第一章線形代数　関連記事

１．線形代数

１．１　固有値分解１

$A=\begin{bmatrix}5 & 2\\ 2 & 8\end{bmatrix}$　対称行列の固有値分解

公式Ａ－λＩ＝０より固有値を求める

（1）固有値

$\begin{bmatrix}5-\lambda & 2\\ 2 & 8-\lambda\end{bmatrix}=(5-\lambda)(8-\lambda)-2*2=40-13*\lambda+\lambda*\lambda-4=\lambda*\lambda-13*\lambda+36=0$

$13\mp\sqrt{(169-4*1*36)}$　　$13\mp\sqrt{25}$

$2*1$　　$=2$　　$=9,4$

（2）固有ベクトル

$\begin{bmatrix}5 & 2\\ 2 & 8\end{bmatrix}\begin{bmatrix}x1\\ x2\end{bmatrix}=9\begin{bmatrix}x1\\ x2\end{bmatrix}$

5*x1+2*x2=9*x1	-4*x1+2*x2=0	-2*x1+x2=0	x1=1/2*x2	$\begin{bmatrix}1\\ 1/2\end{bmatrix}$
2*x1+8*x2=9*x2	2*x1-x2=0		x1=1/2*x2	

$\begin{bmatrix}5 & 2\\ 2 & 8\end{bmatrix}\begin{bmatrix}x1\\ x2\end{bmatrix}=4\begin{bmatrix}x1\\ x2\end{bmatrix}$

5*x1+2*x2 = 4 *x1	x1+2*x2=0	x1=-2*x2	$\begin{bmatrix}1\\ -2\end{bmatrix}$
2*x1+8*x2 = 4*x2	2*x1+4*x2=0	x1=-2*x2	

$D=\begin{bmatrix}9 & 0\\ 0 & 4\end{bmatrix}$

◎一般にｎ次正方行列Ａに対して Av＝λｖを満たすようなスカラーλを固有値、非零ベクトルを固有ベクトルという。

◎固有値λは、Ａの固有方程式 det(λ-A)=0 の解となっている。

◎Ａの固有値λ1,…..,λｎを対角成分に持つ対角行列

$$D=\begin{bmatrix}\lambda 1 & 0 & 0 & 0\\ 0 & \ddots & & \\ 0 & & \ddots & \\ 0 & 0 & 0 & \lambda n\end{bmatrix}$$

と、対応する固有ベクトルを列として並べた行列

P=(v1,….,vn)

に対いて、$A=PDP^{-1}$ が成立する。これを行列Ａの「固有値分解」、又は「対角比」と呼ぶ。

さらにＡが対称行列であるとき、対応する大きさ１の固有ベクトルを列として並べた行列Ｐは直行行列（$P^{-1}=P^{T}$ ）になるため、$A=PDP^{T}$ が成立する。

１．２　固有値分解２

$$A=\begin{bmatrix}0 & 0 & 1\\ 0 & 0 & -1\\ 1 & -1 & 1\end{bmatrix}$$　対称行列の固有値分解

公式Ａ－λＩ＝０より固有値を求める

（1）固有値

$$\begin{bmatrix}-\lambda & 0 & 1\\ 0 & -\lambda & -1\\ 1 & -1 & 1-\lambda\end{bmatrix} = -\lambda\begin{bmatrix}-\lambda & -1\\ -1 & 1-\lambda\end{bmatrix} -0\begin{bmatrix}0 & 1\\ -1 & 1-\lambda\end{bmatrix} + 1\begin{bmatrix}0 & 1\\ -\lambda & 1\end{bmatrix} = -\lambda*((-\lambda+\lambda*\lambda)-(1))+1*(\lambda)=-\lambda*(\lambda*\lambda-\lambda-1)+\lambda$$

$=-\lambda*\lambda*\lambda+\lambda*\lambda+2*\lambda=0$

$\lambda*\lambda*\lambda-\lambda*\lambda-2*\lambda=0$

$\lambda(\lambda*\lambda-\lambda-2)=\lambda((\lambda+1)(\lambda-2))=0$

$\lambda=-1,0,2$

$$\begin{bmatrix}0 & 0 & 1\\ 0 & 0 & -1\\ 1 & -1 & 1\end{bmatrix}\begin{bmatrix}x1\\ x2\\ x3\end{bmatrix} = 2\begin{bmatrix}x1\\ x2\\ x3\end{bmatrix}$$

X3=2x1	x1=1/2*x3	1
-x3=2x2	x2=-1/2*x3	-1
X1-x2+x3=2x3	2	

（2）固有値λ=2 に対する固有ベクトル v=(1/√6,vy,2/√26)

$$\begin{bmatrix}0 & 0 & 1\\ 0 & 0 & -1\\ 1 & -1 & 1\end{bmatrix}\begin{bmatrix}1/\sqrt{6}\\ vy\\ 2/\sqrt{6}\end{bmatrix} = 2\begin{bmatrix}1/\sqrt{6}\\ vy\\ 2/\sqrt{6}\end{bmatrix}$$

$$\begin{bmatrix}2/\sqrt{6}\\ -2/\sqrt{6}\\ 1/\sqrt{6}-vy+2/\sqrt{6}\end{bmatrix} = \begin{bmatrix}2/\sqrt{6}\\ 2vy\\ 4/\sqrt{6}\end{bmatrix}$$　$2vy=-2/\sqrt{6}$　$vy=-1/\sqrt{6}$

◎３次正方行列の行列式は

$$\det\begin{bmatrix}a11 & a12 & a13\\ a21 & a22 & a23\\ a31 & a32 & a33\end{bmatrix} = a11a22a33+a12a23a31+a13a21a32-a11a23a32-a12a21a33-a13a22a31$$

これをサラスの公式を呼称する。

◎高次方程式の解には因数分解を用いる。一般に高次方程式 f(x)=0 が $x=\alpha$ を解に持つとき、

f(x)は $x-\alpha$ を因数に持つ。

１．３　固有値分解３

$A=\begin{bmatrix}1 & 2\\ 2 & 1\end{bmatrix}$　対称行列の固有値分解

（1）固有値

$\begin{bmatrix}1-\lambda & 2\\ 2 & 1-\lambda\end{bmatrix}=(1-\lambda)(1-\lambda)-4=1-\lambda-\lambda+\lambda*\lambda-4=0$

$\lambda*\lambda-2*\lambda-3=(2\mp\sqrt{4-4*1*(-3)})/2=(2\mp\sqrt{16})/2=3,-1$

固有値ベクトル

$\begin{bmatrix}1 & 2\\ 2 & 1\end{bmatrix}\begin{bmatrix}x1\\ x2\end{bmatrix}=3\begin{bmatrix}x1\\ x2\end{bmatrix}$

X1+2x2=3x1　2x1=2x2　x1=x2　1

2x1+x2=3x2　2x1=2x2　1

$\begin{bmatrix}1 & 2\\ 2 & 1\end{bmatrix}\begin{bmatrix}x1\\ x2\end{bmatrix}=-1\begin{bmatrix}x1\\ x2\end{bmatrix}$

X1+2x2=-x1　x1=-3x2　1

2x1+x2=-x2　x1=-x2　1

１．４　特異値分解１

行列

$A=\begin{bmatrix}1 & 0 & 1\\ -1 & 1 & 0\end{bmatrix}$

を特異値分解し $A=U\Sigma V^{T}$ の形で表す

$AA^{T}=\begin{bmatrix}1 & 0 & 1\\ -1 & 1 & 0\end{bmatrix}\begin{bmatrix}1 & -1\\ 0 & 1\\ 1 & 0\end{bmatrix}=\begin{bmatrix}1+0+1 & -1+0+0\\ -1+0+0 & 1+1+0\end{bmatrix}=\begin{bmatrix}2 & -1\\ -1 & 2\end{bmatrix}$

（1）固有値分解

$\begin{bmatrix}2 & -1\\ -1 & 2\end{bmatrix}=\begin{bmatrix}2-\lambda & -1\\ -1 & 2-\lambda\end{bmatrix}=(2-\lambda)(2-\lambda)-1=4-2\lambda-2\lambda+\lambda*\lambda-1=\lambda*\lambda-4\lambda+3$

$\lambda=(4\mp\sqrt{(16-4*3)})/2=3,1$

特異値は$\sqrt{3},\sqrt{1}$

◎任意の零行列でない m×n 行列 A に対して、$Av=\sigma u$,$A^{T}u=\sigma v$ を満たすような正の数 σ を特異値、m 次元ベクトル u を左特異ベクトル、n 次元ベクトル v を右特異ベクトルと呼ぶ。

また、定義式から $A^{T}Au=\sigma^{2}u$,$AA^{T}v=\sigma^{2}v$ が得られるので、行列 AA^{T} と $A^{T}A$ の固有値ベクトルを求めることで、行列 A の特異値・（左右）特異ベクトルを知ることができる。

このようにして求められた特異値を大きい順に(i,j)で成分に並べ、その他の成分を 0 で埋めた m×n 行列 Σ と、左特異ベクトルを列として横に並べた m 次正方行列 U、右特異ベクトルを列として横に並べた n 次正方行列 V を用いて $A=U\Sigma V^{T}$ と書ける。これを A の特異値分解と呼ぶ。なお特異値を降順に並べることにすると Σ は一意に定まるが U 及び V は一意に定まらない。

１．５　特異値分解２

行列

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 2 \end{bmatrix}$$

を特異値分解し $A=U\Sigma V^T$ の形で表す

$$AA^T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 2 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 2 \end{bmatrix} = \begin{bmatrix} 1+0+0 & 0 \\ 0 & 0+1+4 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 5 \end{bmatrix}$$

（1）固有値分解

$$\begin{bmatrix} 1-\lambda & 0 \\ 0 & 5-\lambda \end{bmatrix} = (1-\lambda)(5-\lambda)=0 \qquad \lambda=1,5$$

特異値は $1,\sqrt{5}$

（2）固有値ベクトル

$$\begin{bmatrix} 1 & 0 \\ 0 & 5 \end{bmatrix}\begin{bmatrix} x1 \\ x2 \end{bmatrix} = 5\begin{bmatrix} x1 \\ x2 \end{bmatrix}$$

X1=5*x1

5*x2=5*x2

$$\begin{bmatrix} 1 & 0 \\ 0 & 5 \end{bmatrix}\begin{bmatrix} x1 \\ x2 \end{bmatrix} = \begin{bmatrix} x1 \\ 1\ x2 \end{bmatrix}$$

X1=1*x1

5*x2=1*x2

（3）

A の特異値は $\sigma=1,\sqrt{5}$ なので、これを降順に並べれば

$$\Sigma = \begin{bmatrix} \sqrt{5} & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

U の第一列は AA^T の固有値 5 に対応する固有ベクトル $u=(ux,uy)^T$ のうち、大きさが 1 のもの

$$\begin{bmatrix} 1\ 0 & ux \\ 0\ 5 & uy \end{bmatrix} = 5\begin{bmatrix} ux \\ uy \end{bmatrix}$$

ux=5ux　　ux=0,uy は任意の実数

5uy=5uy

V の各列は A^TA の固有ベクトルだが、これらは正規直行性を満たす必要がある。

$v1=(0,1/\sqrt{5},2/\sqrt{5}),v3=(0,-2/\sqrt{5},vz)$ と置けばそれらの直交性から $v1^Tv3=0$ になる。

$0*0+1/\sqrt{5}*-2/\sqrt{5}+2/\sqrt{5}*vz=0$

$Vz=1/\sqrt{5}$

１．６　L1 ノルム、L2 ノルム

（1）Ｌ１ノルム

$$x\ 1 = \sum_{i=1}^{n} |xi|$$

（2）Ｌ２ノルム

$$x\ 2 = \sqrt{\sum_{i=1}^{n} xi(2)}$$

（3）最大値ノルム

$$x\ \infty = \max_{I} |xi|$$

１．７　距離

（1）マンハッタン距離

$$d(x,y) = \sum_{i=1}^{n} |xi-yi|$$

（2）ユークリッド距離

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (xi-yi)2\text{ 乗}}$$

（3）マハラノビス距離

$$D(x,y) = \sqrt{\Sigma (x-y)T\Sigma\text{-}1\text{ 乗}(x-y)}$$

第二章 確率・統計　関連記事

１．確率・統計

１．１　ベルヌーイ分布１

確率ｐで表が出て、確率１－ｐで裏が出るコイン投げを考える。

表が出た時にX=1をとり、裏が出た時にX=０を取るとする。

（1）X=x(ただし、x=0,1)となる確率

$p^x (1-p)^{1-x}$ となる。

これをベルヌーイ分布という。

（2）ベルヌーイ分布に従う確率変数の期待値

$$E[X]=\sum_{x=0}^{1} x p^x (1-p)^{1-x} = p$$

（3）分散

$$Var[X]=E[X^2]-E[X]^2 = \sum_{x=0}^{1} x^2 p^x (1-p)^{1-x} - p^2 = p-p^2 = p(1-p)$$

◎ベルヌーイ分布の確率関数は、$P(X=x)=p^x (1-p)^{1-x}$

で与えられ、その期待値はｐ、分散はp(1-p)である。

◎離散確率変数Ｘに対して、P(X=x)=f(x)である時、f(x)を確率関数（又は確率分布）と呼ぶ。

この時Ｘの期待値E[X]は

$E[X]=\Sigma f(x)$で定義され

分散Var[X]は

$Var[X]=E[(X-E[X])^2]$で定義される

$Var[X]=E[X^2]-E[X]^2$

１．２　ベルヌーイ分布２

{0,1}を取りうる２値データ D={x1,….,xn}がベルヌーイ分布 f(x;p)=px 乗(1-p)(1-x)乗に独立に従うと仮定する。このとき最尤法によりパラメータ p を決定する。

（1）尤度関数

$$LD(p)=\prod_{i=1}^{n} f(xi;p)=\prod_{i=1}^{n} pxi \text{ 乗}(1-p)(1-xi)\text{乗}$$

（2）負の対数尤度

$$-\log LD(p)=-\log\prod_{i=1}^{n} f(xi;p)=-\sum_{i=1}^{n}\log f(xi;p)=-\sum_{i=1}^{n}\log pxi \text{ 乗}(1-p)(1-xi)\text{乗}=-\sum_{i=1}^{n}(xi\log p+(1-xi)\log(1-p))$$

この式は、２クラス分類のニューラルネットワークの学習に適応されることの多い損失関数

（3） p の最尤推定量

負の対数尤度の最小値が満たすべき必要条件は、

$$\frac{d}{dp}(-\log LD(p))=0$$

の解である。

-logp と-log(1-p)は凸関数なので、-logLD(P)も凸関数であり、解は唯一で最小解である。

$$\frac{d}{dp}(-\log LD(p))=-\sum_{i=1}^{n}(xi\frac{d}{dp}\log p+(1-xi)\frac{d}{dp}\log(1-p))=-\sum_{i=1}^{n}(zi*1/p-(1-xi)*1/(1-p))=-\sum_{i=1}^{n}(xi-p)/p(1-p)$$

$$=-1/p(1-p)(\sum_{i=1}^{n}xi-p\sum_{i=1}^{n}1)=1/p(1-p)(np-\sum_{i=1}^{n}xi)$$

$$1/p(1-p)(np-\sum_{i=1}^{n}xi)=0$$

$$\hat{p}=1/n\sum_{i=1}^{n}xi$$　　$\hat{p}$ はデータによって決定されるパラメータ p の推定量

・最尤推定量

「尤度関数が最大になる（負の対数尤度関数が最小になる）」ように決められる「確率分布がデータに最もよく当てはまる」可能性があるパラメータの推定量

◎パラメータ θ によって定まる確率（密度）関数 $f(x;\theta)$に対して、それに独立に従うと仮定されるデータ D={x1,x2,….,xn}が与えられたとき、尤度関数は

$$LD(\theta)=\prod_{i=1}^{n} f(xi;\theta)$$

で定義される。

最尤推定量とは尤度関数を最大化、すなわち負の対数尤度関数$-\log LD(\theta)$を最小化するパラメータの推定量

$\theta=\hat{\theta}$

◎ベルヌーイ分布 f(x;p)=p　(1-p)　のパラメータ p の最尤推定量は、データの平均、

すなわち「１が出現する確率」で与えられる。

１．３　マルチヌーイ分布(カテゴリカル分布)

（1）ワンホットベクトル

ベクトルｘの成分のうち、ただ１つが１であり、その他が全て０であるようなもの

（2）マルチヌーイ分布

ｋ次元のワンホットベクトルで構成されるデータ

D={x1,x2,….,xn}がマルチヌーイ分布

$$F(x;p)=\prod_{j=1}^{k} pj^{xj}$$

に従っていると仮定する。Xi の j 成分を xij と書く。

ただし、

$p=(p1,p2,….,pk)^T$ 、$\sum_{j=1}^{k} pj=1$、$0 \leqq pj \leqq 1 (j=1,….,k)$

出る目の確率が

$p=(p1,p2,p3,p4,p5,p6)^T$　(ただし$\Sigma pj=1$、$0 \leqq p1,….,p6 \leqq 1$)

で与えられる歪んだ６面サイコロについて、１が出る事象を$(1,0,0,0,0,0)^T$、２が出る事象を$(0,1,0,0,0,0)^T$、…

６が出る事象を$(0,0,0,0,0,1)^T$、と、ワンホットベクトルの確率変数によって表現するすると、これに従う確率分布は

$$f(x;p)=\prod_{j=1}^{6} pj^{xj}$$

と書ける。

マルチヌーイ分布は３値以上のカテゴリ変数に関するモデリングに用いるのに適した分布である。

ここで、尤度関数は確立関数にデータの各値を代入したものの積で書けるので、

この時尤度関数

$$LD(p)=\prod_{i=1}^{n} f(xi;p)=\prod_{i=1}^{n}\prod_{j=1}^{k} pj^{xij}$$

負の対数尤度

$$-\log LD(p)=-\log \prod f(xi;p)$$

$$=-\Sigma \log \prod pj^{xij}$$

$$=-\Sigma \Sigma \log pj$$

$$=-\Sigma \Sigma xij \log pj$$

この式は、分類問題のためのニューラルネットワークの学習に適応されることの多い損失関数である公差エントロピーである。

◎ベクトルｘの成分のうち、ただ１つが１であり、その他が全て０であるようなものをワンホットベクトルと呼ぶ。

◎ワンホットベクトルの従う確率分布としてマルチヌーイ分布があり、その確率関数はパラメータベクトル p を用いて

$F(x;p)=\prod pj$　（ただし、$\Sigma pj=1, 0 \leqq pj \leqq 1$、$j=1,….,k$）

により与えられる。

◎マルチヌーイ分布の最尤推定は、交差エントロピーの最小化に対応している。
◎マルチヌーイ分布の負の対数尤度
$-\log LD(p)=-\Sigma\Sigma k_{ij}\log p_j$ を
制約（$\Sigma p_j=1, 0\leqq p_j\leqq 1$、$j=1,\ldots,k$）のもとで最小化する問題を解き、p の最尤推定量を p と求めることができる。
求解にはラグランジェの未定乗数法を用いる。
つまり、ベルヌーイ分布と同様パラメータ p の最尤推定量はデータの平均、すなわち「各次元において 1 が出現する頻度」で与えられる。

１．４　一変量正規分布

平均が μ、分散が１の１変量正規分布の確率密度関数は、

$$f(x;\mu)=1/\sqrt{2\pi}\exp(-1/2(x-\mu)^2)$$

よって与えられる。

この時、実数データ D={x1,x2,…,xn}によって与えられ、データ D が平均 μ、分散１の１変量正規分布に独立に従うとして、最尤推定によってパラメータ μ を求める。

（1）尤度関数

１変量正規分布は、実数値確率変数が従う代表的な確率分布の１つで、左右対称なベルカーブを描く。

①．確率密度関数

$f(x;\mu,\sigma^2)=1/\sqrt{2\pi\sigma^2}\exp(-1/(2\sigma^2)*(x-\mu)^2)$

で与えられる。

②．分散 $\sigma^2=1$ で固定されているため、密度関数は

$$f(x;\mu)=1/\sqrt{2\pi}\exp(-1/2(x-\mu)^2)$$

③．尤度関数

$$L(\mu)=\prod_{i=1}^{n}f(xi;\mu)$$

$$=\prod_{i=1}^{n}(1/\sqrt{2\pi})*\exp(-1/2(xi-\mu)^2)$$

最尤推定では L(μ)の最大化問題を考えるが、ベルヌーイ分布と同じように、「負の対数尤度の最小化問題」に書き換える。

④．負の対数尤度

$$-\log L(\mu)=-\log(\prod(1/\sqrt{2\pi})\exp(-1/2(xi-\mu)^2))$$

$$=-\Sigma\log((1/\sqrt{2\pi})\exp(-1/2(xi-\mu)^2))$$

$$=-\Sigma(\log(1/\sqrt{2\pi})-1/2(xi-\mu)^2)$$

$$=-n\log(1/\sqrt{2\pi})+1/2\Sigma(xi-\mu)^2$$

実際に最小化する関数は

$$g(\mu)=1/2\Sigma(xi-\mu)^2$$

これは、回帰問題のモデルの学習に用いられる損失関数である、二乗和誤差に対応している。

ここで g(μ)は凸な二次関数の和なので、

$$\frac{d}{d\mu}g(\mu)=0$$

を解けば最小解が求められる。

$$\frac{d}{d\mu}g(\mu)=1/2\Sigma\frac{d}{d\mu}(xi-\mu)^2$$

$$=1/2\Sigma(-2(xi-\mu))$$

$$=\Sigma\mu-\Sigma xi$$

$$=n\mu-\Sigma xi$$

μ の最尤推定量、μ=1/n Σxi が得られる。
・「正規分布の最尤推定量は二乗和誤差の最小化」
・「正規分布の平均の最尤推定量はデータの平均」

◎平均が μ、分散が σ^2 である一変量正規分布は、次のようになる

$$f(xi;\mu, \sigma^2 = 1/\sqrt{2\pi\ \sigma^2} \exp((-1/2\ \sigma^2)(x-\mu)^2\ )$$

◎正規分布をモデルとした最尤推定は、二乗和誤差の最小化に対応している。
◎正規分布の平均の最尤推定量は、データの平均によって与えられる。
◎平均ベクトルが μ、分散共分散行列が Σ である多変量正規化分布の確率密度関数は、

$$f(x;\mu, \Sigma)=1/((\sqrt{2\pi})^n\ \sqrt{\det(\Sigma)})\exp(-1/2(x-\mu)^T\ \Sigma^{-1}\ (x-\mu))$$

いま、分散が単位行列で与えられているとする（つまり、各変数間が無相関で、各変数の分散が 1 であることを意味する）。
このとき確率密度関数は、

$$f(x;\mu,)=1/((\sqrt{2\pi})^n\ )\exp(-1/2(x-\mu)^T\ (x-\mu))$$

これについて、負の対数尤度を導出すると、一変数の場合と同様に多変量の場合の二乗和誤差を導出することができる。

１．５　条件付き確率（ベイズの定理）

事象 A の起こる確率を P(A)、事象 B の起こる確率を P(B)、これらの同時確立を P(A,B)と書く。

(1) 条件付き確率の定義から P(A|B)は

=P(A,B)/P(B)

(2) P(B|A)

=P(A,B)/P(A)

(3) ベイズの定理 P(A|B)

(1) より P(A,B)=P(A|B)P(B)

(2) より P(A,B)=P(B|A)P(A)

P(A|B)P(B)=P(B|A)P(A)

P(A|B)=(P(B|A)P(A))/P(B)

P(B|A)=(P(A|B)P(B))/P(A)

１．６　母集団

母集団に属する人が疾患 X に罹患している確率を 0.010 とする。簡易検査薬 Y は、疾患 X に感染している人に適用した場合に確率 0.90 で陽性を示し、疾患 X に感染していない人に適用した場合に確率 0.10 に陽性を示すことが知られている。母集団に属する人のうち、ある 1 名 Z に対して簡易検査薬 Y を適用したところ、陽性を示した。この時、Z が疾患 X に罹患している確率は？

（1）母集団に属する人が疾患 X に罹患している確率 P(罹患)=0.010

（2）罹患していない確率 P(非罹患)=0.99

（3）疾患 X に罹患している人に対して簡易検査薬 Y が陽性を示す確率 P(陽性|罹患)=0.90

（4）疾患 X に罹患していない人に対して簡易検査薬 Y が陽性を示す確率 P(陽性|非罹患)=0.10

（5）疾患 X に「罹患していること」と「罹患していないこと」は排反事象

P(陽性)=P(陽性,罹患)+P(陽性,非罹患)
=P(陽性|罹患)P(罹患)+P(陽性|非罹患)P(非罹患)
=0.90×0.010+0.10×0.99
P(罹患|陽性)=P(陽性|罹患)P(罹患)/P(陽性)
=0,90×0.010/(0.90×0.010+0.10×0.99)

第三章　情報理論　関連記事

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 $q(x;\theta)$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 $q(x;\theta)$の交差エントロピーは

連続型確率分布 p(x)と $q(x;\theta)$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

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

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

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

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

$$H(p,q)=-\sum_x p(x)\log q(x;\mu)$$

$$=-\sum p(x)\log\prod_{j=1}^{k}\mu_j^{x_j}$$

$$=-\sum p(x)\sum_{j=1}^{k}\log\mu_j$$

$$=-\sum p(x)\sum x_j\log\mu_j$$

モンテカルロ積分を用いれば

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

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。