目　次

内容

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」ような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; θ)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-logLD(θ)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

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

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

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

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

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、$\hat{\mu}=1/n\,\Sigma xi$ 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)