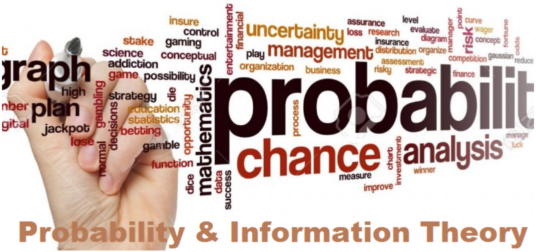
Review of Probability and Information Theory



Leon Guo

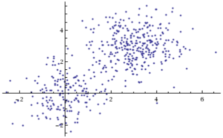
November 6, 2016

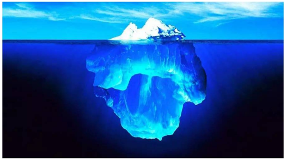
Overview

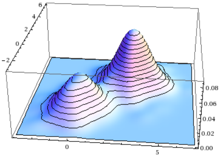
- Probability Theory
  - Basics
  - Common Discrete or Continuous Distribution
  - Useful Functions
  - Bayesian Rule & Statistics
- Probabilistic Graphical Model
- Information Theory

Probability Theory

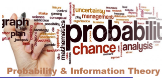
- Uncertainty
- What? Why? How?
  - Inherent stochasticity
  - Incomplete observability
  - Incomplete modeling
- vs. Statistics
- vs. Stochastic Process







Probability Theory - Basics



- Axioms of Probability, event(sum,product)
- Random Variables
- Probability Distributions
  - Joint, Marginal Probability
  - Conditional Probability
    - Chain Rule
    - Independency
- Characteristics of Probability Distribution
  - Expectation, Variance & Covariance, Moment
  - Expectation: Long term certainty underlaying the short term uncertainty
  - Variance: Measuring the uncertainty
  - Covariance: The correlation of two or more RVs,
- univariate $\rightarrow$ multivariate distribution
- Measure-theoretic definition of RV, PDF and $\cdots$, is REALLY cool but a little bit nerdy, and beyond today's topic.

Common Discrete Distributions

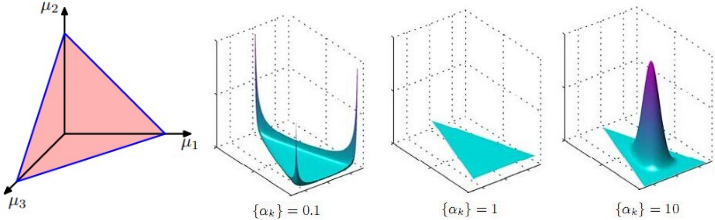
- Bernoulli, $f(k;p) = p^k(1-p)^{1-k}$ for $k \in \{0,1\}$
- Binomial, $f(k;n,p) = \Pr(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$
- Multinomial

$$
\begin{aligned}
& f(x_1, \ldots, x_k; n, p_1, \ldots, p_k) \\
= & \Pr(X_1 = x_1 \text{ and } \ldots \text{ and } X_k = x_k) \\
= & \begin{cases} \dfrac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}, & \text{when } \sum_{i=1}^k x_i = n \\ 0 & \text{otherwise,} \end{cases}
\end{aligned}
$$

  The PMF can be expressed using the gamma function as:
  $f(x_1, \ldots, x_k; p_1, \ldots, p_k) = \frac{\Gamma(\sum_i x_i + 1)}{\prod_i \Gamma(x_i + 1)} \prod_{i=1}^k p_i^{x_i}$.
- Categorical: one-hot of Multinomial; multivariate Bernoulli. PMF is just $f(x_1, \ldots, x_k; p_1, \ldots, p_k) = \prod_{i=1}^k p_i^{x_i}$.
- Poisson : $P(x=k) = \frac{\lambda^k e^{-\lambda}}{k!}$, k events occur in interval.

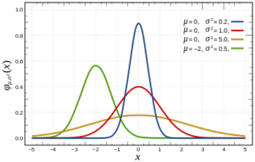
Common Continuous Distributions

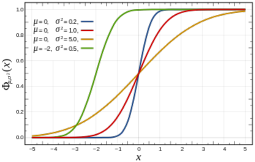
- Dirichlet distribution
  - very important in Bayesian Statistic, as a prior for Multinomial.
  - PDF: $f(x_1, \cdots, x_K; \alpha_1, \cdots, \alpha_K) = \frac{1}{\mathrm{B}(\boldsymbol{\alpha})} \prod_{i=1}^{K} x_i^{\alpha_i - 1}$ and
    $$\mathrm{B}(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}, \qquad \boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_K).$$
  - Support: $\|\boldsymbol{x}\|_1 = 1$, $N-1$ dimension Simplex.
  - Conjugacy: Dirichlet-Multinomial.

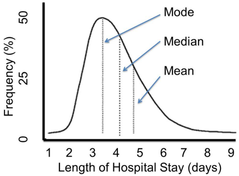


Common Continuous Distributions

- Beta, Gamma, exponential.
- Gaussian
  - Symmetric & Bell shape vs. Skewed shape



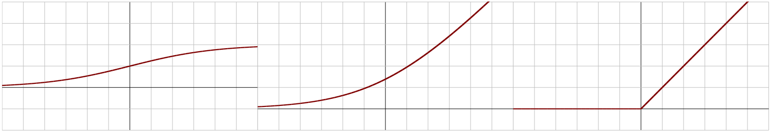




  - PDF: $f(x|\mu,\tau) = \sqrt{\frac{\tau}{2\pi}}\, e^{\frac{-\tau(x-\mu)^2}{2}}$, precision τ (width of distribution).
  - Standardization: $Y = \frac{X-\mu}{\sigma}$. Normal distribution, $Y \sim \mathcal{N}(0,1)$.
  - Perfect & elegant Properties. **CLOSED-FORMED** analytical solution
- Conjugacy: Beta-Bernoulli, Dirichlet-Multinomial, Gamma-Poisson, Normal-inverse-Gamma(Wishart).
- Exponential Family
- Laplace, Student's t, Dirac, Empirical.

Probability Theory - Miscellaneous

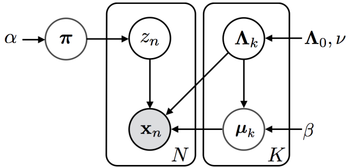
- Activation function
  - Logistic, Softplus, ReLU

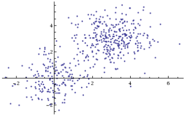


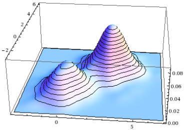
- Change of Variables
  - Eq.3.46 & Eq.3.47
- Bayesian Theorm
  - Posterior $\propto$ Likelihood $\times$ Prior.

Probabilistic Graphical Model

- A modelling approach about (Probability + Graph)
- 3 kinds of graphical model: Direct, Undirect, Factor Graph
- Mixture of Gaussian as an example.



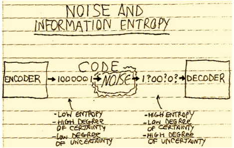


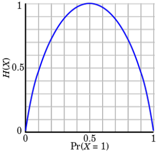


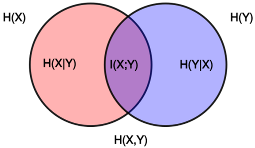
- K-L divergence and Variational Inference

Information Theory - Basics

- What does IT concern about?
- How does IT relate with machine learning, probability?







- 3 important Terms: entropy, relative entropy(KL), conditional entropy

Information Theory - Basics

$$
\begin{aligned}
H(X) &= \int \mathrm{P}(x)\,\underline{\mathrm{I}(x)}\;dx = -\int \mathrm{P}(x)\log_b \mathrm{P}(x)\;dx && (1)\\
D_{\mathrm{KL}}(P\|Q) &= \int_{-\infty}^{\infty} p(x)\,\log\frac{p(x)}{q(x)}\,\mathrm{d}x, && (2)\\
H(X|Y) &= \sum_{i,j} p(x_i,y_j)\log\frac{p(y_j)}{p(x_i,y_j)} && (3)\\
I(X;Y) &= \underline{H(X) - H(X|Y)} = D_{\mathrm{KL}}(p(x,y)\|p(x)p(y)) && (4)
\end{aligned}
$$


Embrace The Uncertainty, Enjoy IT.