ML101: Introduction to Machine Learning **Lecture Date: January, 2021**

A Brief Review of Probability Theory ©GuangBing Yang, 2021. All rights reserved.

Lecturer: Guangbing Yang *Scribe: Guangbing Yang*

1 General Concepts

- The **probability theory** studies the **uncertainty**.
- **Elements of probability**
- if $w \in \Omega$, then $P(w)$ is the probability of event w

1. **Sample space** - the set of all possible outcomes which randomly occur, such situations are called experiments. The sample space is denoted by Ω, Ω is the event sets. Each outcome, denoted as $w \in \Omega$, then $P(w)$ is the probability of event w. 0 is called empty space, and $P(0) = 0$, and $0^c = \Omega$.
2. **Events** - a particular subset of Ω, denoted as A, $A \subseteq \Omega$.
3. **Probability measure** - A function $P : F \rightarrow R$ that satisfies the following properties,
   (a) $P(w) \geq 0$
   (b) $\sum_{w \in \Omega} P(w) = 1$
   (c) If A_1 and A_2 are disjoint, then $P(A_1 \cup A_2) = P(A_1) + P(A_2)$, more generally, if $A_1, A_2, ...A_n$ are mutually disjoint, then
   $$P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i) \tag{1}$$
   (d) The first two axioms are obvious. The first axiom simply states that a probability is non-negative. The second exists because Ω consists of all possible outcomes, $P(\Omega) = 1$. The third axiom states that if A and B are disjoint–that is, have no outcomes in common–then $P(A \cup B) = P(A) + P(B)$ or $P(A, B) = P(A) + P(B)$. Sometimes, we also use $P(A, B)$ to denote joint distribution. It is the same definition as the notation $P(A \cup B)$.
   (e) Addition Law $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. $P(A \cap B)$ is called **intersection** of two events A and B.
4. **Properties**
   (a) If $A \subseteq B, P(A) \leq P(B)$
   (b) $P(A \cap B) = min(P(A), P(B))$
   (c) $P(A \cup B) \leq P(A) + P(B)$ - (Union Bound)
   (d) $P(A^c) = 1 - P(A)$ since A^c is called A's complement, A^c and A are disjoint. $A^c \cup A = \Omega$, and $P(A^c \cup A) = P(\Omega) = 1 = P(A^c) + P(A)$
   (e) (Law of Total Probability) If $A_1, ..., A_k$ are a set of disjoint events such that $\sum_{i=1}^{k} A_k = \Omega$, then $\sum_{i=1}^{k} P(A_k) = P(\Omega) = 1$

- **Product Law** – Let A and B be events and assume $P(B) \neq 0$, then $P(A, B) = P(A|B)P(B)$.

- A ***Random Variable*** X is a function from Ω to some set of values χ.

  1. Let us consider an experiment in which a coin is flipped three times, and the sequence of heads and tails is observed as: $\Omega = \{hhh, hht, htt, hth, ttt, tth, thh, tht\}$.
  2. Define the random variables such as (1) the total number of heads, (2) the total number of tails, and (3) the number of heads minus the number of tails. Each of these is a real-valued function defined on Ω. In other words, each of them is a rule that assigns a real number to every point $w \in \Omega$.
  3. In general, a random variable in probability is a function from Ω to a real number. Because the outcome of the experiment with sample space Ω is random, the number produced by the function is also random as well.
  4. Normally, we denote random variables using upper case letters $X(w)$ or more simply X (where the dependence on the random outcome w is implied). We will denote the value that a random variable may take on using lower case letters x. A discrete random variable is a random variable that can take on a finite.
  5. A continuous random variable is a random variable that takes on an infinite number of possible values. Let?s denote the probability that X takes on a value between two real constants a and b (where $a < b$) as:
  6. if χ is countable then X is a *discrete* random variable
  7. if χ is continuous the X is a *continuous* random variable

- if x is a possible value for X, then $P(X = x) = \sum_{w \in \Omega, X(w)=x} P(w)$

- **Cumulative distribution functions (CDF)** is a function $F_X : \mathbb{R} \to [0, 1]$ that specifies a probability measure as,

  $$F_X(x) = P(X \le x) \tag{2}$$

  This function is used to calculate the probability of any event in $\mathbb{F}$.

  Properties:

  1. $0 \le F_X(x) \le 1$.
  2. $lim_{x \to -\infty} F_X(x) = 0$.
  3. $lim_{x \to \infty} F_X(x) = 1$.
  4. $x \le y \implies F_X(x) \le F_X(y)$.

- **Probability mass functions (PMF)** – To discrete random variables, the PMF is used to measure the probability. It is a function $p_X : \Omega \to \mathbb{R}$ in which

  $$p_X(x) = P(X = x) \tag{3}$$

  Properties:

  1. $0 \le p_X(x) \le 1$.
  2. $\sum_{x \in X} p_X(x) = 1$
  3. $\sum_{x \in A} p_X(x) = P(X \in A)$

- **Probability density functions (PDF)** – To continuous random variables, the PDF is defined as the derivative of the CDF since the cumulative distribution function $F_X(x)$ is differentiable.

  $$f_X(x) = \frac{dF_X(x)}{dx} \tag{4}$$

Note that the PDF may not always exists (the cumulative function $F_X(x)$ is not differentiable everywhere. Also, very important point is that the value of PDF at any given point x is not the probability of that event, which is $f_X(x) \neq P(X = x)$, and $f_X(x)$ can be a value larger than 1. We know no probabilistic value is larger than 1.

Properties:

1. $f_X(x) \geq 0$.
2. $\int_{-\infty}^{\infty} f_X(x) = 1$.
3. $\int_{x \in A} f_X(x)dx = P(X \in A)$.

- **Independence and conditional distributions**

  1. Two random variables (RVs), X and Y are *independent* **iff** $P(X,Y) = P(X)P(Y)$
  2. The *conditional distribution* of Y given X is: $P(Y|X) = \frac{P(Y,X)}{P(X)}$, so X and Y are independent iff $P(Y|X) = P(Y)$
  3. The joint distribution of a sequence of RVs can be interpreated as a product of conditionals: $P(X_1, ..., X_n) = P(X_1)P(X_2|X_1)P(X_3|X_2, X_1)...P(X_n|X_{n-1}, .., X_1)$. The probability of generating $X_1, ..., X_n$ "at once" is the same as generating them one at a time if each X_i is conditioned to the $X_1, ..., X_{i-1}$ that preceded it.

- **Conditional distributions**

  1. It's always possible to factor any distribution over $X = (X_1, ..., X_n)$ into a product of conditionals $P(X) = \prod_{i=1}^{n} P(X_i|X_1, ..., X_{i-1})$
  2. However, in many interesting cases, X_i depends only on a subset of $X_1, ..., X_{i-1}$, i.e., $P(X) = \prod_i P(X_i|X_{Pa(i)})$, where $Pa(i) \subseteq \{1, ..., i-1\}$ and $X_S = \{X_j : j \in S\}$
  3. X and Y are *conditionally independent* given Z iff $P(X, Y|Z) = P(X|Z)P(Y|Z)$ or equivalently, $P(X|Y, Z) = P(X|Z)$
  4. Note: the "parents" $Pa(i)$ of X_i depend on the order in which the variables are enumerated!

- **Bayesian Networks**

  1. A Bayesian network is a graphical depiction of a factorization of a probability distribution into products of conditional distributions $P(X) = \prod_i P(X_i|X_{Pa(i)})$
  2. A Bayesian network has a node for each variable X_i and an arc from X_j to X_i iff $j \in Pa(i)$. It represents the full joint distribution.
  3. Inference in Bayesian networks means computing the probability distribution of a set of query variables, given a set of evident variables. Prior, likelihood, posterior, and evidence are important concepts in the Bayesian networks.
  4. The probabilistic inference is to compute the posterior probability distribution for a set of query variables (unknown variables), given some observed event (evidence variable).
  5. **Bayes Rule** is defined as: Let A and $B_1, ..., B_n$ be events where the B_i are disjoint, $\cup_{i=1}^{n} B_i = \Omega$, and $P(B_i) > 0$ for all i, Then

     $$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)} \quad (5)$$

     where $\sum_{i=1}^{n} P(A|B_i)P(B_i) = P(A)$ is called evidence or marginal distribution of joint probability of A and B over B.

2 Expected Values

2.1 Expectation

- Definition of Expectation (or Expected value) of discrete random variable: If X is a discrete random variable with PMF $p_X(x)$ and $g : \mathbb{R} \to \mathbb{R}$ is an arbitrary function. In this case, g(X) can be considered a random variable, the expected value of g(X), denoted by $E(g(X))$ is

$$E(g(X)) = \sum_{x \in X} g(x) p_X(x) \tag{6}$$

provided that $\sum_{x \in X} |g(x)| p_X(x) < \infty$. If the sum diverges, the expectation is undefined.

- For the expectation of random variable X, $E(X) = \sum_i x_i p(x_i)$ is also referred to as the **mean** of X and often denoted by μ.

- Intuitively, the expectation of g(X) can be thought of as a 'weighted average' of the values that g(x) can taken on for different values of x, where the weights are given by $p_X(x)$ or $f_X(x)$. The **mean** of X can be treated as a special case when g(x)=x.

- Definition of Expectation (or Expected value) of continuous random variable If X is a continuous random variable with PDF $f_X(x)$, then the expected value of g(X) is defined as,

$$E(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x) dx \tag{7}$$

provided that $\int |g(x)| f_X(x) dx < \infty$. If the integral diverges, the expectation is undefined.

Properties:

1. $E[a] = a$ for any constant $a \in \mathbb{R}$.
2. $E[af(X)] = aE[f(X)]$ for any constant $a \in \mathbb{R}$.
3. $E[f(X) + g(X)] = E[f(X)] + E[g(X)]$.
4. $E[1X = k] = P(X = k)$ for a discrete random variable X.

2.2 Variance

The definition of variance of random variable X with expected value E(X) is given as:

$$Var(X) = E\{[X - E(X)]^2\} \tag{8}$$

or

$$Var(X) = E[X^2] - [E(X)]^2 \tag{9}$$

provided that the expectation exists. The standard deviation of X is the square root of the variance.

properties:

1. $Var[a] = 0$ for any constant $a \in \mathbb{R}$.
2. $Var[af(X)] = a^2 Var[f(X)]$ for any constant $a \in \mathbb{R}$.

3 Common probability distributions

3.1 Discrete distribution

- **Bernoulli distribution** - A Bernoulli random variable takes on only two values: 0 and 1, with probabilities $1 - p$ and p, respectively. Its frequency function is thus

$$p(x) = \begin{cases} p & \text{if p = 1} \\ 1 - p & \text{if p = 0} \end{cases} \tag{10}$$

  An alternative and very useful representation of this function is

$$p(x) = \begin{cases} p^x(1-p)^{1-x} & \text{if x = 0 or x=1} \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

  This is actually the cost function of the binary classification. Thus, the Bernoulli distribution is the mathematical theory behind the binary classification.

- **Binomial distribution** - Suppose that n independent experiments, e.g. throw a coin n times, are performed, where n is a fixed number, and that each experiment results in a 'success' with probability p and a 'failure' with probability $1 - p$. The total number of successes, X, is a binomial random variable with parameters n and p.

  The probability that $X = k$, or $p(k)$, is given as

$$p(k) = \begin{pmatrix} n \\ k \end{pmatrix} p^k (1-p)^{n-k} \tag{12}$$

  The binomial distribution is the mathematical backend of the logit model and binomial regression in machine learning.

3.2 Continuous distribution

- **Uniform (a, b)** - (where $a < b$): equal probability density to every value between a and b on the real line.

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if a } \leq \text{x} \leq \text{b} \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

  A uniform random variable on the interval [0, 1] is a model for what we mean when we say 'choose a number at random between 0 and 1.' For a uniform density on [0,1], the function is given as

$$f(x) = \begin{cases} 1 & \text{if 0 } \leq \text{x} \leq 1 \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

- **Normal distribution** or **Gaussian distribution** - the density function is given as,

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \tag{15}$$

  It depends on two parameters μ and σ, where $-\infty < \mu < \infty$, and $\sigma > 0$. The parameters μ and σ are called **mean** and **standard deviation** of the normal density. Here the parameter μ determines the density position and the parameter σ determines the shape of the density (e.g., how wide or narrow the curve).

  If $\mu = 0$ and $\sigma = 1$, the normal density is called the **standard normal density**.