作成日：2019 年 9 月 11 日
輪講予定日：2019 年 9 月 19 日
慶應義塾大学理工学部物理学科
岡崎健人

卒研ゼミ「深層学習」

3　確率と情報理論

3.1　なぜ確率なのか（要約）

コンピュータ科学ではほとんどの場合、完全に決定論的で、確実な対象を扱うが、機械学習では不確実・非決定論的な量を扱わなければならない。不確実性を生み出す要因は３つある。１つめは扱う対象そのものが確率性を持つ場合（例：量子力学や完璧にシャッフルされたカードなど）、２つめは扱う対象の振る舞いを決める変数を観測できない・しない場合（例：モンティ・ホール問題において回答者の回答が当たりかハズレかは決定論的に決まっているが、ドアを開けるまで結果は不確実である）、最後は観測した情報を破棄した場合（例：物体の位置を離散化したとき、正確な位置は不確実となる）である。

確実だが複雑な規則よりも、不確実だが単純な規則を用いるほうが実用的であるため、確率論が用いられる。

頻度確率（frequentist probability）とは、反復可能な試行を無限回くりかえしたときに、ある事象が現れる頻度という意味での確率のことである。いっぽう反復可能な試行でなくとも、信念の度合い、可能性の大きさという意味での確率をベイズ確率（Bayesian probability）という。確率論ではこれら２つの意味での確率を、同等に扱う。

確率論はある命題のもっともらしさが与えられたときに、対象の命題が真であるもっともらしさを決定するための、不確実性を扱うための論理学の拡張とみなすことができる。

3.2　確率変数

確率変数（random variable）とは起こりうる事象に対応した、ランダムな値を取れる変数のことである[*1]。確率変数を x のようにローマン体の小文字で書き[*2]、x がとりうる値は x のようにイタリック体で書く。確率変数がベクトル値のときはそれらは太字で $\mathbf{x}$, $\boldsymbol{x}$ のように書く。確率変数は離散値でも連続値でもよい。

3.3　確率分布

3.3.1　離散変数と確率質量関数

確率質量関数（probability mass function, PMF）とは、離散型確率変数に対してその値をとるときの確率を対応させた関数のことであり、大文字 P で書き表す。たとえば確率 p で表（$\mathrm{x}=1$）、確率 $1-p$ で裏

[*1] たとえばサイコロを振ったときの「6 の目が出る」という事象は $\mathrm{x}=6$ に対応させる、コイントスをしたときの「裏が出る」「表が出る」という事象をそれぞれ $\mathrm{x}=0$, $\mathrm{x}=1$ に対応させるなど。このように確率変数は、事象の集合（標本空間）から数の集合への関数と考えることができる。

[*2] 多くの本では X のようにイタリック体の大文字で書かれることが多いようです。

（$\mathrm{x}=0$）が出るようなコインでコイントスをするとき、確率質量関数は

$$P(\mathrm{x}=x)=\begin{cases} p, & (x=1) \\ 1-p & (x=0) \end{cases}$$

とかける（これはベルヌーイ分布とよばれる）。このコイントスを n 回行ったうち k 回だけ表が出る確率は

$$P(\mathrm{y}=k)=\binom{n}{k}p^k(1-p)^{n-k}$$

となる（これは二項分布とよばれる）。このとき「確率変数 y は母数 (n,p) の二項分布 $\mathrm{B}(n,p)$ に従う」といい、これを $\mathrm{y}\sim\mathrm{B}(n,p)$ と表記する。

$\mathrm{x}=x$ かつ $\mathrm{y}=y$ であるときの確率 $P(\mathrm{x}=x,\mathrm{y}=y)=P(x,y)$ は同時確率分布とよばる。

関数 P が離散型確率変数 x の確率質量関数であるためには以下の性質を満たさなければならない。

- 定義域は確率変数のとりうる値すべての集合である。
- その集合を A とすると、$\forall x\in A(0\leq P(x)\leq 1)$.
- 正規化されている（normalized）：$\sum_{\forall x\in A}P(x)=1$.

たとえば、確率変数 x が k 個の異なる離散値 $x_1,x_2,\ldots,x_k$ をとるとき、確率質量関数を

$$P(\mathrm{x}=x_i)=\frac{1}{k}$$

とすれば一様分布を定義することできる。すべての i で和をとれば

$$\sum_{i=1}^{k}P(\mathrm{x}=x_i)=\frac{1}{k}\sum_{i=1}^{k}1=\frac{1}{k}\cdot k=1$$

となって、正規化されていることがわかる。

3.3.2 連続変数と確率密度関数

離散型確率変数に対して確率質量関数とよんでいたものを、連続型確率変数に対しては確率密度関数（probability density function, PDF）とよぶ。確率密度関数 p は以下の性質を満たさなければならない。

- 定義域は確率変数のとりうる値すべての集合である。
- その集合を I とすると、$\forall x\in I(0\leq p(x))$. ただし $p(x)\leq 1$ である必要はない。
- $\int_I p(x)\,\mathrm{d}x=1$.

$\mathrm{x}=x$ であるときの確率は $p(x)$ と直接得られるわけではなく、代わりに x が微小区間 $[x,x+\delta x]$ の値をとるときの確率が $p(x)\delta x$ で与えられる。区間 $[a,b]$ に x が存在する確率は $\int_{[a,b]}p(x)\,\mathrm{d}x$ で求められる。

確率密度関数の例として、一様分布がある。

$$u(x;a,b)=\begin{cases} \dfrac{1}{b-a}, & x\in[a,b], \\ 0, & x\notin[a,b]. \end{cases}$$

積分すれば 1 になる。x が区間 $[a,b]$ において一様分布にしたがうことを、$\mathrm{x}\sim U(a,b)$ と表す。