情報量と符号化

I. ここでの目的
情報量の単位はビットで、2種の文字を持つ記号の情報量が１ビットです。ここでは、一般にn種の文字を持つ記号の情報量を定義します。次に、出現する文字に偏りがある場合の平均情報量を定義します。
　この平均情報量は、記号を適当に0,1で符号化する場合の平均符号長にほぼ等しくなることがわかります。

II. 情報量とは
　A. bit
　情報量の単位としてbitが利用されます。1bitは0か1の情報を運びます。2者択一の問題があるとき「0」か「1」などの1bitの情報があれば答えを表現できます。

　B. 三択問題
　では3択問題の場合はどうでしょう。1bitでは答えは表現できません。0,1,2の3種を区別する必要があるので、2bitが必要です。では4択問題ではどうでしょう。やはり0,1,2,3の4種が表現できれば良いので、2bitで表現できます。すると、3択問題も、4択問題も必要な情報量は同じでしょうか？

　C. 答えが推測できる場合
　今日が良い天気の場合、明日は「晴れか雨か」の問いには、多分晴れと答えるでしょう。でも、今日が雨の場合、晴れと雨の確率は五分五分でしょう。この二つの場合、答えの持つ情報量は同じでしょうか？
　3択でも、情報科学部の学生に「情報科学部の学科数は１，２，３どれでしょう」の問題を出したとき、多くは3を出すでしょう。答え１を出す確率はほとんどないでしょう。仮に、10%が１，30%が2、60%が3と答えるたしましょう。
　しかし、一般の人は余り知らないから、30%が１と３，40%が2と答えるものとしましょう。この二つの場合の答えのもつ情報量は同じでしょうか？

III. 情報量の定義（シャノンの定義）
　A. n個の選択肢からの情報量
　　1. n=2の場合
　　たとえば、コインを投げたとき、裏と表の二つの選択肢となる。これは1bitで表現できます。

　　2. n=4の場合
　　4種類だから2bitで表現できます。これは、コインを2回投げた場合と同じでやはり2bitとなる。

　　3. 一般の場合
　　nbitで、2^nの場合を表現できます。従って、n個の選択肢の場合、$\log_2 n$

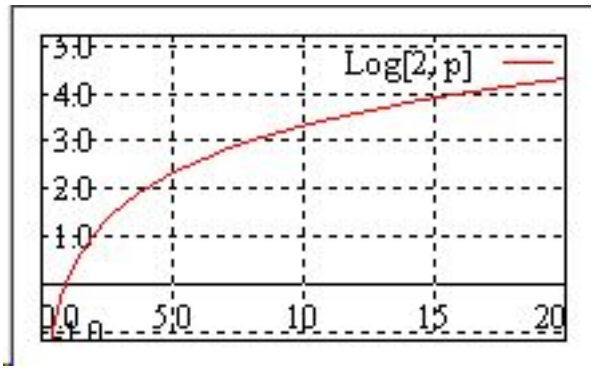
ビットとなる。

4. logて何？
   一般に、$2^n=m$　のとき、$\log_2$（m）=n　と表現します。logの次の2をlog（対数）の底と呼び、省略する場合があります。一般に、log(1/p)は-log p 、$\log_2$（m）は任意の底で log m/log 2 で計算できます。

   a. なぜ?（底の変換）
      $p=a^r$ -> $r=\log_a p$
      底をbとして両辺の対数をとる。-> $\log_b P=\log_b a^r$ -> $\log_b p=r\log_b a$ -> $r=\log_b p/\log_b a$

   b. なぜ?（積の対数は、対数の和）
      $a^r = p$, $a^s = q$ とおく。
      $a^{(r+s)} = a^r a^s=pq$ ->両辺のlogをとる　$\log_a(pq) = r + s = \log_a p + \log_a q$

   c. $\log_2$pのグラフ
      横軸p、縦軸が$\log_2$pのグラフです。縦軸をxとすると、横軸は2^xのグラフです。

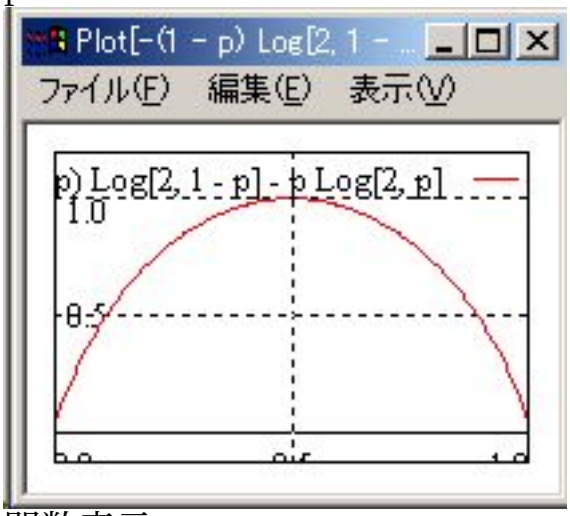


5. 英字、漢字
   記号を含む英字（記号を含めて約32文字）と漢字（約3000字）の1文字当たりのおおよその情報量を求めなさい。また、多くの場合英字は8bit、漢字は16bitで表現されます。この理由を考えなさい。必要なら、次の値を利用しなさい。
   $2^5=32$、$2^{10}=1024$、$2^{12}=4800$

B. 確率を考慮する
   1. シャノンの定義
      シャノンはいくつかの事象Eiがある確率Piで起こる場合、一つの事象Eiが持つ情報量を次のように定義しました。
      I(pi)=log(1/pi)
      そして、1記号当たりの平均情報量を次のように定義しました。
      I(p1,..pn)=p1*log(1/p1) + p2*log(1/p2) +.. + pn*log(1/pn)
      ここでlogの底は２とします。

2. 性質
   一般にn個の事象の生起確率が等しい場合に1記号の平均情報量は最大になり、log nとなります。

3. n=2の場合
   事象が２の場合、一方が確率pの場合、他方の確率は1-pとなります。平均情報量は
   　I(p,1-p)=p*log(1/p) +　(1-p)*log(1/(1-p))
   となります。これを、確率pのグラフにすると、次のようになります。p=0.0、1.0のときは、確実に予想出来ますから情報量0です。確率1/2の時は、全く答えの予想が出来ません。このとき情報量が最大で1となります。

   a. pを0..1に変化したときの平均情報量
      

   b. 関数表示ツール
      1. フリーウエア：FunctionView
         以下からダウンロードできます。
         http://hp.vector.co.jp/authors/VA017172/
         自己解凍後、陽関数に関数を設定し、設定メニューで変数の範囲を指定します。関数の様子を見るのにおすすめソフトです。

4. n=3の場合
   a. 問題
      3個の事象の平均情報量の最大値は幾らですか？
      1. 解答例
         各確率が1/3の場合が最大で、次の値になります。計算には、Mathmediaを利用しています。
         In[12]:=N(-3*(1/3)*log(2,(1/3)))
         Out[12]:=1.58

   b. 問題
      p1=0.3、p2=0.4、p3=0.3 の場合の平均情報量をいくらか。
      1. 解答例
         N(　)は、値の数値を求めます。

```
In[10]:=N(-2*0.3*log(2,0.3)-0.4*log(2,0.4))
Out[10]:=1.57
```

c. 問題
p1=0.1、p2=0.3、p3=0.6 の場合の平均情報量をいくらか。
1. 解答例

```
In[11]:=N(-0.1*log(2,0.1)-0.3*log(2,0.3)-0.6*log(2,0.6))
Out[11]:=1.29
```

d. 問題
先の例で3.bの場合より3.cの場合の方が情報量が低い理由を推測しやすさの観点から説明しなさい

C. 平均情報量を計算するプログラム（アプレット）
1. 機能
ここで、平均情報量を求めるプログラム（アプレット）を紹介しましょう。

2. アプレットの実行
下のウインドウでtextの欄に文字を入力します。たとえば、0011と入力します。「計算ボタン」を押すと、各文字の出現確率を求め、これから平均情報量を計算して表示します。この場合、0,1の出現確率はともに2/4で、平均情報量は1になります。0011110022では、平均情報量は1.52となります。

3. 計算プログラム（Java言語：アプレット）
アプレットのプログラムは、Cとよく似ています。ただし、文字列や表示関数はJava独特です。
node[]はtextの各文字の出現回数を記録する配列です（これを　new　で確保します）。最初のfor文で、node[]を0に初期化します。
　次に、st=text.getText()　で、textの記号をstring型の文字列stに読み込み、次のfor文で、各文字の出現回数をnode[]に記録します。st.charAt(c)はstのc番目の文字、st.length()はstの文字数を返す関数です。

　最後のfor文で、平均情報量を計算しています。Math.log(x) は、xの対数の値を返す数学関数です。prbは、各文字の確率を記録する文字列で、
　prb = prb + (char)i+":"+node[i]+"/"+st.length()+" ";
は、i番目の文字とその確率を示す文字を、prbに追記します。文字列の　+　は文字列を繋ぐ演算となります。infSumは平均情報量を累計します。Double.toString(infSum)　はinfSum　を小数の文字列に変換し、avrinfo.setText() で、計算した値を表示します。

```
 void button1_actionPerformed(ActionEvent e) {
  int CHAR_SIZE=256;
```

```
    int i,c;
    int node[]=new int[CHAR_SIZE]; //配列を定義する
    for (i = 0; i < CHAR_SIZE; i++)
      node[i] = 0;

    String st;
    st=text.getText();//入力した文字列を読む

    for (c = 0; c < st.length(); c++){//各文字の個数を求める
      node[(int)st.charAt(c)]++;
    }

    double infSum=0.0,pi;
    String prb="";
    for(i=0;i<CHAR_SIZE;i++){
      if(node[i] != 0) {
        pi=(double)node[i]/(double)st.length();//i番目の文字の出現確率を計
算
        prb=prb+(char)i+":"+node[i]+"/"+st.length()+" ";//i番目の文字の出現
確率の表示準備
        infSum=infSum+pi*(Math.log(1/pi)/Math.log(2.0));//平均情報量を
求める
      }
    }
    avrinfo.setText(Double.toString(infSum));//平均情報量を表示する
    label1.setText(prb);//各文字の出現確率の表示
  }
```

4. Javaソース
   アプレットはJava言語のプログラムで、ホームページから実行できるプログラムです。
   ソースプログラムを参照して下さい。
   くわしくアプレットを勉強したい人は、こちらのプログラミング＞Javaを参考にしてください。

D. 課題
1. 問題１
   0.3、0.2、0.15、0.15、0.1、0.1　の確率を持つ事象の平均情報量を求めなさい。

2. ヒント
   logは、windowsのアクセサリの電卓で計算できます。
   表示メニューで「関数電卓」に切り替えます。数字の次に　log　キーを押すと、10を底とするlogの値が表示されます。これを　log　2　で割ると 2 を底とするlogの値になります。

トップに戻る