The benefit of this is that we can use multiple random variables on the same set to map to different outcomes



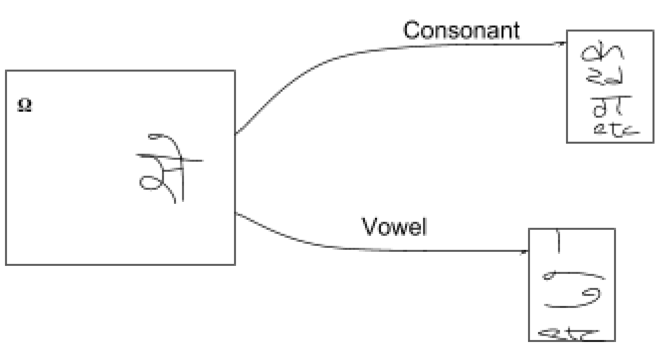
3.4.3: Random Variable Formal Definition

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 = ी)

3.4.4: 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

3.4.5: 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)

3.4.6: 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)

3.4.7: 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)

3.4.8: 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**	$\hat{Y}$ **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.

3.5: Information Theory

3.5.1: Expectation

What is the expectation of a distribution

1. Let us consider the random variable X that maps to the winning team amongst the 4 teams: A, B, C, D
2. P(X = x) represents the probability of team x winning where $x \in \{A, B, C, D\}$
3. Consider G(X=x), the gain associated with each of the teams if they win, where $x \in \{A, B, C, D\}$
4. Now, the expectation E(x) is given by $\Sigma_{i \in \{A,B,C,D\}} P(X = i) * G(X = i)$
5. Consider the following data

X	P(X = x)	G(X = x)
A	0.4	10000
B	0.2	2000
C	0.1	-8000
D	0.3	5000

6. Therefore, E(X) = $(0.4 * 10000) + (0.2 * 2000) + (0.1 * -8000) + (0.3 * 5000)$ = 5100

3.5.2: Information Content

What is Information content?

1. Consider the Random variable SR which maps to the direction in which the sun rises: East, West, North & South.
   a. Now, we are told that P(SR=East) is 1.
   b. Here, this is almost a blatantly obvious truth, thus we can say that the Information Gained here is very low.
2. Consider another Random variable ST, which maps to whether there is going to be a storm today: Yes, No.
   a. Now, we are told that P(ST=Yes) = 1
   b. Here, the information gained is very high as this is a rather surprising(low probability) event
   c. We can almost say that $Information\ Content \propto Surprise$
   d. Or in other words $Information\ Content \propto \frac{1}{P(X=Surprise)}$
   e. Thus, it can be inferred that the information content is a function of the probability of the event
   f. $IC(P(X = S))$ Where IC is information content
3. Now, consider two separate events
   a. X maps to which cricket team won the match: A, B, C, D
   b. Y maps to the state of a light switch: On, Off
   c. Now we are told that Team B won the match AND the light switch is On
   d. The total Information gained is $IC(X = B \cap Y = On) = IC(X = B) + IC(Y = On)$

4. Combining the points from above, we have
   a. $IC(P(X = S))$ (Information Content is a function of probability)
   b. $IC(P(X \cap Y)) = IC(P(X)) + IC(P(Y))$ (From the previous example)
   c. From probability theory, if P(X) and P(Y) are disjoint, then $(P(X \cap Y)) = P(X) . P(Y)$
   d. Therefore $IC(P(X).P(Y)) = IC(P(X)) + IC(P(Y))$
   e. Therefore we need a family of function that satisfy $f(a.b) = f(a) + f(b)$
   f. The log functions satisfy this $log(a.b) = log(a) + log(b)$
5. Now we can write the IC function as follows
   a. $IC(X = A) = log(\frac{1}{P(X=A)})$
   b. $IC(X = A) = log(1) - log(P(X = A))$
   c. $IC(X = A) = - log_2 P(X = A)$ (All the logs use base 2)

3.5.3: Entropy

What is Entropy

1. First, a quick recap of the concepts we've studied so far

Random Variable: X	**Probability Distribution: P(X=?)**	**Information Content: IC(X=?)**	**Expectation E(Gain)**
A	P(X=A)	$-log_2P(X=A)$	$\Sigma_{i \in \{A,B,C,D\}} P(X = i) * Gain(X = i)$
B	P(X=B)	$-log_2P(X=B)$	
C	P(X=C)	$-log_2P(X=C)$	
D	P(X=D)	$-log_2P(X=D)$	

2. Based on these four concepts, we can talk about Entropy
3. Entropy H(X) is the Expected Information Content of a Random Variable
4. $H(X) = -\Sigma_{i \in \{A,B,C,D\}} P(X = i) * log_2 P(X = i)$
5. Basically, substitute Gain for Information Content in the Expectation Equation

3.5.4: Relation to Number of Bits

Relation between number of bits and entropy

1. Consider the Entropy equation from the previous section using shorthand P_i for P(X=i)
2. $H(X) = -\Sigma_{i \in \{A,B,C,D\}} P_i * log\ P_i$
3. Suppose there is a message X that you want to transfer that can take 4 values: A, B, C, D

4. For 4 values, we would use 2 Bits to transfer each message

Random Variable: X	2 Bit version	Probability Distribution: P(X=?)	Information Content: IC(X=?)
A	00	1/4	$-\log_2 2^2 = 2$ (ie $\log_a a^n = n$)
B	01	1/4	$-\log_2 2^2 = 2$
C	10	1/4	$-\log_2 2^2 = 2$
D	11	1/4	$-\log_2 2^2 = 2$

5. Now we can make the connection that the number of bits required to transfer a message is equal to the information content of that message
6. Consider another message X with 8 values: A, B, C, D, E, F, G, H

Random Variable: X	3 Bit version	Probability Distribution: P(X=?)	Information Content: IC(X=?)
A	000	1/8	$-\log_2 2^3 = 3$ (ie $\log_a a^n = n$)
B	001	1/8	$-\log_2 2^3 = 3$
C	010	1/8	$-\log_2 2^3 = 3$
D	100	1/8	$-\log_2 2^3 = 3$
E	011	1/8	$-\log_2 2^3 = 3$
F	101	1/8	$-\log_2 2^3 = 3$
G	110	1/8	$-\log_2 2^3 = 3$
H	111	1/8	$-\log_2 2^3 = 3$

7. While sending a continuous stream of messages, we would be interested in minimizing the stream of bits that we send
8. Consider the same 4 valued example but with a different distribution

Random Variable: X	Probability Distribution: P(X=?)	Information Content: IC(X=?)
A	1/2 (High prob)	$-\log_2 2^1 = 1$ (ie $\log_a a^n = n$)
B	1/4 (Medium prob)	$-\log_2 2^2 = 2$
C	1/8 (Low prob)	$-\log_2 2^3 = 3$
D	1/8 (Low prob)	$-\log_2 2^3 = 3$

9. This situation is considered favourable only if the average number of bits is less that the value it takes for an equally distributed set of values
10. The average is calculated using Entropy $H(X) = -\Sigma_{i \in \{A,B,C,D\}} P_i * log\ P_i$
11. Average/Entropy $= \frac{1}{2}(1) + \frac{1}{4}(2) + \frac{1}{8}(3) + \frac{1}{8}(3) = 1.75\ which\ is < 2$
12. Thus, the Entropy gives us the ideal number of bits that should be used to transmit the message

3.5.5: KL- Divergence and Cross Entropy

How we deal with true and predicted distributions

1. Consider the following data:

X	True Distribution: y	True IC(X)	Predicted Distribution: y	Predicted IC(X)
A	y_1	$-\log y_1$	$\hat{y}_1$	$-\log\hat{y}_1$
B	y_2	$-\log y_2$	$\hat{y}_2$	$-\log\hat{y}_2$
C	y_3	$-\log y_3$	$\hat{y}_3$	$-\log\hat{y}_3$
D	y_4	$-\log y_4$	$\hat{y}_4$	$-\log\hat{y}_4$

2. Initially, we do not know the values of the True distribution and thereby the True Information Content
3. Hence, we generate a Predicted distribution and use that to compute the predicted information content.
4. But, the actual message will come from the True distribution y.
5. So therefore, the No. of bits will **not be** $-\Sigma\hat{y}_i\, log\hat{y}_i$ but **instead** $-\Sigma y_i\, log\hat{y}_i$
6. This is because the value associated with each of these messages comes from the predicted distribution $-log\hat{y}_i$ but the messages themselves comes from the True distribution y
7. Now, we have formed to the basis to talk about KL-Divergence:
   a. $H_y = -\Sigma y_i\, log y_i$ is called the entropy
   b. $H_{y,\hat{y}} = -\Sigma y_i\, log\hat{y}_i$ is called the cross entropy
   c. Now we want to find the difference/distance between the predicted case and the true case, using something more efficient than the squared error
   d. So y||ŷ = $H_{y,\hat{y}} - H_y$
   e. y||ŷ = $-\Sigma y_i\, log\hat{y}_i + \Sigma y_i\, log y_i$
   f. This is called the KL-Divergence
8. Thus, we now have **KLD(y||ŷ) =** $-\Sigma y_i\, log\hat{y}_i + \Sigma y_i\, log y_i$
9. Now, we have a way of computing the difference between two distributions.

Fin