Goodfellow：深層学習

Part1：応用数学と機械学習の基礎

第３章：確率と情報理論

情報理論の基礎

- **情報理論における基本的な直観**
  - 起こりそうもない事象が起こったことを学習することは、
    起こりそうな事象が起こったことを学習することよりも価値がある。

- **情報の価値の定式化**
  - 起こりやすい事象の情報量が少ない。間違いなく起こる事象の情報量はゼロ。
  - 起こりにくい事象ほどその情報量は多い。
  - 独立な事象は付加情報を持つ。
    Ex.) コインを投げて表が1回出るのを見るよりも、表が2回出るのを見る方が、
    2倍の情報を伝達する。
  - 上記３つの特性を満足する量として、事象 $X = x$ の自己情報量を
    以下のように定義する。

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

エントロピー

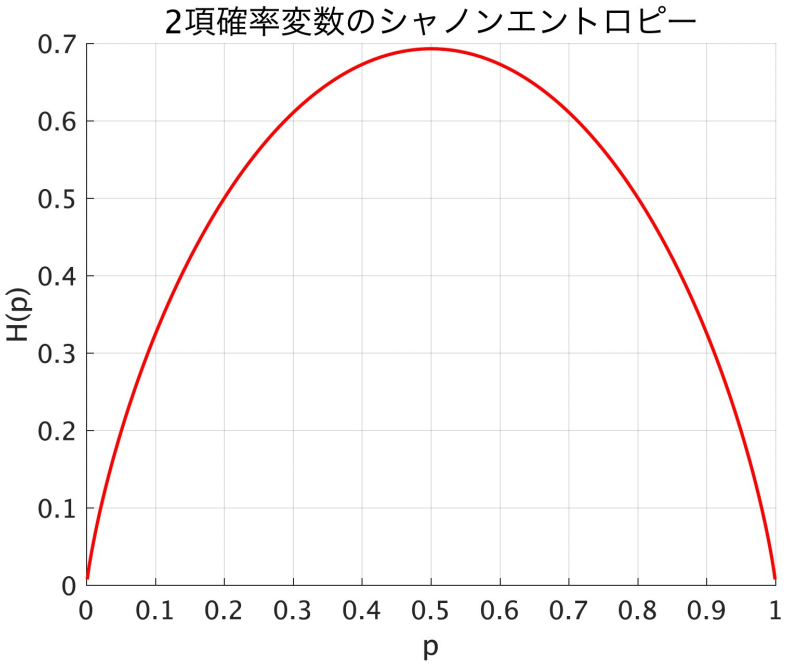
- 自己情報量 $I(x)$ はある１つの事象に関する結果である。
  確率密度全体の不確実性を表現する量として**エントロピー**が定義される。
  - 分布の（シャノン）エントロピー
    → その分布から抽出される事象に期待される情報量

$$H(x) = \mathbb{E}_{X\sim P}[I(x)] = -\mathbb{E}_{X\sim P}[\log P(x)]$$

- **二項確率変数のシャノンエントロピー**
  - 確率変数 X が 0, 1 の２値をとる。
  - X = 1となる確率を p とした時の
    シャノンエントロピーの p 依存性

$$H(p) = -p\log p - (1-p)\log(1-p)$$

p=0.5 の時にH(p)が最大となる



KLダイバージェンス

- 確率分布 P(x) と Q(x) に対する **KLダイバージェンス**
  - 2 つの分布の距離を測る指標

$$D_{\mathrm{KL}}(P\|Q) = \mathbb{E}_{X\sim P}\left[\log \frac{P(x)}{Q(x)}\right] = \mathbb{E}_{X\sim P}[\log P(x) - \log Q(x)]$$

- **KLダイバージェンスの性質**
  - ① KLダイバージェンスは非負。
    KLダイバージェンスが 0 となるのは、P(x)=Q(x) の時に限る。
  - ② KLダイバージェンスは非対称。 $D_{\mathrm{KL}}(P\|Q) \neq D_{\mathrm{KL}}(Q\|P)$
    → KLダイバージェンスは厳密には距離ではない。

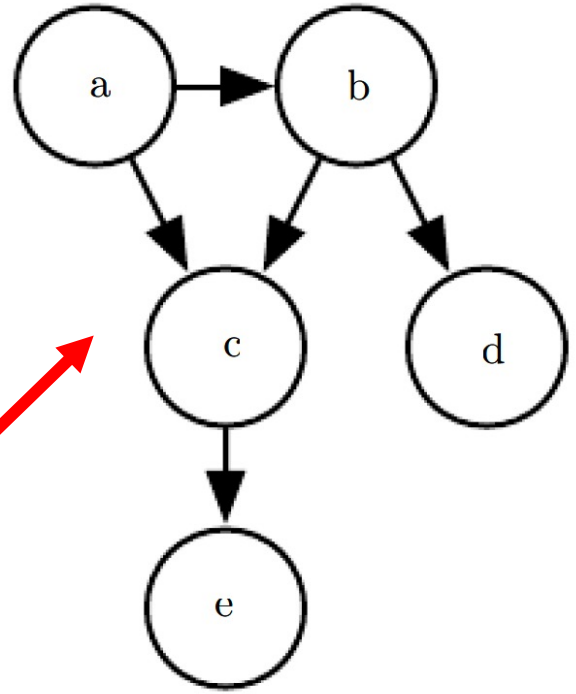
構造化確率モデルのモチベーション

- 確率分布の因数分解
  - 例）３つの確率変数 a, b, c があり、aはbの値に影響を与え、bはcの値に影響を与えるが、bが与えられた下でaとcは独立である場合

$$p(a, b, c) = p(a)p(b|a)p(c|b)$$

- **構造化確率モデル**
  - 確率分布の因数分解をグラフを使って表現する。
  - 各ノードが確率変数に対応。
  - ２つの確率変数が辺で結ばれていることは、確率分布がこの２つの変数の間の直接的な関係を表現できることを意味する。
  - 有向モデル：条件付き確率分布への分解を表現。

$$p(a, b, c, d, e) = p(a)p(b|a)p(c|a, b)p(d|b)p(e|c)$$



無向グラフィカルモデル

- **無向グラフ**
  - 向きのないグラフを利用。関数の集合への因数分解を表現。
  - 各集合（クリーク）は非負の因子 $\phi^{(i)}(\cdot)$ を使って表される。
  - 確率分布は正規化定数 Z により総和、あるいは積分値が１になるように補正されている。

$$p(a, b, c, d, e) = \frac{1}{Z}\phi^{(1)}(a, b, c)\phi^{(2)}(b, d)\phi^{(3)}(c, e)$$