目　次

内容

１．情報理論 2
１．１　情報量 2
１．２　エントロピー 2
１．３　交差エントロピー（クロスエントロピー） 2
１．４　KL ダイバージェンス 3
１．５　連続型確率分布 4
１．６　マルチヌーイ分布 5

１．情報理論

１．１　情報量

　事象 A が起こる確率を P(A)とする。この時、事象 A が起こることの自己情報量は、

$I(A)=-\log_2 P(A)$

１）小さな確率の事象が起こることを、大きな情報量で表現したい

　式が単調減少であること

２）複数の事象が発生する確率は、積で表現されるが、情報量においては和で表現したい

　事象 A,B が独立である時

$$
\begin{aligned}
I(A\cap B) &= -\log_2 P(A\cap B)\\
&= -\log_2 P(A)P(B)\\
&= -\log_2 P(A)-\log_2 P(B)\\
&= I(A)+I(B)
\end{aligned}
$$

確率で起きる事象の情報量は$-\log_2 1=0$

■事象 A の起こる確率が P(A)の時、事象 A が起こることの情報量は$-\log_2 P(A)$

■

１．２　エントロピー

　離散確率変数 X において、X=x となる確率が p(x)で与えられたとする。

確率変数 X のエントロピーは？

$H(X)=-\sum_x p(x)\log_2 p(x)$

エントロピー（平均情報量）は、情報量（すなわち、「事象の起こりにくさ・珍しさ」）の期待値で与えられる。そのため、確率変数のランダム性の指標として用いられる。

■事象の集合Ωについて $A\in\Omega$が起こる確率を P(A)とすると、P は確率分布であるとみなせる。確率分布 P のエントロピーは

$$H(A)=-\sum_{A\in\Omega} P(A)\log_2 P(A)$$

１．３　交差エントロピー（クロスエントロピー）

　２つの確率分布 p(x)と q(x)の交差エントロピーは

$H(p,q)=-\sum_x p(x)\log_2 q(x)$

分類問題を解くための損失関数として用いられる。

２つの確率分布が全く同じときに交差エントロピーが最小になる。

ここで p がデータによって近似される真の分布であり、q がモデルの分布である。

１．４　KL ダイバージェンス

２つの確率分布 p(x)と q(x)に対して

$D(p||q)=\sum_x p(x)\log_2 (p(x)/q(x))$

KL（カルバック・ライブラー）ダイバージェンスと呼ぶ

KL ダイバージェンスは、２つの確率分布の近さを表現する最も基本的な量で、統計学・情報理論において非常に重要な役割を果たすものである。

・p のエントロピーを H(p)

・p と q の交差エントロピーを H(p,q)

・p と q の KL ダイバージェンスを $D_{KL}(p||q)$

$$
\begin{aligned}
H(p)+D_{KL}(p||q) &= -\Sigma p(x)\log_2 p(x)+\Sigma p(x)\log_2 (p(x)/q(x)) \\
&= -\Sigma p(x)\log_2 p(x)+\Sigma p(x)\log_2 p(x)-\Sigma p(x)\log_2 q(x) \\
&= -\Sigma p(x)\log_2 q(x) \\
&= H(p,q)
\end{aligned}
$$

p と q が全く同じ分布である時に交差エントロピーは最小になり、同時に KL ダイバージェンスは最小になる。

このとき

$$
\begin{aligned}
D_{KL}(p||p) &= \Sigma p(x)\log_2 (p(x)/p(x)) \\
&= \Sigma p(x)\log_2 1 \\
&= 0
\end{aligned}
$$

■２つの確率分布 p(x)と q(x)に対して JS ダイバージェンスは次のようになる

$D_{JS}(p||p)=1/2(\Sigma p(x)\log_2 (p(x)/r(x)))+\Sigma q(x)\log_2 (q(x)/r(x))$

ただし

$r(x)=(p(x)+q(x))/2$

JS ダイバージェンスは KL ダイバージェンスとは異なり、２つの分布に対して対称な量である。

JS ダイバージェンスは敵対的ネットワークの損失関数に用いられる。

■一般化 KL ダイバージェンスや IS ダイバージェンスは、正規化されていない（合計が１にならない）ような分布同士の近さを測る量である。

どちらも非負値行列因子分解の損失関数として、音響信号処理の領域でよく用いられる。

１．５　連続型確率分布

データ D={x1,x2,…,xn}が、p(x)を確率密度関数とする連続型確率分布に独立に従っている。

この時、モデル q(x;θ)によって、交差エントロピーが最小になるように p(x)を推定することを考える。

１）P(x)と q(x;θ)の交差エントロピーは

連続型確率分布 p(x)と q(x;θ)の交差エントロピーは

$$H(p,q)=-\int_x p(x)\log q(x;\theta)dx$$

エントロピー、KL ダイバージェンスについても同様の拡張が可能である。

２）真の分布 p(x)での期待値をデータ D による平均に置き換えた量は

モンテカルロ積分

$$\tilde{H}(p,q)=-1/n\sum_{i=1}^{n}\log q(ki;\theta)$$

これに置き換えることで、実際にはわからない真の分布 p(x)を含むような計算を回避している。

３）尤度関数は $L_D(\theta)=\Pi q(x,\theta)$であるから、等式が成り立つ

$$\begin{aligned}\tilde{H}(p,q)&=-1/n\Sigma\log q(xi;\theta)\\&=-1/n\log\Pi q(xi;\theta)\\&=-1/n\log L_D(\theta)\end{aligned}$$

交差エントロピーが最小となる推定は､「負の対数尤度が最小となる尤度」すなわち、最尤推定と等価である。

■連続型確率分布 p のエントロピーは

$$H(p)=-\int p(x)\log p(x)dx$$

■連続型確率分布 p と q の交差エントロピー

$$H(p,q)=-\int p(x)\log q(x)dx$$

■連続型確率分布 p と q の KL ダイバージェンス

$$D_{KL}(p||q)=-\int p(x)\log(q(x)/p(x))dx$$

■定義域[a,b]の実数値確率変数 X において、X=x となる確率密度関数を p(x)と書くとき、関数 f(X)の期待値は

$$E[f(X)]=\int_a^b p(x)f(x)dx$$

■確率変数 X の観測として独立なデータ D={x1,x2,…,xn}が与えられたとき、E[f(X)]はモンテカルロ積によって近似できる

$$E[f(X)]\fallingdotseq 1/n\sum_{i=1}^{n}f(xi)$$

１．６　マルチヌーイ分布

　k 次元ワンホットベクトルが従う確率分布 p(x)を、マルチヌーイ分布

$$q(x;\mu)=\prod_{j=1}^{k}\mu j^{xj}$$

によって推定することを考える。

ただし、x,μ の j 成分を xj,μj とかく。

この時 p(x)と q(x;μ)の交差エントロピーは

$$H(p,q)=-\sum_{x} p(x)\log q(x;\mu)$$

$$=-\Sigma\, p(x)\log\prod_{j=1}^{k}\mu j^{xj}$$

$$=-\Sigma\, p(x)\sum_{j=1}^{k}\log\mu j$$

$$=-\Sigma\, p(x)\,\Sigma\, xj\log\mu j$$

モンテカルロ積分を用いれば

$$H(p,q)\fallingdotseq -1/n\sum_{i=1}^{n}\sum_{j=1}^{k}xij\log\mu j$$

ここで μ をロジスティック回帰やニューラルネットワークの出力、xi を正解ラベルと置き換えれば、交差エントロピーは分類問題において典型的な損失関数となる。