Probability and Statistics

Probability and Key Concepts

- **Probability** is a measure of the likelihood of an event occurring;
- $P(A)$ is the probability of event A;
- $P(A) = \frac{N_A}{N}$ where N_A is the number of ways event A can occur and N is the total number of possible outcomes;
- **Frequentist definition**: $P(A) = \lim_{N\to\infty} \frac{N_A}{N}$ - the probability of an event is the limit of its relative frequency in a large number of trials.
- **Sample space** is the set of all possible outcomes of an experiment;
- **Event** is a subset of the sample space.

Kolmogorov Axioms

- $P(A) \geq 0$ for all events A
- $P(X) = 1$ where X is the sample space
- $P(A \cup B) = P(A) + P(B)$ for all disjoint events A and B

From these axioms, we can derive the following:

- $P(\emptyset) = 0$
- $C \subseteq D \implies P(C) \leq P(D)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Conditional Probability and Independence

- **Conditional probability** of event A given event B is $P(A|B) = \frac{P(A \cap B)}{P(B)}$, where $P(B) > 0$;
- Events A and B are **independent** if $P(A \cap B) = P(A)P(B)$;
  - If A and B are independent, then $P(A|B) = P(A)$ and $P(B|A) = P(B)$.

Law of Total Probability and Bayes Theorem

- **Law of Total Probability**: $P(A) = \sum_i P(A|B_i)P(B_i)$ where B_i are disjoint events such that $\cup_i B_i = X$;
- **Bayes Theorem**: $P(B|A) = \frac{P(A|B)P(B)}{P(A)}$.

Random Variables

Random variable is a function $X : \Omega \to \mathbb{R}$ that maps each outcome to a real number;

Discrete Random Variables

- **Discrete random variable** is a random variable that takes on a finite or countably infinite number of values;
- **Distribution function** of a discrete random variable X is $F_X(x) = P(X \leq x)$;
- **Probability mass function** of a discrete random variable X is $p_X(x) = P(X = x)$.

There are many discrete probability distributions, including:

- **Uniform**: $f_X(x_i) = \frac{1}{n}$ for $i = 1, \ldots, n$;
- **Bernoulli**: $f_X(x) = p^x(1-p)^{1-x}$ for $x \in \{0, 1\}$, or:

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

- **Binomial** is the sum of n independent Bernoulli trials: $f_X(x) = Binomial(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x}$ for $x \in \{0, 1, \ldots, n\}$;
  - The binomial coefficient $\binom{n}{x} = \frac{n!}{x!(n-x)!}$ is the number of ways to choose x items from n items.

Continuous Random Variables

- **Continuous random variable** is a random variable that takes on an uncountably infinite number of values;
- **Distribution function** of a continuous random variable X is $F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)dt$;
- **Probability density function** of a continuous random variable X is $f_X(x)$ such that $P(a \leq X \leq b) = \int_a^b f_X(x)dx$.

There are many continuous probability distributions, including:

- **Uniform**:

$$f_X(x) = Uniform(x; a, b) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

- **Normal** (or Gaussian):

$$f_X(x) = N(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- **Exponential**:

$$f_X(x) = Exponential(x; \lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Expectation of Random Variables

- **Expectation** of a random variable X is:

$$E[X] = \begin{cases} \sum_x x p_X(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} x f_X(x)dx & \text{if } X \text{ is continuous} \end{cases}$$

- **Linearity of expectation**: $E[X + Y] = E[X] + E[Y]$;

- $E[aX+b] = aE[X]+b$ for constants a and b;
- The **expectation of a function of a random variable** $g(X)$ is:

$$E[g(X)] = \begin{cases} \sum_x g(x)p_X(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} g(x)f_X(x)dx & \text{if } X \text{ is continuous} \end{cases}$$

Two (or More) Random Variables

- **Joint distribution** of two random variables X and Y is $F_{XY}(x,y) = P(X \leq x, Y \leq y)$;
- **Joint probability mass function** of two discrete random variables X and Y is $p_{XY}(x,y) = P(X=x, Y=y)$;
- **Joint probability density function** of two continuous random variables X and Y is $f_{XY}(x,y)$ such that $P((X,Y) \in A) = \iint_A f_{XY}(x,y)dxdy$;
- **Marginalization** is the process of obtaining the distribution of one variable from the joint distribution of two variables:

$$f_X(x) = \begin{cases} \sum_y f_{XY}(x,y) & \text{if } X \text{ and } Y \text{ are discrete} \\ \int_{-\infty}^{\infty} f_{XY}(x,y)dy & \text{if } X \text{ and } Y \text{ are continuous} \end{cases}$$

- **Independence** of two random variables X and Y is $F_{XY}(x,y) = F_X(x)F_Y(y)$ for all x and y.

There are many joint distributions, including:

- **Multinomial** is the generalization of the binomial distribution to more than two outcomes:

$$f_{X_1,\ldots,X_k}(x_1,\ldots,x_k; n, p_1,\ldots,p_k) = \frac{n!}{x_1!\ldots x_k!} p_1^{x_1}\ldots p_k^{x_k}$$

- **Multivariate Gaussian** is the generalization of the normal distribution to more than one dimension:

$$f_{X_1,\ldots,X_k}(x_1,\ldots,x_k; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^k|\Sigma|}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$$

Conditionals and Bayes' Theorem

- **Conditional pmf** of X given Y is $p_{X|Y}(x|y) = P(X = x|Y = y) = \frac{p_{XY}(x,y)}{p_Y(y)}$;
- **Conditional pdf** of X given Y is $f_{X|Y}(x|y) = \frac{f_{XY}(x,y)}{f_Y(y)}$;
- **Bayes' Theorem** for two random variables X and Y is:

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)f_X(x)}{f_Y(y)}$$

Covariance and Correlation

- **Covariance** of two random variables X and Y is $cov(X,Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$;
  - $cov(X,X) = var(X)$;
- **Covariance matrix** of $X = (X_1, \dots, X_k)$ is:

$$\Sigma = \begin{bmatrix} var(X_1) & cov(X_1,X_2) & \dots & cov(X_1,X_k) \\ cov(X_2,X_1) & var(X_2) & \dots & cov(X_2,X_k) \\ \vdots & \vdots & \ddots & \vdots \\ cov(X_k,X_1) & cov(X_k,X_2) & \dots & var(X_k) \end{bmatrix}$$

- The covariance of Gaussian: $N(x;\mu,\Sigma)$ is Σ.

Entropy

- **Entropy** of a discrete random variable X is the expected value of the information content of X - is the **uncertainty/randomness** of X:

$$H(X) = E[I(X)] = E[-\log_2 p_X(x)] = -\sum_x p_X(x)\log_2 p_X(x)$$

- **Positivity**: $H(X) \geq 0$;
- **Maximum entropy**: $H(X) \leq \log_2 n$ where n is the number of possible values of X;

The entropy of a continuous random variable X is:

$$H(X) = E[I(X)] = E[-\log_2 f_X(x)] = -\int_{-\infty}^{\infty} f_X(x)\log_2 f_X(x)dx$$

Kullback-Leibler Divergence

The **Kullback-Leibler divergence** of two distributions p and q is the expected value of the information gained when one revises one's beliefs from the prior probability distribution q to the posterior probability distribution p:

$$D_{KL}(p||q) = E_{x\sim p}[\log_2 \frac{p(x)}{q(x)}] = \sum_x p(x) \log_2 \frac{p(x)}{q(x)}$$

- **Positivity**: $D_{KL}(p||q) \geq 0$;
- **Non-negativity**: $D_{KL}(p||q) = 0$ if and only if $p = q$.

For continuous distributions, the KL divergence is:

$$D_{KL}(p||q) = E_{x\sim p}[\log_2 \frac{p(x)}{q(x)}] = \int_{-\infty}^{\infty} p(x) \log_2 \frac{p(x)}{q(x)} dx$$

- **Positivity**: $D_{KL}(p||q) \geq 0$;
- **Non-negativity**: $D_{KL}(p||q) = 0$ if and only if $p = q$.