Mathematics for Machine Learning

Lecture 8
(13.06.2024)

Probability Distributions

Mahammad Namazov

Table of contents

Introduction

Randomness

Foundation

Mathematics

Randomness

- Basically, all our experiments till now;
- In Probability Theory, we specify the randomness as a nature of the variables;
- Random variable is a quantitative variable which values are related with chance;
- We will classify them into two main groups:
  - Discrete and Continuous Random Variables
- *Range – Domain* relations will help us to define distribution functions in the similar manner;

Foundation

- When we have multiple events (i.e., more than 3):
  - Generalization;
  - Classification;
  - Visualization;
  - Inference;

- Remember the mapping principle of functions?

$$f: A \to B$$

  - A: ***Domain*** (i.e., range of variable)
  - B: ***Co-domain*** (i.e., probabilities)

- Random variable X is:
  - ***Discrete:*** when the range is countable (e.g., $\{x_1, x_2, x_3, \dots, x_n\}$)
  - ***Continuous:*** when the range is uncountable (i.e., any interval (e.g., $[a, b], where\ a < b$))

Mathematics

- We have a set: $\{a, b, c\}$
  - Size of the set is 3 (n) in this case;
  - All combinations that:
  - Includes 3 elements;
  - Includes each element:

$$\{\{a,b,c\},\{a,c,b\},\{b,a,c\},\{b,c,a\},\{c,a,b\},\{c,b,a\}\}$$

  - Number of elements is simply in the simple set can be found as below:

$$n! = 3! = 3 * 2 * 1 = 6$$

- What if I want to have specific number of elements in each combination:
  - E.g., I want binary combinations (2 elements) of these values:

$$P(n,k) = \frac{n!}{(n-k)!} = \frac{3!}{1!} = \frac{6}{1} = 6$$

This is called, Permutation of n elements for given k elements' arrangement (order matters!)

Mathematics

- What if I do not care about the order?

- What I need to know the possible combination of:
  - Specific number of elements from the set;
  - I don't want to focus on arrangement of such (a, b) or (b, a)

- We have the set as before: $\{a, b, c\}$
  - We want to check binary combinations again, but don't care about the order:
  - We have all binary combinations as P(3, 2) = 6
  - We have all binary possibilities as 2! = 2
  - Then we can get our scenarios:

$$C(n,k) = C_k^n = \binom{n}{k} = \frac{P(n,k)}{k!} = \frac{n!}{(n-k!)k!} = \frac{6}{2} = 3$$

Univariate Probability Distributions

PMF

PDF

CDF

Some distributions

Investigate a scenario

- Bernoulli Trial: Suppose you have a single event, which has two possible mutually exclusive outcomes of this event: Success and Failure;

- When probability of success is p, then probability of failure will be 1-p;

- Now let's introduce a problem to apply our knowledge to:
  - Chat GPT says that %68 of Canadians own home. We ask 2 random people (blind questionnaire). How would be the probability distribution of our questionnaire?
  - Questions to be asked:
    - What are possible scenarios?
    - How can we formulate this problem in terms of one random variable?
    - How to find probability distribution?

Solve the problem

- What is success, what is failure?
- What are possible scenarios in our problem?
- Samples are chosen randomly, and they are independent: Bernoulli Trial
- Once you consider ”having home” as a success, then you will have the following table as a distribution:

Scenario	SS	SF	FS	FF
Probability	0.68 x 0.68	0.68 x 0.32	0.32 x 0.68	0.32 x 0.32

- Is it a valid distribution?

Probability Mass Function (PMF)

- We use it when the available data is discrete;

- For any discrete random variable (X), we use *Probability Mass Function* (Why Mass?)
  - We deal with corresponding ***probability*** for each value: $x \in X$;
  - What part of the main ***mass*** does each x convey?;
  - Using the ***function***, we map values from the range into probabilities;

- PMF $P(X)$ must be defined in the range of X;
  - $P(X = x)$, for any $x \in X$;

- Properties for PMF to be held:
  - $0 \le P(X = x) \le 1$ for any $x \in X$;
  - $\sum_{x \in X} P(X) = 1$

Probability Mass Function (PMF)

- We use it when the available data is discrete;

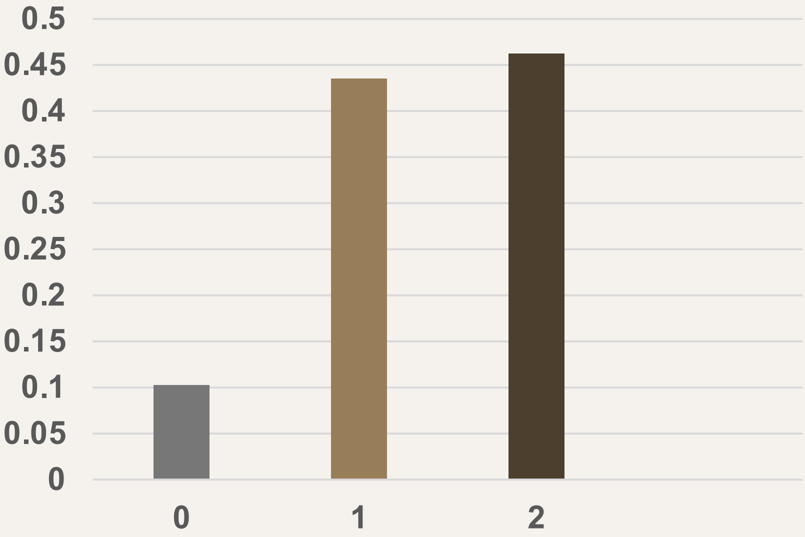
- For any discrete random variable (X), we use *Probability Mass Function* (Why Mass?)
  - We deal with corresponding ***probability*** for each value: $x \in X$;
  - What part of the main ***mass*** does each x convey?;
  - Using the ***function***, we map values from the range into probabilities;

- PMF $P(X)$ must be defined in the range of X;
  - $P(X = x)$, for any $x \in X$;

- Properties for PMF to be held:
  - $0 \leq P(X = x) \leq 1$ for any $x \in X$;
  - $\sum_{x \in X} P(X) = 1$

Was it valid?

- Then Probability Distribution for success count on the event that people own home:
  - Horizontally: $x \in X = \{0, 1, 2\}$
  - Vertically: $P(X = x)$



Continuous Domain

- What if our variable's range is continuous?
- In this case we will work with intervals, but not with single points in the domain;
- Once we talk about continuous domain, distribution will be shown with PDF. Why P, D, F?
  - We deal with corresponding ***probability*** for given range: $[x_i, x_k] \in X$;
  - How ***dense*** is distribution in the given range?;
  - Using the ***function***, we map values from the specific range in the domain into probabilities;
- Some details we need to know:
  - Variable is defined by range, not set of discrete values;
  - P(X=x) = 0, always;
  - Boundaries do not matter in terms of probability;

Probability Density Function (PDF)

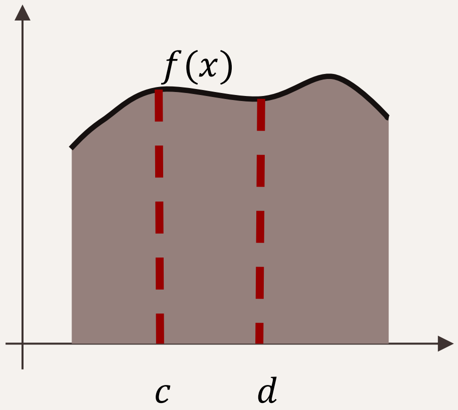
- Suppose we obtain the following PDF $f(x)$
- $f(x)$ represents the height of the curve at x;
- $f(x) = 0$, when $x < a$ and $x > b$;
- For specific region of $[c, d] \in [a, b]$, PDF will be:

$$P(c \leq x \leq d) = \int_{x=c}^{x=d} f(x)dx$$

- Including or excluding boundaries does not change the PDF value:

$$(c, d) \equiv [c, d] \equiv (c, d] \equiv [c, d)$$

  - *It is applicable for PDF! Why?*



Cumulative Distribution Function (CDF)

- Applicable in both domains: Discrete and Continuous;
- For simplicity let’s remember the Canada scenario (to have home or not to have):
  - Same initial conditions remain, but we ask 20 people instead;
  - We want to solve another problem: What is the probability of at most x people has home:

$$F(X = x) = F(x) = \sum_{X \leq x} P(X = x)$$

- CDFs are non-decreasing functions and the upper bound for CDFs is 1;
- In continuous domain, the idea is same but computation is slightly different:

$$F(X = x) = F(x) = \int_{-\infty}^{x} f(x)dx$$

- What if we ask not at most scenario but at least?

Specific Parameters

Expected Value

Variance

Standard Deviation

Expected Value (i.e., mean)

- Specifies the center of the distribution (i.e., mean)
- Why do we need to know, where is the center of our distribution?
- For a discrete random variable X, expected value $E(X)$ is computed as:

$$E(X) = \begin{cases} \sum_{x \in X} xP(X = x) \\ \int_{x \in X} xp(x)dx \end{cases} = \mu$$

  - It is literally weighted average of values which are weighted with corresponding probabilities

- Now think that there is a function which changes values with respect to this variable (e.g., $E(X)$). Expected value of this function will be **(same applies for continuous as well)**:

$$E(g(X)) = \sum_{x \in X} g(x)p(x) = \sum_{x \in X} g(x)P(X = x)$$

Variance & Standard Deviation

- Variance simply answers the following question:
  - How does the random variable X vary with respect to the mean?

- We can formulate it as following:
  - The expected value of the squared distance of the variable from the mean:

$$\sigma^2 = E[(X-\mu)^2] = \begin{cases} \sum_{x \in X} (x-\mu)^2 p(x) \\ \int_{x \in X} (x-\mu)^2 p(x) dx \end{cases}$$

  - If you don't waste your time, then use this one (if you have relevant information):

$$\sigma^2 = E(X^2) - [E(X)]^2$$

- Once you have variance, you have also standard deviation σ;

Joint Probability Distributions

Sum Rule

Product Rule

Bayes Theorem

Scenario Investigation

- Let's have 2 variables for a person where each specifies different characteristics of a person:
  - X = {0, 1} specifies whether a person is a fan of football;
  - Y = {0, 1} specifies whether a person watches games in the stadium;

	Y	
X	$P(X = 0, Y = 0)$	$P(X = 0, Y = 1)$
	$P(X = 1, Y = 0)$	$P(X = 1, Y = 1)$

- Notice that:

$$\sum_{x_i \in \{0,1\}} \sum_{y_i \in \{0,1\}} P\left(X = x_i, Y = y_j\right) = P(0,0) + P(0,1) + P(1,0) + P(1,1) = 1$$

Sum Rule (Marginalization)

- All distributions must be interconnected with each other (i.e., Marginal probability of a single variable cannot be computed without considering others);

- For specific scenario of $X = x_i$:

$$P(X = x_i) = \sum_{y_j \in Y} P\left(X = x_i, Y = y_j\right)$$

	Y	
$X = 0$	$P(X = 0, Y = 0)$	$P(X = 0, Y = 1)$
$X = 1$	$P(X = 1, Y = 0)$	$P(X = 1, Y = 1)$

- **Warning: It is an irreversible act!**

Product Rule and Conditionality

- Joint probability of 2 variables can be given by their relations:

$$p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{x}|\boldsymbol{y})p(\boldsymbol{y}) = p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})$$

- In ML and Bayesian statistics, we want to infer about the value of an unknown variable, based on the given information: **prior, likelihood and evidence:**

$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{y})p(\boldsymbol{y})}{p(\boldsymbol{x})}$$

- **Notice:** Conditioning of random variable to the fixed value of another random variable, provides us another probability distribution (How?)

- From conditional probability to Marginality:
  - **Marginal Probability of X for specific x_i is a weighted average of all y_j values in Y, where weights are $P(X = x_i | Y = y_j)$ for each y_j**

Covariance / Correlation

Covariance

Correlation

More than 2

Introduction

- Now that we deal with 2 and more variables, we need to analyze their relations;
- Covariance: How vary these variables together?
- Correlation: How strong and in what direction is the relationship between these variables?
- Road Map:
  - Introduce a problem, where we can build notion together;
  - Build step by step;
  - Know how to compute these parameters

Problem Definition

- To see steps granularly, we will use discrete domain;
- Assume you have 2 variables $X = \{0, 1, 2\}, Y = \{0, 1, 2, 3\}$ with the following Joint Probability Distribution Table:

X \ Y	0	1	2	3
0				
1				
2				

Details about $E(X, Y)$:
Magnitude: How strong?
Sign: Increase or Decrease?

- Ingredients:
  - Expected values of variables: $E(X), E(Y)$
  - Expected value of variables together to quantify variation of variables together: $E(X, Y)$

Step 0: Expected values

- Expected value for any variable Z can be computed as following:

$$E(Z) = \sum_{i=1}^{n} z_i P(Z = z_i) = \sum_{i=1}^{n} z_i p(z_i)$$

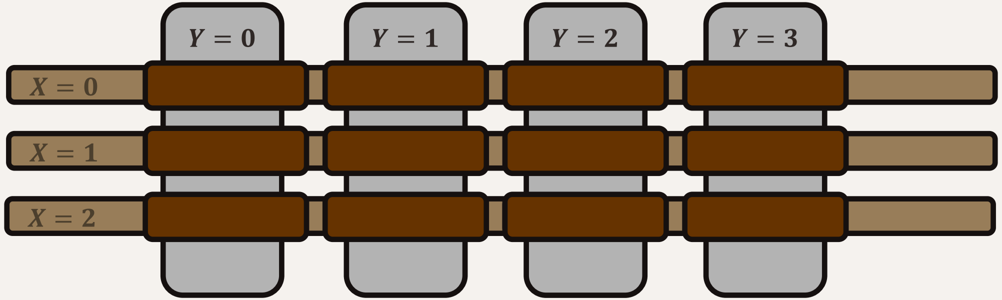
- We deal with joint probability, thus for marginality:

$$P(Z = z_i) = \sum_{j=1}^{m} P\left(Z = z_i, K = k_j\right)$$

- Note: Z and K are used for generic representations;

Step 1: Marginal Probabilities

- Sum all values of the variable with respect to specific value of the asked variable:



Step 2: Compute Expected Values

- Now we have $E(X)$ and $E(Y)$:

$$E(X) = \sum_{i=0}^{2} x_i p(x_i) =$$

$$E(Y) = \sum_{j=0}^{3} y_j p(y_j) =$$

- Now we need to compute $E(XY)$:

$$E(XY) = \sum_{i=1}^{3} \sum_{j=1}^{4} x_i y_j p(x_i, y_j)$$

  - What is $p(x_i\ y_j)$ and how to compute it?

Step 3: E(X,Y)

- Then we have these Marginal Probabilities:

XY	0	1	2	3	4	6
$P(XY)$						
$XYp(XY)$						

- Now using these probabilities in the recent equation:

$$E(XY) = \sum_{i=1}^{3} \sum_{j=1}^{4} x_i y_j p(x_i y_j) =$$

Step 4: Covariance Cov(X,Y)

- Idea:
  - How those two variables vary with respect to each other;
  - There are two methods to compute the covariance between two variables:

- Using probabilities and mean values:
  - Find how distant the specific $x_i \in X$ from the variable's mean:
$$x_i - \bar{x} = x_i - E(X)$$
  - Find the same for specific $y_j \in Y$:
$$y_j - \bar{y} = y_j - E(Y)$$
  - Multiply them together and weight the result with probability. For x_i and y_j:
$$p(x_i, y_j)(x_i - E(X))(y_j - E(Y))$$
  - Sum them all:
$$Cov(X,Y) = \sum_{x_i \in X} \sum_{y_j \in Y} p(x_i, y_j)(x_i - E(X))(y_j - E(Y)) = E(X)E(Y) - E(X)E(Y)$$
  - Or:
$$Cov(X,Y) = \frac{\sum_i (x_i - E(X))(y_i - E(Y))}{n-1}$$

Warning: Don't use the second!

Step 5: Correlation Corr(X,Y)

- To see the strength of the relation, we need to use correlation;
- It is simply normalizing covariance with variances of each variable;
- Mathematically:

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

- Correlation of variables can lie within the following range:

$$Corr(X,Y) \in [-1,1]$$

- If two variables:
  - Increase or decrease together, then $Corr(X,Y) \in (0,1]$;
  - One increase and the other decrease, then $Corr(X,Y) \in [-1,0)$
  - There is not any trend: $Corr(X,Y) = 0$

Conclusion

Summary

Takeaways

References

References

- Further information to read:
  - Deisenroth, M. P., Faisal, A. A., & Ong, C. S. (2020). *Mathematics for machine learning*. Cambridge University Press.
    - Chapter 6, Section 1 only
    - Author's suggestion: Chapter 2 from (Walpole et al., 2011)
  - https://www.probabilitycourse.com
    - Chapter 1

Summary

- To see specific scenario's possibility, we can use PMF in this case;
- To see several scenarios' possibilities as a distribution, we can use CDF;
- Expected value, Mean, Standard Deviation and Mode are specific parameters in Probability Theory that can speak for data
- In PDF we cannot get specific value's probability but for interval;
- Since we can compute it for specific range, then it can be seen as likelihood of events' density in the given range;
- Joint distributions are utilized to analyze common relations of several variables;
- This will help us to analyze and read relations among parameters (or features of our data)
- Covariance is an indicator to show whether there is a trend or not;
- Correlation measures how strong is the trend (if exists)

Takeaways

- Probability distribution provides us distribution of possibilities of several outcomes of an event;
- Joint probability distribution tells us how two such events collaborate;
- Specific parameters enables us to understand what data want to tell us;
- Covariance shows how 2 variable change together;
- This change’s strength and direction is determined by correlation;

References

- Further information to read:
  - Deisenroth, M. P., Faisal, A. A., & Ong, C. S. (2020). *Mathematics for machine learning*. Cambridge University Press.
    - Chapter 6, Sections 6.6 and 6.7 are optional

- This channel provides significant information on Probability Distribution:
  https://www.youtube.com/watch?v=oHcrna8Fk18&list=PLvxOuBpazmsNIHP5cz37oOPZx0JKyNszN&ab_channel=jbstatistics

The End

Thanks for your attention!

Mahammad Namazov