パターン認識と機械学習

1.6.1 相対エントロピーと相互情報量

吉田拓夢

相対エントロピー

- 未知の分布p(x)←近似的にq(x)でモデル化
- 真の分布p(x)の代わりにq(x)でxを符号化
- xの値を特定するために必要な追加情報量の平均は

$$\begin{aligned}\mathrm{KL}(p \,\|\, q) &= -\int p(\mathbf{x}) \ln q(\mathbf{x}) \mathrm{d}\mathbf{x} - \left(-\int p(\mathbf{x}) \ln p(\mathbf{x}) \mathrm{d}\mathbf{x}\right) \\ &= -\int p(\mathbf{x}) \ln \left\{\frac{q(\mathbf{x})}{p(\mathbf{x})}\right\} \mathrm{d}\mathbf{x}\end{aligned}$$

個分布p(x)とq(x)の間の相対エントロピー

相対エントロピー

$$\begin{aligned}\mathrm{KL}(p \,\|\, q) &= -\int p(\mathbf{x}) \ln q(\mathbf{x}) \mathrm{d}\mathbf{x} - \left(-\int p(\mathbf{x}) \ln p(\mathbf{x}) \mathrm{d}\mathbf{x}\right) \\ &= -\int p(\mathbf{x}) \ln \left\{\frac{q(\mathbf{x})}{p(\mathbf{x})}\right\} \mathrm{d}\mathbf{x}\end{aligned}$$

個相対エントロピー(relative entropy)

別称：KL(カルバック-ライブラー)ダイバージェンス

(Kullback-Leibler divergence)

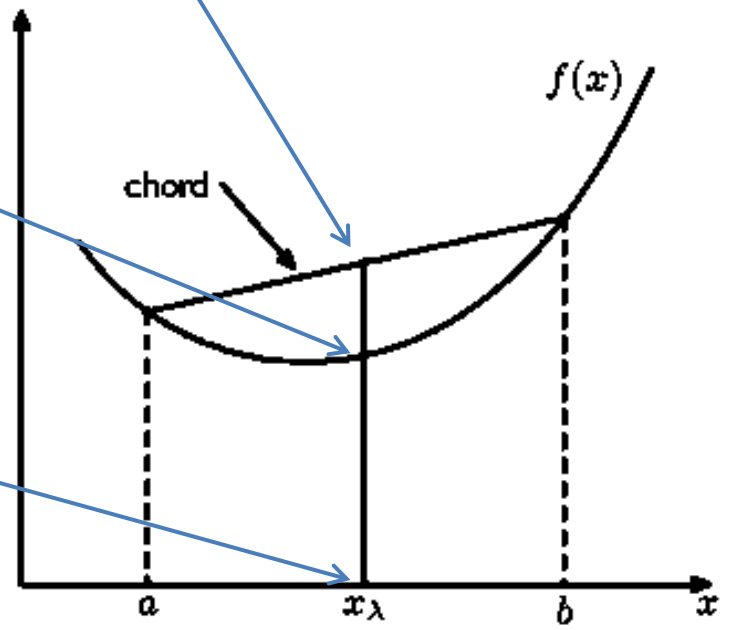
- 対称な量ではない　$\mathrm{KL}(p \,\|\, q) \neq \mathrm{KL}(q \,\|\, p)$
- 右を満たす　$\mathrm{KL}(p \,\|\, q) \geq 0$

凸関数

- 凸性は次式で表現される

$$f(\lambda a + (1-\lambda)f(b)) \leq \lambda f(a) + (1-\lambda)f(b)$$

$$\lambda a + (1-\lambda)b \qquad 0 \leq \lambda \leq 1$$



- 凸関数(convex function)
  f(x)は全ての弦(chord)が関数に乗っているか
  それより上にあるとき凸である

凸関数

- 凸関数の例 : $x \ln x (x > 0), x^2$
- λ=0とλ=1だけで成立→真に凸(strictry convex)
- 凸性と逆の性質→凹関数(concane function)
- f(x)が凸関数→-f(x)は凹関数

イェンセンの不等式

$$f\left(\sum_{i=1}^{M}\lambda_i x_i\right) \leq \sum_{i=1}^{M}\lambda_i f(x_i) \qquad \lambda_i \geq 0 \quad \sum_i \lambda_i = 1$$

個凸関数f(x)は任意の点集合{xi}に対して上を満たす

λiを値{xi}を取る離散確率変数x上の確率分布とすると

$$f\left(\mathbb{E}[x]\right) \leq \mathbb{E}[f(x)]$$

連続変数に対しては

$$f\left(\int \mathbf{x}p(\mathbf{x})\mathrm{d}\mathbf{x}\right) \leq \int f(\mathbf{x})p(\mathbf{x})\mathrm{d}\mathbf{x}$$

イェンセンの不等式

- KLダイバージェンスに適用すると

(-ln xが凸関数であること、規格化条件 $\int q(\mathbf{x})\mathrm{d}\mathbf{x} = 1$ を使って)

$$\mathrm{KL}(p \,||\, q) = -\int p(\mathbf{x}) \ln\left\{\frac{q(\mathbf{x})}{p(\mathbf{x})}\right\}\mathrm{d}\mathbf{x} \geq -\ln \int q(\mathbf{x})\mathrm{d}\mathbf{x} = 0$$

- 等式は全てのxについてp(x)=q(x)のときのみ成り立つ
- KLダイバージェンスを2つの分布p(x)とq(x)の隔たりの尺度として扱える

データ圧縮と密度推定

- 最も効率的な圧縮←真の分布が分かる
- 真の分布でない分布→非効率な符号化
  追加情報量の平均≧2分布間のKLダイバージェンス

データ圧縮と密度推定

- データが未知の分布p(x)から生成されるとき
  それをモデル化

→可変パラメータθをもつパラメトリックな分布q(x|θ)で近似

→θの決定：p(x)とq(x|θ)間のKLダイバージェンスをθで最小化

p(x)から得られた有限個の訓練点の集合Xn(n=1,…N)で近似

$$\mathrm{KL}(p \,\|\, q) \simeq \frac{1}{N} \sum_{n=1}^{N} \{-\ln q(\mathbf{x}_n|\boldsymbol{\theta}) + \ln p(\mathbf{x}_n)\}$$

訓練集合で評価した分布q(x|θ)下でのθの負の対数尤度

→KLダイバージェンスの最小化=尤度の最大化

相互情報量

- 2つの変数集合x,yの同時分布p(x,y)

→独立なら周辺分布の積に分解p(x,y)=p(x)p(y)

- 独立でない→どれぐらい独立に「近い」か？

→同時分布と周辺分布の積の間のKLダイバージェンスを考える

$$\begin{aligned} \mathrm{I}[\mathbf{x},\mathbf{y}] &\equiv \mathrm{KL}(p(\mathbf{x},\mathbf{y}) \,\|\, p(\mathbf{x})p(\mathbf{y})) \\ &= -\int\int p(\mathbf{x},\mathbf{y}) \ln\left(\frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x},\mathbf{y})}\right) \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} \end{aligned}$$

KLダイバージェンスの性質から $I(\mathbf{x},\mathbf{y}) \geq 0$

相互情報量

- 相互情報量は条件付きエントロピーと関係

  (確率の加法・乗法定理より)

$$\mathrm{I}[\mathbf{x},\mathbf{y}] = \mathrm{H}[\mathbf{x}] - \mathrm{H}[\mathbf{x}|\mathbf{y}] = \mathrm{H}[\mathbf{y}] - \mathrm{H}[\mathbf{y}|\mathbf{x}]$$

- yの値を知ることでxに関する不確実性がどれだけ減少するかを表す
- p(x):xの事前分布

  p(x|y):新たなデータyを観測した後の事後分布

→新たなyを観測した結果としてxに関する不確実性が減少した度合い