「パターン認識と機械学習」
1.6 情報理論

小野寺喜行

情報量を得る尺度【$h(x)$】

- 確率分布 $p(x)$ に依存
- 事象 x, y が無関係のとき

$$h(x, y) = h(x) + h(y)$$

$$p(x, y) = p(x)\,p(y)$$

$$\Downarrow$$

$$h(x) = -\log_2 p(x)$$ 単位:ビット

情報の平均量【$\mathrm{H}[x]$】

$$\mathrm{H}[x] = -\sum_{x} p(x)\log_2 p(x)$$

エントロピー(entropy)と呼ばれる

エントロピー

- 8個の可能な状態を等確率で取るとき

$$\mathrm{H}\left[x\right] = -8 \times \frac{1}{8} \log_2 \frac{1}{8} = 3 \text{ ビット}$$

エントロピー

- 8個の可能な状態$\{a,b,c,d,e,f,g,h\}$

  各確率は$\left(\frac{1}{2},\frac{1}{4},\frac{1}{8},\frac{1}{16},\frac{1}{64},\frac{1}{64},\frac{1}{64},\frac{1}{64}\right)$

  とする時

$$\mathrm{H}[x] = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{8}\log_2\frac{1}{8}$$

$$-\frac{1}{16}\log_2\frac{1}{16} - \frac{4}{64}\log_2\frac{1}{64} = 2 \text{ ビット}$$

エントロピー

- 非一様な分布のエントロピーは一様な場合より小さい
- 確率変数の状態を送信するときの平均符号長と同じになる

多重度

- N個の同じ物体をたくさんの箱に入れる場合、i番目の箱に入る個数をn_iとしたときの箱への入れ方の総数W

$$W = \frac{N!}{\prod_i n_i!}$$

多重度(multiplicity)

多重度のエントロピー

- 多重度の対数を適当に定数倍したもの

$$\mathrm{H} = \frac{1}{N}\ln W = \frac{1}{N}\ln N! - \frac{1}{N}\sum_i \ln n_i!$$

多重度のエントロピー

- n_i / Nを一定で$N \to \infty$とし、スターリングの近似式

$$\ln N! \cong N \ln N - N$$

を使い、$\sum_i n_i = N$より

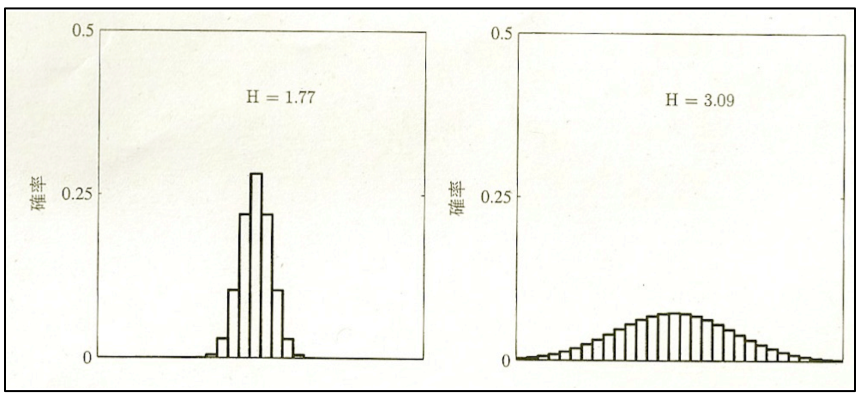
$$\mathrm{H} = -\lim_{N\to\infty} \sum_i \left(\frac{n_i}{N}\right) \ln\left(\frac{n_i}{N}\right) = -\sum_i p_i \ln p_i$$

$$\left(p_i = \lim\nolimits_{N\to\infty}(n_i / N)\right)$$

確率変数X

- 箱を離散確率変数 X の状態 x_i と解釈すると
  $p(X = x_i) = p_i$
- 確率変数X のエントロピー
  $$\mathrm{H}\left[p\right] = -\sum_i p\left(x_i\right)\ln p\left(x_i\right)$$

確率分布とエントロピー



- 少ない値で鋭いピークを持つような分布のエントロピーは低い（左）
- たくさんの値に広がっている分布のエントロピーは高い（右）

確率分布とエントロピー

- $p_i = 1, p_{j \neq i} = 0$ となる分布でエントロピーは最小値0となる
- エントロピーの最大値は

$$\tilde{\mathrm{H}} = -\sum_i p(x_i) \ln p(x_i) + \lambda \left(\sum_i p(x_i) - 1 \right)$$

を最大化する時、つまり$p(x_i)$ すべて等しい分布で $\mathrm{H} = \ln M$

(M は x_i の状態の総数)

確率分布とエントロピー

- 実際に最大値であることを確かめるため、エントロピーを2階微分

$$\frac{\partial \tilde{\mathrm{H}}}{\partial p\left(x_i\right)\partial p\left(x_j\right)} = -I_{ij}\frac{1}{p_i}$$

(I_{ij} は単位行列の成分)

連続変数 x の分布のエントロピー

- xを等間隔の区間Δ に分けると

$$\int_{i\Delta}^{(i+1)\Delta} p(x)\mathrm{dx} = p(x_i)\Delta$$

となるx_i が必ず存在

- x_i の値を観測する確率$p(x_i)\Delta$

連続変数 x の分布のエントロピー

- 離散分布のエントロピー

$$H_\Delta = -\sum_i p(x_i)\Delta \ln\left(p(x_i)\Delta\right)$$

$$= -\sum_i p(x_i)\Delta \ln p(x_i) - \ln\Delta$$

連続変数 x の分布のエントロピー

- $-\ln\Delta$ を無視し、$\Delta \to 0$ の極限をとると

$$\lim_{\Delta\to 0}\left\{-\sum_{p} p(x_i)\Delta \ln p(x_i)\right\} = -\int p(x)\ln p(x)\,\mathrm{d}x$$

微分エントロピー

- 離散と連続のエントロピーは$\ln\Delta$ だけ異なる

連続変数 x の分布のエントロピー

- 複数の連続変数$\mathbf{x}$の、その密度に対する微分エントロピー

$$\mathrm{H}[\mathbf{x}] = -\int p(\mathbf{x}) \ln p(\mathbf{x}) \, \mathrm{d}\mathbf{x}$$

連続変数の場合のエントロピー最大化

- 次の3つの制約のもとで最大化する

$$\int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = 1$$

$$\int_{-\infty}^{\infty} xp(x)\,\mathrm{d}x = \mu$$

$$\int_{-\infty}^{\infty} (x-\mu)^2\, p(x)\,\mathrm{d}x = \sigma^2$$

連続変数の場合のエントロピー最大化

- ラグランジュ乗数法を使い、以下の汎関数を$p(x)$について最大化して求める

$$
-\int_{-\infty}^{\infty} p(x)\ln p(x)\,\mathrm{d}x + \lambda_1\left(\int_{-\infty}^{\infty} p(x)\,\mathrm{d}x - 1\right)
$$

$$
+\lambda_2\left(\int_{-\infty}^{\infty} xp(x)\,dx - \mu\right)
$$

$$
+\lambda_3\left(\int_{-\infty}^{\infty} (x-\mu)^2\, p(x) - \sigma^2\right)
$$

連続変数の場合のエントロピー最大化

- 変分法により汎関数の微分を0とおくと

$$p(x) = \exp\left\{-1 + \lambda_1 + \lambda_2 x + \lambda_3 (x-\mu)^2\right\}$$

$$\Downarrow$$

制約式に代入し、最終的に

$$p(x) = \frac{1}{\left(2\pi\sigma^2\right)^{1/2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

（ガウス分布）

ガウス分布の微分エントロピー

$$\mathrm{H}\left[x\right] = \frac{1}{2}\left\{1 + \ln\left(2\pi\sigma^2\right)\right\}$$

- σ^2 が増えて分布が幅広くなるにつれて大きくなる
- $\sigma^2 < 1/\left(2\pi e\right)$ のとき $\mathrm{H}\left[x\right] < 0$ となる

同時分布$p(\mathbf{x},\mathbf{y})$

- $\mathbf{x}$ が既知のとき、対応する$\mathbf{y}$の値の特定に必要な付加情報

$$\mathrm{H}[\mathbf{y}\mid\mathbf{x}] = -\iint p(\mathbf{y},\mathbf{x})\ln p(\mathbf{y}\mid\mathbf{x})\,\mathrm{d}\mathbf{y}\,\mathrm{d}\mathbf{x}$$

⇩

条件付きエントロピー（conditional entropy）

同時分布$p(\mathbf{x}, \mathbf{y})$

- 確率の乗法定理より、条件付きエントロピーは以下の関係を満たす

$$\mathrm{H}[\mathbf{x}, \mathbf{y}] = \mathrm{H}[\mathbf{y} \mid \mathbf{x}] + \mathrm{H}[\mathbf{x}]$$