Part I

Machine Learning

Machine Learning

Boltzman Machines

Network of symetrically coupled stocastic binary units. Energy of state $\{v, h\}$ is defined as:

$$E(v, h; \theta) = -\frac{1}{2}v^T L v - \frac{1}{2}h^T J h - v^T W h \tag{1}$$

where $\theta = \{W, L, J\}$ represents model parameters. diagonal terms of L and J are set to 0.

---

[0]bias terms have been omitted for clarity

Probability that model assigns a visible vector v is

$$p(v;\theta) = \frac{p^*(v;\theta)}{Z(\theta)} \tag{2}$$

$$= \frac{1}{Z(\theta)} \sum_n \exp(-E(v,h;\theta)) \tag{3}$$

$$Z(\theta) = \sum_v \sum_h \exp(-E(v,h;\theta)) \tag{4}$$

The conditional distribution over hidden and visible units are given by

$$p(h_j = 1|v, h_{-j}) = \sigma\left(\sum_{i=1}^{D} W_{ij}v_i + \sum_{m=1\backslash j}^{P} J_{jm}h_j\right)$$

$$p(v_i = 1|h, v_{-i}) = \sigma\left(\sum_{j=1}^{P} W_{ij}h_j + \sum_{k=1\backslash i}^{D} L_{ik}v_j\right)$$

Parameter updates that are needed to perform gradient ascent in the log-likelihood from Eq. 3:

$$\Delta W = \alpha(\mathbb{E}_{P_{data}}[vh^T] - \mathbb{E}_{P_{model}}[vh^T])$$
$$\Delta L = \alpha(\mathbb{E}_{P_{data}}[vv^T] - \mathbb{E}_{P_{model}}[vv^T])$$
$$\Delta J = \alpha(\mathbb{E}_{P_{data}}[hh^T] - \mathbb{E}_{P_{model}}[hh^T])$$

$\mathbb{E}_{P_{data}}[\cdot]$ represents expectation w.r.to complete data distribution $P_{data}(h, v; \theta) = p(h|v; \theta)P_{data}(v)$, with $P_{data}(v) = \frac{1}{N}\sum_n \delta(v - v_n)$ representing the empirical distribution, and $\mathbb{E}_{P_{model}}$ is an expectation w.r.to. distribution defined by the model.

Reduce the expectation of the model distribution and the data distribution

Graphical Models

A graphical model or probabilistic graphical model (PGM) is a probabilistic model for which a graph expresses the conditional dependence structure between random variables.

Gaussian Bernoulli RBM

$$E(v,h) = \|v-a\|^2 - b^T h - v^T W h$$

$$F(v) = -\ln\left(\sum_h e^{-E(v,h)}\right) \qquad =$$

Bayes Rule

Bayes rule

$$P(A|B) = P(B|A)\frac{P(A)}{P(B)} \tag{5}$$

Law of total probability

$$P(B) = P(B|A_1)P(A_1) + \dots P(B|A_n)P(A_n) \quad (6)$$

where $\{A_i\}_{i=1}^{n}$ are partitions of sample space

Probability distributions

Gaussian/Normal Distribution

$$pdf \equiv \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$cdf \equiv \frac{1}{2}\left[1 + erf\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$$

$$erf(x) = \frac{1}{\sqrt{\pi}} \int_{-x}^{x} e^{-t^2/2} dt$$

Bernoulli Distribution

$$P(X = 1) = p$$

$$E[X] = 1 \cdot P(X == 1) + 0 \cdot P(X == 0) = p$$
$$Var[X] = E[X^2] - E[X]^2 = p - p^2 = p(1 - p) = pq$$

---

[0] Denoted by $Bern(p)$

Binomial Distribution

$$B(n, p)$$

Probability distribution of number of successes in n Bernoulli trials Bern(p)

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$E[X] = \sum_i E[X_i] = np$$

$$Var[X] = Var\left(\mathbf{\Sigma}_i X_i\right) = \mathbf{\Sigma}_i Var(X_i) = npq$$

$X \sim B(n, p)$ and $Y \sim B(m, p)$ are independent binomial variables , $X + Y \sim B(n + m, p)$

Poisson

Dirichlet

Beta

Central Limit Theorem

Central Limit Theorem (CLT) states that, given certain conditions, the arithmetic mean of a sufficiently large number of iterates of independent random variables, each with a well-defined expected value and well-defined variance, will be approximately normally distributed, regardless of the underlying distribution.

A simple example of this is that if one flips a coin many times, the probability of getting a given number of heads should follow a normal curve, with mean equal to half the total number of flips.