

CPSC 532P / LING 530A: Deep Learning for Natural Language Processing (DL-NLP)

Muhammad Abdul-Mageed

muhammad.mageed@ubc.ca

Natural Language Processing Lab

The University of British Columbia

Table of Contents

1 Probability

- Probability
- Random Variables
- PMS & PDS
- Marginal Probability
- Conditional Probability
- The Chain Rule
- Independence
- Expectation
- Variance & Covariance

2 Distributions

Many of the current slides are a summary Chapter 3 in Goodfellow et al. (2016). More information can be found therein. Note: The authors credit Pearl (1988) for a lot of the content of the chapter. Other sources used here are credited where appropriate.

Why Probability?

- Nearly all activities require some ability to reason in the presence of **uncertainty**.
- There are three possible sources of uncertainty:

Three Possible Sources of Uncertainty

- **Inherent stochasticity in the system** being modeled. For example, most interpretations of quantum mechanics describe the dynamics of subatomic particles as being probabilistic.
- **Incomplete observability**: When we cannot observe all of the variables that drive the behavior of a system
- **Incomplete modeling**: When we use a model that must discard some of the information we have observed, the discarded information results in uncertainty in the model's predictions.

- Probability can be seen as the extension of **logic** to deal with uncertainty.
- Logic provides a set of formal rules for determining what propositions are implied to be **true** or **false** given the assumption that some other set of propositions is true or false.
- Probability theory provides a set of formal rules for determining the **likelihood** of a proposition being true given the likelihood of other propositions.

A Random Variable

- A **random variable** is a variable that can take on different values randomly.
- E.g., both x_1 and x_2 are possible values that the random variable x can take on.
- **Vector-valued variables:** We write the random variable as \mathbf{x} (**bolded**) and one of its values as \mathbf{x} (*italicized*).
- On its own, a **random variable** is just a description of the states that are possible; it must be coupled with a **probability distribution** that specifies how likely each of these states are.

A Discrete Random Variable

- A **Discrete random variable** is one that has a finite or countably infinite/distinct/separate number of states (e.g., 1, 2, 3, 4,5).
- Note: **these states are not necessarily the integers**
- They can also just be named states (e.g., "head", "tail") that are not considered to have any numerical value.

A Continuous Random Variable

- A **continuous random variable** is associated with a real value.
- The data can take infinitely many values (e.g., height of a tree).
- Continuous random variables describe outcomes in probabilistic situations where the possible values some quantity can take form a **continuum**, which is often (but not always) the entire set of real numbers \mathbb{R} . . .
- They are a **generalization of discrete random variables to uncountably infinite sets of possible outcomes**. [link].

PMF

- A probability distribution over discrete variables may be described using a **probability mass function (PMF)**.
- The PMF maps from a state of a random variable to the probability of that random variable taking on that state.
- Suppose we want the **probability of it raining in Vancouver in July**.
- We will call this probability x .
- We have two states: x_1 (rain) and x_2 (no_rain).
- We would say $P(x_1)=0.3$ (or 30%).
- And $P(x_2)=0.7$ (or 70%).

Discrete Variables and Probability Mass Functions II

- The probability that $x = x$ is denoted as $P(x)$, with a probability of 1 indicating that $x = x$ is certain and a probability of 0 indicating that $x = x$ is impossible.
- Probability mass functions can act on many variables at the same time (**joint probability distribution**).
- $P(x = x, y = y)$ denotes the probability that $x = x$ and $y = y$ simultaneously.
- We may also write $P(x, y)$ for brevity.

1: Properties of PMF P

- The domain of P must be the set of all possible states of x .

$$\forall x \in \mathcal{X}, 0 \leq P(x) \leq 1$$

$$\sum_{x \in \mathcal{X}} P(x) = 1$$

Probability Density Function

- For continuous random variables, we describe probability distributions using a **probability density function (PDF)**, which must satisfy the following:

2: Properties of PDF I

- The domain of p must be the set of all possible states of x .

$$\forall x \in x, P(x) \geq 0$$

- **Note:** We do not require

$$P(x) \leq 1$$

- And:

$$\int p(x) dx = 1$$

Probability Density Function II

PDF

- A probability density function $p(x)$ does not give the probability of a specific state directly, instead the **probability of landing inside an infinitesimal region with volume δx** (read: "delta x") is given by $p(x)\delta x$.
- We can **integrate the density function to find the actual probability mass of a set of points**.
- Specifically, the probability that x lies in some set \mathbb{S} is given by the **integral of $p(x)$ over that set**.
- In the univariate example, the **probability that x lies in the interval $[a, b]$** is given by $\int_{[a,b]} p(x)dx$.

Joint Probability

X	Y		
	Y=0	Y=1	Y=2
X=0	0.1	0.2	0.3
X=1	0.3	0.4	0.2

3: Joint Probability

What is $P(x = 0, y \leq 1)$?

$$P_{x,y} = P_{x,y}(0,0) + P_{x,y}(0,1)$$

Marginal Probability

- Sometimes we know the probability distribution over a set of variables and we want to know the **probability distribution over just a subset** of them.
- The probability distribution over the subset is known as the **marginal probability** distribution.

Symptom (y)	Disease (x)	
	Yes	No
Yes	0.4	0.3
No	0.2	0.1

Figure: Setup for marginal probability

Marginal Probability Contd.

Symptom (y)	Disease (x)	
	Yes	No
Yes	0.4	0.3
No	0.2	0.1

Symptom (y)	Disease (x)	
	Yes	No
Yes	0.4	0.3
No	0.2	0.1

Figure: **Top:** $P(y)=\text{yes}$. **Bottom:** $P(x)=\text{no}$.

Marginal Probability Contd.

Symptom	Disease		
	<i>Yes</i>	<i>No</i>	<i>Total</i>
<i>Yes</i>	0.4	0.3	0.7
<i>No</i>	0.2	0.1	0.3
<i>Total</i>	0.6	0.4	1

Figure: Summing over the margins (for discrete random variables). Note: We integrate for continuous random variables.

- For example, suppose we have discrete random variables x and y and we know $P(x, y)$. **We can find $P(x)$ with the sum rule:**

4: Marginal Probability

$$\forall x \in x, P(x = x) = \sum_y P(x = x, y = y).$$

- For continuous variables, we need to **use integration** instead of summation:

$$p(x) = \int p(x, y) dy.$$

Conditional Probability

- The probability of some event, given that some other event has happened.
- Conditional probability denoted: $y=y$ given $x=x$ as $P(y=y \mid x=x)$.

5: Conditional Probability

$$P(y = y \mid x = x) = \frac{P(y = y, x = x)}{P(x = x)}$$

- Conditional probability is **only defined when** $P(x=x) > 0$.
- We **cannot compute** the conditional probability conditioned on an event that **never happens**.

The Chain Rule I

- Any joint probability distribution over **many** random variables may be decomposed into conditional distributions over **only one** variable.
- This is known as the **chain rule** or **product rule**.

6: Chain Rule

$$P(X^{(1)}, \dots, X^{(n)}) = P(X^{(1)}) \prod_{i=2}^n P(X^{(i)} | X^{(1)} \dots X^{(i-1)}).$$

- With 4 variables A_4, A_3, A_2, A_1 , we get:

7: Chain Rule

$$P(A_1, A_2, A_3, A_4) = \\ P(A_1)P(A_2|A_1)P(A_3|A_2, A_1)P(A_4|A_3, A_2, A_1)$$

- x and y are **independent** if the realization of one does not affect the probability distribution of the other.
- In other words, we can express their probability distribution as a **product** of two factors, one involving only x and one involving only y :

8: Independence

$$P(x, y) = P(x)P(y)$$

Conditional Independence

- Two random variables x and y are **conditionally independent** given z if, once z is known, the value of y does not add any additional information about x (Wikipedia).
- In other words, the **conditional probability distribution over x and y factorizes** as follows, for every value of z :

9: Independence

$$P(x, y|z) = P(x|z)P(y|z)$$

Independence Notation

- We can **denote independence and conditional independence with compact notation**:
 - $x \perp y$: x and y are independent
 - $x \perp y|z$: x and y are conditionally independent given z . (See LaTeX symbols [\[link\]](#).)

Expectation: Discrete Random Variables

- The **expectation** or expected value of some function $f(x)$ with respect to a probability distribution $P(x)$ is **the average or mean value that f takes on when x is drawn from P .**
- For **discrete variables** this can be **computed with a summation**:

10: Expectation for Discrete Variables

$$\mathbb{E}_{x \sim P}[f(x)] = \sum_x P(x)f(x).$$

Expectation: Continuous Random Variables

- For **continuous variables**, expectation is **computed with an integral**:

11: Expectation for Continuous Variables

$$\mathbb{E}_{x \sim p}[f(x)] = \int p(x)f(x)dx.$$

- The **variance** gives a measure of how much the values of a function of a random variable x vary as we sample different values of x from its probability distribution:

12: Variance

$$\text{Var}(f(x)) = \mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2]$$

- When the **variance** is low, the **values of $f(x)$** cluster near their expected value.
- The square root of the variance is known as the **standard deviation**.
- σ : standard deviation; σ^2 : variance.

- The **covariance** gives a sense of
 - how much two values are *linearly* related to each other
 - the *scale* of these variables:
- As below, the **covariance between x and y is the expected product of their deviations from their individual expected values.**

13: Covariance

$$\text{Cov}(f(x), g(y)) = \mathbb{E}[(f(x) - \mathbb{E}[f(x)])(g(y) - \mathbb{E}[g(y)])]$$

On Covariance

- **High absolute values** of the covariance mean that the values change very much and are both far from their respective means at the same time.
- If the **sign of the covariance is positive**, then both variables tend to take on relatively high values simultaneously.
- If the **sign of the covariance is negative**, then one variable tends to take on a relatively high value at the times that the other takes on a relatively low value and vice versa.

Bernoulli Distribution I

- A distribution over a **single *binary* random variable**.
- Controlled by a single parameter $\phi \in [0, 1]$, which (i.e., ϕ) gives the probability of the random variable being equal to 1.
- You can think about $\phi = 1$ as **success**, and $\phi = 0$ as **failure**.
- The **Bernoulli distribution** is a special case of the **binomial distribution** (where we run **many** trials, rather than just 1).

14: Properties of Bernoulli Distribution

$$P(x = 1) = \phi$$

$$P(x = 0) = 1 - \phi$$

$$P(x = x) = \phi^x(1 - \phi)^{1-x}$$

$$\mathbb{E}_x[x] = \phi$$

$$\text{Var}_x(x) = \phi(1 - \phi)$$

Multinomial Distribution I

Multinomial Distribution

- The **multinomial distribution** models the probability of counts for rolling a *k*-sided die for *n* times.
- **Probability mass function** can be calculated as follows: ($n!$ the factorial of n , the product of numbers from 1 to $n = 1 \times 2 \times 3 \dots \times n$).

15: Multinomial Distribution

$$p = \frac{n!}{x_1! \dots x_k!} p_1^{x_1} \dots p_k^{x_k}$$

Multinomial & Bernoulli

- **Recall:** When *k* (possible outcomes) is 2 and *n* (number of trials) is 1, the multinomial distribution is the **Bernoulli distribution**.

Example

- Suppose that two chess players had played numerous games and it was determined that the probability that Player A would win is 0.40, the probability that Player B would win is 0.35, and the probability that the game would end in a draw is 0.25. The multinomial distribution can be used to answer questions such as: "If these two chess players played 12 games, what is the probability that Player A would win 7 games, Player B would win 2 games, and the remaining 3 games would be drawn?"
- For more, see the original [example] and [Wikipedia]...

Multinomial Distribution III

- n = total # of events 12 (12 games are played)
- $n_1 = 7$ (number of times Outcome A occurs; games won by Player A)
- ...
- $p_1 = 0.40$ (probability of Outcome A; that player A wins)
- ...

16: Chess Game Solution

$$p = \frac{n!}{(x_1!)(x_2!)(x_3!)}(p_1^{x_1})(p_2^{x_2})(p_3^{x_3})$$

$$p = \frac{12!}{(7!)(2!)(3!)}(.40^7)(.35^2)(.25^3) = 0.0248$$

Multinomial/Categorical Distribution

Multinomial/Categorical Distribution

- Describes a distribution over a single discrete r.v. with k different states, when k is finite and n is 1.
- A **multinomial distribution** is the distribution over vectors in $\{0, \dots, n\}^k$ representing how many times each of the k categories is visited when n samples are drawn from a **multinoulli** distribution.
- Special case of the **multinomial distributions**, with $k > 2$ and $n = 1$.
- Parameterized by a **vector** \mathbf{p} :

17: Multinoulli Distribution

$$\mathbf{p} \in [0, 1]^{k-1}$$

where p_i gives the probability of the i -th state.

18: Gaussian Distribution

$$x \sim \mathcal{N}(\mu, \sigma^2)$$

Can also write it as:

$$\mathcal{N}(x; \mu, \sigma^2)$$

where $\mu \in \mathbb{R}$ and $\sigma \in (0, \infty)$

- μ : mean of the distribution.
- σ : standard deviation
- σ^2 : variance

Standard Normal Distribution

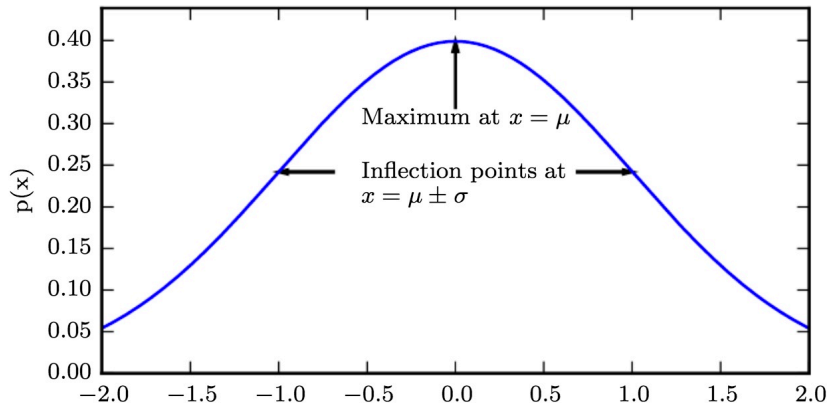


Figure: Standard normal distribution, with $\mu = 0$ and $\sigma = 1$. [From Goodfellow et al. (2016)].

Other distributions and mixtures of distributions are also listed in the Goodfellow et al. (2016).