タイトル：数学的準備体操：数理的基盤と計算の実際
講　　師：村田 昇
レポーター：赤崎孝文、伊藤淳司

数学的な準備体操として、数理的な神経回路モデルの概説と、さらに情報理論の基礎として、エントロピーや情報量などついて説明します。神経回路モデルについては銅谷さんが詳しいことを話されたので、ここでは簡単に述べるにとどめたいと思います。

神経回路モデルのなかでモジュールとしてよく考えられるのが、神経細胞と、それを組み合わせた神経回路網といわれる、いわゆるネットワークモデルです。神経細胞そのものに着目する場合には、どんなものを数理的には注目して扱わなければならないかというのを、以下に列記してあります。

空間的加算	神経細胞の内部状態は入力信号の重ね合わせで決定する
時間的加算	入力信号の影響はある時間持続し、後から入ってきた入力と相互に干渉する
閾値作用	出力は内部状態がある閾値を超えるまで発生しない
不応期	閾値は定数ではなく出力に応じて短期的に変化する
疲労	閾値は出力に応じて長期的に変化する
可塑性	学習や記憶は素子感の結合荷重の変化によってもたらされる

神経細胞の特徴の、もっとも大きな二つは、空間的加算と、時間的加算を行うということです。空間的加算というのは、ある神経細胞の内部状態は、外にある、他の多数の神経細胞からの入力信号の重ねあわせで変わる、ということです。時間的加算というのは、入力が瞬時に決まるわけではなくて、過去の履歴をある程度引きずって変わる、ということです。そういう性質があることにより、数理的には単純な細胞なのですが、いろいろなことができるのです。

もうひとつ大事な性質として、閾値作用というものがあります。特にパルスなどを扱う場合には、内部状態がある値を超えないと出力を出さない、という性質が非常に重要になってきます。

さらに、付随する性質でいろいろなものがあるのですが、ひとつは不応期というものです。これは、閾値が定数ではなく、長いあいだ発火していると閾値が上がって、なかなか発火しなくなるという性質です。不応期と疲労はほとんど同じもので、見方を変えていると言うといいのかもしれません。疲労というのも同じように、閾値が変化するという性質です。特に疲労という場合には長期的に変化するものを指すことが多いようです。

もうひとつは、これがいちばん大事なのでしょうが、可塑性という性質があります。可塑性というのは学習や記憶を実際に実現するために、素子間の結合の状態が変わることをいいます。どのようなしくみで可塑性が実現されているかというのは、まだあまり明確ではありません。

こうした特徴を踏まえて、どういうモデルを使って解析が行われているかというのは、次のように、大体３つに分けられます。

離散時間－離散情報　離散で内部状態を更新し、０、１を出力する

$$u = \sum_{i=1}^{n} w_i x_i \quad , z = 1(u - h) = \begin{cases} 1, & u > h \\ 0, & u \le h \end{cases} \qquad (1)$$

離散時間－連続情報　離散で内部状態を更新し、連続価を出力する

$$z = f(u - h), \quad f(x) = \tanh(x) \qquad (2)$$

連続時間－連続情報　連続に内部状態を更新し、連続価を出力する

$$\tau \frac{du(t)}{dt} = -u(t) + \sum_{i=1}^{n} w_i x_i(t) - h, \quad z(t) = f(u(t)) \qquad (3)$$

時間方向の情報をどう見るかというのと、出力や入力をどう見るかというので、大まかに考えると３つに分かれるのです。組み合わせから考えると４つになるのですが、連続時間で離散情報というのは、あまり意味がないので、ここには書いてありません。

いちばん単純なモデルとしては離散時間‐離散情報モデルというものがあります。これは、時間を、ユニットは何でもかまわないのですが、0,1,2,3,4 というように整数値でとり、それに応じて内部状態を更新しておきます。出力は 0 か１かの２値になってしまうので、離散的な情報しかありません。ここでは 0-1 と書いてありますが、0-1 でなくてたとえば３値とか４値でもかまわないわけですが、0-1 がいちばん楽なので、多くはこれを用います。

離散時間でもうすこし複雑なことをやらせようと思うと、出力を、連続値にしてみるということになるわけです。その場合には、上の離散時間‐離散情報で、内部状態に応じて出力を出す関数としてステップファンクションを使っていた代わりに、何か適当な非線型関数を使ってやります。

計算上の理由からよく使われるのはハイパボリックタンジェント *tanh* や、シグモイド関数といわれるものです。なぜこれがよく使われるかというと、いろいろな理由があると思うのですが、微分が出力値だけで計算できるという非常に単純な性質を持っている、というのが理由のひとつだと思います。ここでは話さないつもりですが、バックプロパゲーションという有名な学習則があります。これは微分を計算しなければいけないのですが、*tanh* やシグモイド関

数と呼ばれる関数を使っているぶんには、微分は直接計算しないで出力値だけを使って全部計算できてしまうのです。ここで *tanh* が嫌いな人は tan^{-1} で使うとか、いろいろなことをやってかまわないわけです。

これよりもうすこし複雑なものにしようと思うと、時間のほうを連続的に変化させるということがあるわけです。ということで最後が連続時間-連続情報で、銅谷さんの話の中で出てきた integrate-and-fire model の、内部状態の変化のところに出てきた式と同じものです。これは非常に単純な線形の作用をしている式しか書いてありませんが、もっと複雑にしたければ、この線形作用ではなくて非線型作用を入れたりするなどして、どれだけでも複雑にすることができます。

質問：
連続時間-離散情報が意味がないと言われた理由がよくわかりません。スパイク列というのはまさに連続時間-離散情報だと思うのですが。

普通はそれは連続時間-連続情報に入れてしまいます。大事じゃないと言うとやはり語弊があるのかもしれません。この分け方は、寄らば大樹の陰ということで、甘利先生の分け方に従ったというのがあるんですが、ただ扱い自体は、連続時間-離散情報は非常に難しくなるのです。数理解析の上から言うと，離散-離散というのはわかりやすくていいのですが、連続-離散にすると，常微分方程式の解の一意性がなくなってしまい，解けなくなってしまうのです。なので、普通は使いません。数理解析を目的とするモデルとしてはやりづらいというので書いてないということです。

ただ実際には，そういうモデルでスパイク列などの話をしなければいけないのですが、そのときには、たとえば Genesis とかいわれるシミュレーターツールを使ったりするわけです。具体的に解ける場合が非常に少ない、というのが一番の問題だと思います。

次に回路網モデルとして，今話したような単細胞モデルをたくさんつなげて，ネットワークにします。ネットワークにするときどういう構造を考えるかというと，大まかに分けると次のように二通りあります。

ひとつは，カスケード，フィードフォワードと言われるタイプの，順番に並べて下から上へ向かう結合しかないような，そういう一方方向の結合だけでできているモデルです。多層回路網とか，フィードフォワードネットワークとか呼ばれ，conventional には下から上に向けて描いてありますが，人によっては、情報の流れとして、上から下に描いたりします。

もうひとつは再帰的回路網というもので，リカレントネットなどと言われるタイプです。有名なものとしてはホップフィールドネットワークがこれにあたります。あるいはボルツマンマシンもこれにあたります。いわゆる上で説明したような層という概念があまりなく，全部が，一様に結合しています。もちろん全部が結合している必要はなく，適当なトポロジーを入れて，近いところだけ結合するというようなことは考えられるわけですが，いずれにせよ，層状結合というよりは，全体の中で適当に結合しているというかたちになります。

これが両極端で，これのあいの子みたいなものももちろんあるわけです。二つのリカレントネットワークを層状に結合するという話はもちろんあるわけで、必要に応じてどんどん複雑にすればよいのです。

次は，情報理論のいろいろな言葉の説明をします。

情報理論がどういう所で神経科学と関わるかというと，最近，シグナルのエントロピーを計算する人たちがたくさん出てきているようです。信号そのものがどれくらいの情報量を持っているかというのは非常に大事なことだと思うのですが、それを測るひとつの量としてエントロピーというものがあります。そのエントロピーというのは何なのか、というのがここからの話です。

エントロピーの定義は次のようになります。ある確率変数 X があって，その確率密度が $Px(x)$として与えられたときに，X のエントロピーは,

$$H(X) = -\int P_X(x)\log P_X(x)dx \qquad (1)$$

で定義されます。

これは直感的には確率変数の持つランダムさの目安，のようなものになっています。どういうことかというと，エントロピーは実は情報量とも言われるのですが，情報量とランダムさというのが非常に密接に関与してるわけです。例えば，極端な場合として、これは確率変数とは言わないのですが，1 という値しか出ないような確率変数を考えましょう。X を観測すると，常に 1 しか帰ってこない。ここにはランダムさは一切ないわけです。そのときエントロピーはどうなるか，を調べてみます。(1)式に当てはめるとこれが計算できるのですが、この場合は離散なので，積分を和で置き換えないといけません。X のとりうる値すべてについて summation をとってやる、

$$H(X) = -\sum P(x)\log P(x) \qquad (2)$$

が離散型の場合の定義です。この場合に計算してみるとすぐにわかりますが，X が 1 しかとらないので $H(x) = -P(1)\log P(1)$ になります。つまり、-1 log1です。log1 は 0 ですね。ということで，エントロピーは 0 になります。この場合は，実は何を観測するか前からわかってるわけですね。ということは、1 を受け取

ったとしてもまったく得られる情報がないわけです。そういう意味で，乱雑さがないものは情報がない、そういうことになります。

逆に，確率分布がある程度広がっていて、しかも、めったに出ないはずの値が出ると，非常に情報が豊富になるわけです。めったに起こらない事象が起こった、ということを測るために、*-logP* という、P の値が 0 に近くなると非常に大きな値になる量を用いているわけです。エントロピーとは、めったに起こらない事象が来るとそれだけ大きくなるような値を，すべての取りうる値で平均をとっている量なのです。だからこれは平均的にどれくらいの情報が来ているかということを測る量になるわけです。

今は 1 変数の場合を言いましたが多変数の場合にも自然に拡張できて、

$$H(X_1, X_2) = -\int P_{X_1 X_2}(x_1, x_2) \log P_{X_1 X_2}(x_1, x_2) dx_1 dx_2 \qquad (3)$$

このような定義になります。連続型の 2 変数の場合には，P_{X1X2} で平均をとってやります。離散の場合には 2 変数も 3 変数も関係なくて，X というベクトルでかまわないので、それの取り得る値全部について，足してやればいいのです。

エントロピーの加法性という性質が，非常に大事だと言われています。もうすこし先に行くとなぜ大事かという話をもうすこし別の言い方から説明します。

まず条件つき確率のエントロピーというものを考えます。条件つき確率というのは何かというと，二つの確率変数 X,Y があって、お互いに何らかの関係があったとします。このとき、X が x という値をとったときに，Y がどういう確率で分布するか，というのを確率密度 $P(y|x)$というふうに書きます。こうすると、これは$\int P(y|x)dy$がちゃんと 1 になるような密度になっているので、この確率密度のエントロピーが上と同じ定義で計算できます。

$$\mathrm{H}(\mathrm{Y}|\mathrm{x}) = -\int P(y|x) \log P(y|x) dy \qquad (4)$$

このようにすればいいわけです。Y という確率変数を，ある特定の x で条件づけた時のエントロピーがこういうかたちになります。ここで x は適当な分布にしたがってゆらいでいます。$H(Y|x)$の x の平均を取ったものを、条件つきエントロピーといいます。

条件つきエントロピーは，条件つき確率のエントロピーを平均したもの，条件づけている確率変数で平均したものです。これを定義すると，エントロピーの加法性という，次のような単純な関係が成り立ちます。X1 と X2 という二つの変数があったときの，同時分布に関するエントロピー，同時分布に関するエントロピーというのは，2 変数の場合のエントロピーのことですが，このエントロピーは，分解できるという性質を持っています。どのように分解できるかというと，X1 だけのエントロピーと，X1 で条件づけた X2 の条件つきエントロピーに分けられるのです。

$$H(X1,X2) = H(X1) + H(X2\,|X1) \tag{5}$$

このように分解できたときに $H(X2|X1)$の部分の意味が非常に大事になるわけですが，この部分は、X1 を知ったときの X2 の曖昧さを表しているわけです。

続けざまにもう一個べつの量の定義をします。mutual information あるいは相互情報量といわれる量が，今の条件つきエントロピーを使って定義されます。どのように定義するかというと，X1 のエントロピーから，X2 で条件づけた X1 の条件つきエントロピーを引いてやります。

$$\begin{aligned} I(X1;X2) &= H(X1)-H(X1|X2) \\ &= H(X2)-H(X2|X1) \end{aligned} \tag{6}$$

これは X1 と X2 の役割を入れ替えてもまったく同じ量が出ます。$H(X2|X1)$を書き換えしてみると、エントロピーの加法性から

$$H(X2\,|X1) = H(X1,X2) - H(X1) \tag{7}$$

こうなります。これを相互情報量の定義（(6)式）に代入すると，

$$I(X1;X2) = H(X1) + H(X2) - H(X1, X2) \qquad (8)$$

が出ます。この式を定義にもどって書きなおしてみると，

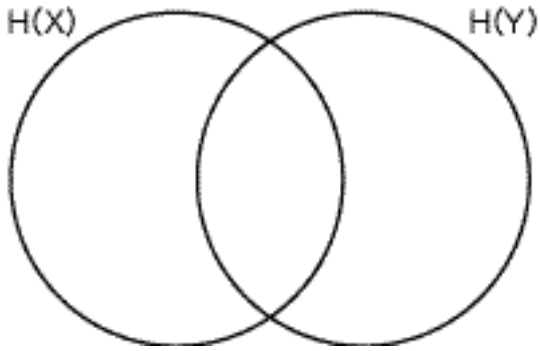
$$I(X1;X2) = E_{P_{X1X2}}\left[\log \frac{P_{X1X2}(X1,X2)}{P_{X1}(X1)P_{X2}(X2)}\right] \qquad (9)$$

このような量になります。この量は実は Kullback-Leibler divergence と呼ばれる、確率密度の間の一種の距離を測るための関数になっています。その中に同時分布と言われる、X1 と X2 の両方を含んでいる確率分布と，X1 だけの周辺分布と X2 だけの周辺分布の積をほおりこんだ形になっています。

このように見ると何がうれしいかというと，もし X1 と X2 が独立であれば，分子の $P_{X1X2}(X1,X2)$と分母の $P_{X1}(X1)P_{X2}(X2)$は，等しくなります。これは独立性の定義です。X1 と X2 が独立であれば，X1 と X2 の同時分布は個々の分布の積で書けるというのが独立性の正確な定義ですから，X1 と X2 が独立であれば分母と分子が同じになって，log の中は 1 になります。log1 はさっきも言ったように 0 で、0 をいくら平均取っても 0 です。ということで、独立な場合には必ず相互情報量は 0 になります。これが一番大事な性質です。

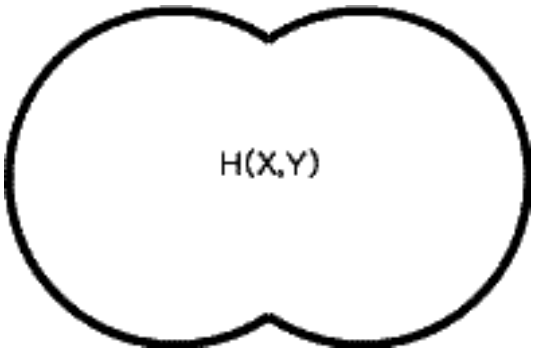
というわけで，この相互情報量というのは，独立性の基準になっているのです。二つの確率変数が独立か独立でないかを知る、かなりよい手がかりになります。ということで、情報理論のなかでは珍重されているわけです。

ここまでは、ずっと式だけで進めてきたわけですが、実は情報理論というのは非常にうまくできあがっていて，次のような集合的な見方をすることができます.すごく直感的な話ですが情報量の空間というものを考えます。X と Y という二つの確率変数を考えることにします。



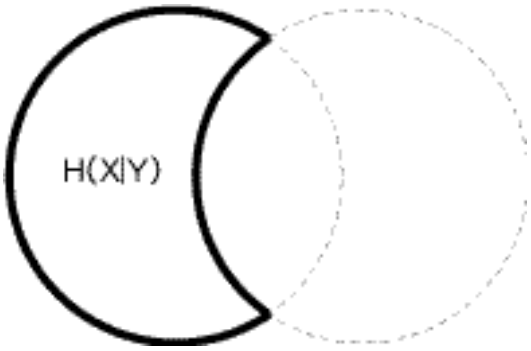
X が持っている情報の量を，丸で描きます。上の定義から言うと，X の持っている情報の量はエントロピーです。これは、この丸の大きさを表しているのがH(X)である、というふうに書けます。同じように，Y が持っている情報量，というのも描けるわけです。まあ丸である必要はないわけですが。このときに、二つの確率変数が持っている情報の量を，このなかにベン図状に描いてみるわけです。

H(X,Y)、つまり、2 変数の情報量はどこになるかというと，集合的に考えれば容易にわかるように，二つが持っている量をまとめあげたものなので,

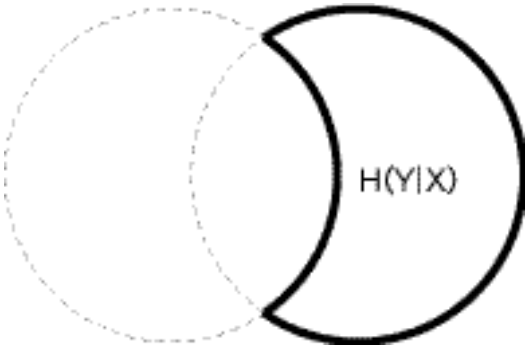


これが、X、Y の同時分布が持っている，情報の量になります。

条件つきエントロピーは何だったかというと，Y で条件付けた X のエントロピー、つまり、Y が何か適当に決まったあと、残る X のあいまいさ、というのが定義でした。Y がわかってしまってもわからない、X の未知の部分はどこになるかというと，

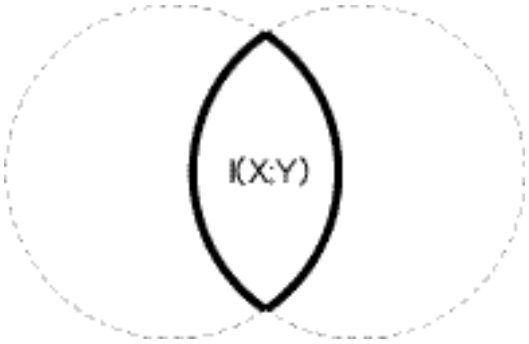


ここが、Y で条件付けた X の条件つきエントロピーになるわけです。集合的に書くと，X から Y を引いているのです。
同じように，X を知ってもなお残る Y の曖昧さということで，X に対する Y の条件つきエントロピーが，



になるわけです。

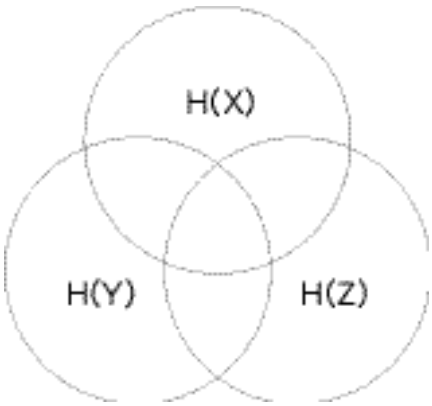
最後に残った真ん中の intersection の部分、ここが相互情報量になります。



先ほどまでの説明は、すべてエントロピーから出発していて、そこから作ってきた量はこのような集合的な関係を持っている，という話でしたが，実は逆にこのような集合的な関係から進めていくと，エントロピーという量は自然に定義されます。

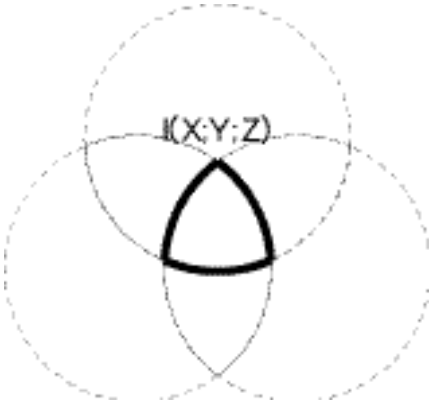
X、Y という量が入ってるからあたかも値ごとに何かあるのではないかというふうに見えてしまいますがそうではありません。実現値と確率変数を混同しないように注意してください。これはすべて確率変数，あるいは確率分布そのものの持っている性質として定義されるものであって，ある確率分布から出てきた実現値について定義される量ではありません。

いま 2 変数の話をしましたが 3 変数の話も同様にできます。いま X、Y、Z という 3 変数のエントロピー、もしくは情報量を，

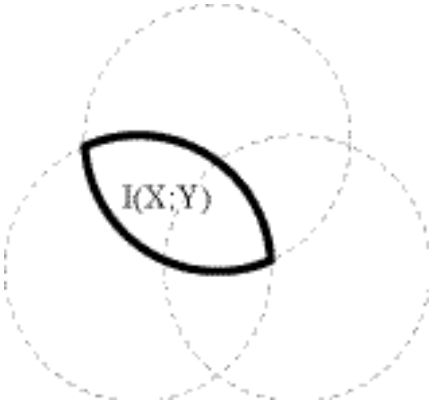


と表します。

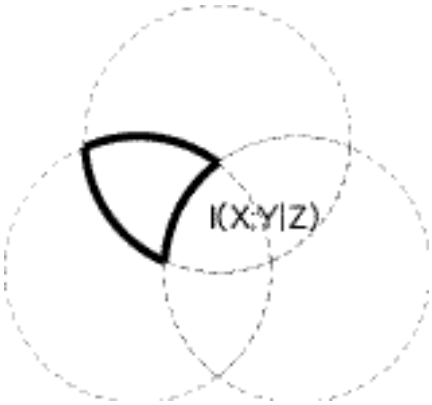
そのときに，三つが本当に交わるところが，三項間の相互情報量と呼ばれる量です。これが正しい，相互情報量の定義です。



次の部分は X と Y の相互情報量です。



それを Z を知ったときに残る部分だけにしたのが、Z で条件付けた時の X と Y の相互情報量であり、ちょうど



になるわけです。ここでもやはり集合的な解釈が成り立っています。
これを式で書くとこうなります。

$$H(X,Y) = H(X) \quad H(Y)$$
$$H(X|Y) = H(X \setminus Y) + H(X,Y) - H(Y) = H(X) \cap \overline{H(Y)}$$
$$I(X,Y) \quad = H(X) - H(X|Y) = H(X) \cap H(Y)$$

3変数の相互情報量を具体的に計算してみると次の式のようになります。

$$\begin{aligned}
&I(X;Y;Z) \\
&= I(X;Y) - I(X;Y|Z) \\
&= \{H(X) + H(Y) - H(X,Y)\} - \{H(X|Z) + H(Y|Z) - H(X,Y|Z)\} \\
&= \{H(X) + H(Y) - H(X,Y)\} - \{H(X,Z) - H(Z) + H(Y,Z) - H(Z) - H(X,Y,Z) + H(Z)\} \\
&= H(X) + H(Y) + H(Z) - H(X,Y) - H(Y,Z) - H(X,Z) + H(X,Y,Z)
\end{aligned}$$

集合論的な話から言うと、、簡単に分かると思います。

問題１

３層パーセプトロンの matlab での実装は、

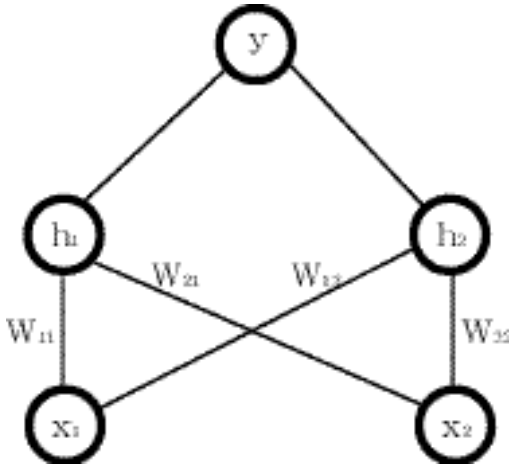
$$y = \phi(\mathbf{V} * \psi(\mathbf{W} * \mathbf{x} - \mathbf{b}) - \mathbf{c}) \qquad (4)$$

のように簡単にできる。例えば

$$\mathbf{W} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \quad \mathbf{b} = \begin{pmatrix} 0.5 \\ 1.5 \end{pmatrix} \quad \psi(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases}$$
$$\mathbf{V} = \begin{pmatrix} 1 & -1 \end{pmatrix} \quad \mathbf{c} = 0 \qquad \phi(x) = x$$

とすれば、exclusive OR が簡単に実現できる。

単純な三層パーセプトロンと呼ばれるものを組んでみましょう。



入力層を x、中間層を h、出力層を y とします。入力層と中間層のあいだの結合係数が W_{ij} です。W_{ij} というのは、入力層の x_j という細胞から，中間層の h_i という細胞への，結合荷重を表しています。

具体例として exclusive OR という論理関数で考えましょう。exclusive OR というのは 2 入力 1 出力型の論理関数です。入力が 0-0 の場合には出力として 0、入力が 0-1 の場合には出力として 1，1-0 の場合にはやっぱり 1，そして両方 1 だったら 0 になるという、そういう論理関数です。今の場合，これを，入力素子が 2 個，中間素子が 2 個，出力素子が 1 個の三層パーセプトロンで組んでみましょうという話です。

x_1 という入力層の1番目の細胞から中間層の1番目の細胞への結合係数を W_{11}、1 から 2 を W_{21}、2 から 1 を W_{12}、2 から 2 を W_{22} とする。それを、いわゆる普通の行列の形に書いてやります。これを W とします。いま W として、

$$W = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$

こういうものを用意しましょう。$b = \begin{pmatrix} 0.5 \\ -1.5 \end{pmatrix}$は，threshold です。中間層の 1 番目の素子に入力がきたら-0.5 してやるための、この 0.5 というのがこのベクトルの第 1 要素です。第 2 素子のほうは、-1.5 してあげます。

入力に対して W をかけて，threshold を引いたものは、中間素子が受け取る入力を，ベクトルに並べたものですね。（４）式のψの中までで，それを計算しています。そこにψという、適当な非線型関数、ここではステップファンクションをかけてやります。それをして出てきた結果が，中間素子の出力に出てくるわけです。その値が、ちょうど$\psi(W \cdot x - b)$のところまでで計算されています。

ここから上に行くのはいまと同じ手続きで、今度は V を、V というのは中間層と出力層の間の結合係数のことですね，かけてやって，さらに threshold、c を引いてやります。

というかたちで，これは３層なんでこれだけですけれど、４層５層にしたければさらに入れ子にしていけばいい、というかたちになっています。

問題 1.1

上の例が確かに exclusive OR に鳴子とを matlab で確かめよ。

（ヒント）step 関数は例えば **ceil(tanh())**などで実現できる。

解答例

```
W=[1 1;1 1];b=[.5;-1.5];V=[1,-1]
X=[0 0 1 1;0 1 01 ]
Y=V*ceil(tanh(W*x-b*ones(1,size(x,2))))
```

問題 2

対数正規分布は密度関数が

$$p(x)=\frac{1}{\sqrt{2\pi}\sigma x}e^{-\frac{(\log x-\mu)^2}{2\sigma^2}},(0<x<\infty)$$

であらわされる確率分布である。

この問題では，具体的にエントロピーを計算してもらいたい。いま定義として，エントロピーを与えたわけですけれど，実際になにか確率変数を扱う場合にわれわれが手に入れられるものは実は観測値しかないわけです。確率分布そのものを扱うことはめったになくて，何かある想定された確率分布から発生された系列だとか、観測値だとかがあって、それからエントロピーとか、相互情報量といわれるものを推定しなければいけないわけです。ここでは理論的に計算できる量と，実際に乱数を発生して，ヒストグラムのようなものを使って計算した相互情報量とかエントロピーというのがどのくらい一致するのか，ということを見ていこう、というわけです。

2.1　対数正規分布のエントロピーを定義に従って計算せよ。

解答例　$\frac{1}{2}\log\left(2\pi e\sigma^2\right)+\mu$

対数正規分布というのは何かというと，ある確率変数の対数が正規分布に従っているという，まあその名のとおりの，log-normal と呼ばれる分布です。僕もあまり良くは知らないのですが，これはけっこう最近の統計学では出てくるらしくて，たとえば、O157 の生存確率の，近似によく使われたらしいです。そういうのに使われたりするということで，けっこう、まっとうな関数らしいです。まずこれのエントロピーを計算しましょう。

正規分布の積分が 1 になるというのは知ってますね。平均μ、分散σ^2に従う正規分布の確率密度は,

$$p(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

と書かれます。

確率分布なので，全域(-∞ < x < ∞)で積分すると当然 1 になります。そのための，正規化因子が$\frac{1}{\sqrt{2\pi}\sigma}$です。

平均というのは，定義によると

$$E(x) = \int x\, p(x)\, dx$$

と書け、これが、μになります。
分散は X から平均値をひいてそれの二乗を平均する。

$$E((x - E(x))^2) = \int (x - E(x))^2\ p(x) dx$$

これが、どうなるかというと、σ^2になります。
もともと平均μで分散σ^2だといったんだからこれが成り立っている。これらの式は使ってかまいません。

2.2　μ=0、σ＝1 として密度関数のグラフをプロットせよ。

解答例

```
x=0:0.1:10;x(1)=eps;
plot(x,1./(2*pi*x).*exp(-.5*log(x).^2))
```

2.3　matlab の正規乱数　randn()　を使って、対数正規分布に従う乱数を適当な数だけ生成せよ。
（ヒント）ヒストグラムを表示するためには hist() という関数を使う。

解答例
tail が長いので必要に応じてトリミングする。

```
x=exp(randon(1,10000));
hist(x,50)
y=sort(x);
hist(y(1:9800),50)
```

*log(x)*が正規分布に従ってほしいわけですから，単純に言ってしまえば、指数関数の肩に正規乱数が乗ればいいわけですね。なので正規乱数を生成して，それを，指数関数の中に入れてやればいい、ということになります。

2.4　上で生成した乱数を用いて尤もらしいと思う方法でエントロピーを数値的に計算し、2.1 の l 結果と比較せよ。
（ヒント）　例えば多項分布だと思えば

```
[n xx]=hist(x)
```

によって代入される頻度を正規化して p_i とすれば、

$$H(X) = -\sum p_i \log p_i$$

がエントロピーを求める式になる。

解答例　bin の数は適当に選ぶこと。ただし次のやり方は一般的には良い推定量ではない。

```
[n xx]=hist(x,50);
m=sort(n(:)/sum(n(:)));
[v I]=max(ceil(m));
m-m(I:length(m));
H=-sum(m.*log(m));
```

何のために問題 2.3 でヒストグラムを作ったかというと，実はこれを使ってエントロピーの推定をしてみようというわけです。上で生成した乱数を用いて尤もらしいと思う方法でエントロピーを計算しろというのが、問題 2.4 です。実際の値は知っているわけです。（問題 2.1 解答参照。）μが 0 で、σが 1 の場合には，log の中身が 2πe という値になります。これを計算してみると、1.4189、理論的にはこういう値になるわけです。では実際にさっきの，乱数生成のルーチンを使って作ったもので，エントロピーを推定してみましょう。

エントロピーの推定はどうやればいいかというと，さっきヒストグラムを作ったのですね。エントロピーの定義は何だったかっていうと，-logP(x)の平均値を求めるわけですね。いまの場合数値的にしかわかっていないわけだから，データ点のいくつかしかわかってないわけなので，この P 自体はわからないわけです。もちろんどういう生成機構から出たか知っているのでわれわれは知っているのですが，データしか与えられなかったとしたらわからないわけです。わからないから P(x)をなんとかいいかげんにでっち上げてやらなければいけないわけです。
でっち上げるひとつの手段としてヒストグラムというものがあるのです。さっきすこし話したように，P(x)を離散分布だと思うと、エントロピーの定義は，P(x)logP(x)の積分がただの和になります。ヒストグラムの各ビンのなかに何個入るかというので，離散分布の各確率が求められます。これを使って，推定してあげよう、ということになるわけです。
ここでヒストグラムを用いて求めた結果は、理論的に求めた値とはかなり異なります。プログラムが間違ってるか，やり方が悪いかのどちらかですね。実はこれはやり方が悪いのです。多くの場合データ処理でヒストグラムが使われるのですが，エントロピーというのは非常に統計屋から見ると，たちの悪い量で，ヒストグラムで推定すると，あまりきれいな値が出ないことがわかっています。これ以外のやり方でエントロピーを推定する方法でいいものがあるかと言われると，はっきり言えばないのです。エントロピーは実は非常に推定しにくい量なのでどうしようもないので、これくらいで我慢してもらうしかない。
これはビンの数を変えて，ぎりぎりまでチューニングすればそこそこ良くはいくのですが，だいたいオーダーとして，全部で何個データ点があるかというのを n とすると、n の 1/3 乗くらいでしか，良くなりません。n の 1/3 乗でしか

良くならないということは 1000 倍データ点を取ってはじめて推定量の良さが 1/10 になるということです。

質問：
それはまったく元の分布によらない性質なのですか？

これは分布によりません。～$n^{(\frac{1}{3})}$の係数が分布によって違うはずです。僕もちょっと詳しくは知らないのですが，だいたい、1/3 だったと思います。漸近オーダーとして 1/3、1000 倍カウントしてはじめて 1/10 になる。

ということなので、気をつけて使ってくださいということです。エントロピーという量を，どこまで信用するかというのは，非常に難しいという話です。