Chapter 3 Probability and Information Theory

3.1 Why probability ?

Source of Uncertainty

- Inherent stochasticity in the system being modeled
- Incomplete observation
- Incomplete modeling

Frequentist and Bayesian Probability

- Frequentist: View the probability as the proportion that would result in certain outcome out of infinite many repetitions
- Bayesian: Use probability to present a degree of brief

3.2 Random Variables

Definition of this book: A random variable is a variable that can take on different value randomly

Formal definition: [From wiki]

3.3 Probability Distributions

3.3.1 Discrete Variables and Probability Mass Function(PMF)

3.3.2 Continuous Variables and Probability Density Function(PDF)

3.4 Marginal Probability

3.5 Conditional Probability

3.6 The Chain Rule of Conditional Probability

3.7 Independence and Conditional Independence

3.8 Expectation, Variance and Covariance

Expectation

The expectation of some function $f(x)$ with respect to a probability $P(x)$ is:

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

- Expectation are linear
  $\mathbb{E}_P[\alpha f(x) + \beta g(x)] = \alpha\,\mathbb{E}_P[f(x)] + \beta\,\mathbb{E}_P[g(x)]$

Variance

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

Covariance

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

- Covariance measure the **linear correlation** between $f(x)$ and $g(x)$

Covariance Matrix of a random vector

For a random vector $\mathbf{x} \in R^n$, covariance matrix is a $n \times n$ matrix, such that

$Cov(x)_{i,j} = Cov(x_i, x_j)$

$Cov(x)_{i,i} = Var(x_i)$

3.9 Common Probability Distribution

3.9.1 Bernoulli Distribution

3.9.2 Multinoulli Distribution

3.9.3 Gaussian Distribution

Gaussian Distribution is a good choice for default distribution without prior knowledge since:

- First, many distributions are close to normal distributions. Due to central limit theorem(CLT), the sum of many independent random variables is approximately normally distributed.
- Second, the Gaussian(normal) distribution encode the maximum amount of uncertainty over the real numbers out of all possible probability distribution with the same variance

Multivariable Distribution

3.9.4 Exponential and Laplace Distribution

Exponential Distribution: $p(x;\lambda) = \lambda e^{-\lambda x}$

- Have a distribution with sharp point at $x = 0$

Laplace Distribution:

- Generalized exponential distribution that allow a sharp point at arbitrary points

3.9.5 The Dirac Distribution and Empirical Distribution

3.9.6 Mixtures of Distribution

- Latent variable
- Gaussian mixture model

3.10 Useful Properties of Common Functions

- Logistic sigmoid

- Softmax
- Softplus

3.11 Bayes' Rules

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

3.12 Technical Details of Continuous Variables

Measure theory

- Measure zero
- Almost everywhere

3.13 Information Theory

Self-information of an event x

$I(x) = -\log P(x)$

Shannon Entropy of a probability distribution

$H(x) = E_{x~P}[I(x)] = E_{x~P}[-\log P(x)]$

KL Divergence of two probability distribution

$D_{KL}(P||Q) = E_P[\log P(x) - \log Q(x)]$

- $D_{KL}(P||Q) \neq D_{KL}(Q||P)$

Cross-entropy of two probability P and Q

- $H(P, Q) = H(P) + D_{KL}(P||Q)$

3.14 Structured Probabilistic Model

Graphical model / Graphical probability model

- Directed

- Undirected