知能機械と自然言語処理
知能機械部　第4回

ソフトウェア情報学部

Goutam Chakraborty

ヒストグラム（棒グラフ）

小学生500人の体重のデータがあります

http://www.chishiki.soft.iwate-pu.ac.jp/ISNLP/isnlp.html

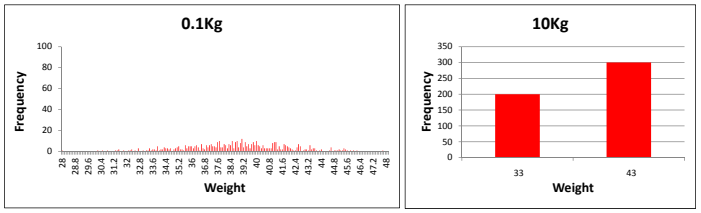
このようなデータから棒グラフを書くとき、最初にデータの最大と最小の値を求めて、全体の範囲（レンジ）を調べる事が重要です。

例えば（レンジ/データ数）の大きさに比例した棒の表示範囲を設定することで、適切なグラフを作る事が出来ます。

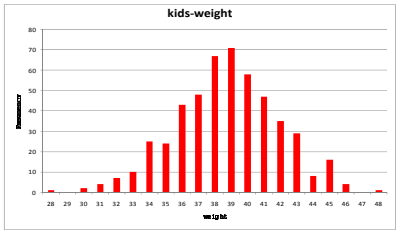
下記データのレンジは20Kgです。そのときの棒の表示範囲が0.1Kg，1Kg，10Kgであるグラフを次のスライドで比較します。

36.48	39.94	42.57	39.53	33.81
43.13	37.97	42.41	39.61	43.30
41.64	39.01	37.77	38.94	41.10
40.37	43.49	37.60	40.14	38.88
38.62	33.43	45.17	42.66	39.98
39.25	41.06	41.17	38.30	38.24
.		.		.
.		.		.
.		.		.

棒の広さによるヒストグラムの形



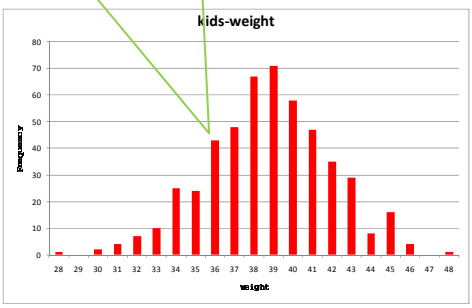
Too short



Too long

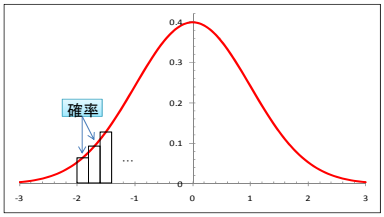
ヒストグラムから確率を求める

500人の子供の内、35.5～36.5Kgである子供は42人です。よってこの位置の確率は42/500=0.084です。



ヒストグラムから正規分布

離散的データの分布は棒グラフ（ヒストグラム）で表現します。
棒グラフの形をまねしてデータの分布を連続関数でも表現出来ます。
（たとえば正規分布、指数分布など）
ほとんどの自然のデータは正規分布になってます。
データの平均と分散から正規分布の関数を求められますし、
その正規分布の関数から確率も求められます。



標準正規分布

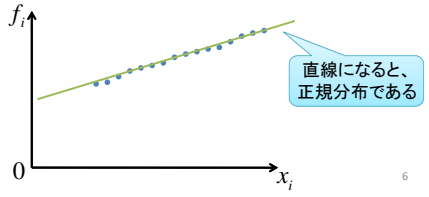
正規分布であることを確かめる方法

ソートを行った後のデータ

i	x_i
1	33.1
2	33.4
3	34.0
⋮	⋮
500	43.7

データ数が少ない場合は、下記の方法で確認できる

1. ソートを行う
2. $f_i = \dfrac{i-0.375}{n+0.25}$　（n：データ数　）
3. グラフにする



正規分布の確率密度関数

正規分布の形は $y(x)=\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$

確率密度関数 $f(x)$ の場合 $\int_{x_{\min}}^{x_{\max}} f(x)\,dx=1$

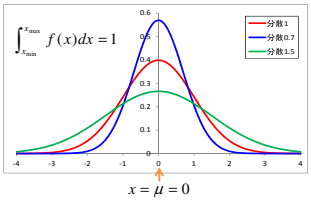
$$\int_{x_{\min}}^{x_{\max}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)dx=\sqrt{2\pi\sigma^2} \Rightarrow \int_{x_{\min}}^{x_{\max}} \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)dx=1$$

上の結果から正規分布の
確率密度関数は

Very Important equation
1-feature normal distribution

$$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

μ:平均値 ;σ:標準偏差



$x=\mu=0$

正規分布の場合確率密度関数から確率を求める手法

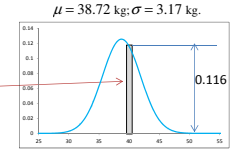
子供500人の内39.5～40.5Kgの人数は58人です。このデータから
体重39.5～40.5Kgの確率は $P(39.5\le x<40.5)=\frac{58}{500}=0.116$

同じ500件のデータの $\mu=38.72\ \mathrm{kg};\sigma=3.17\ \mathrm{kg}$

正規分布の確率密度関数 $f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ からも下記通り確率を求められる。

先ず $f(x=40)=\frac{1}{\sqrt{2\pi(3.17)^2}}\exp\left(-\frac{(x-38.72)^2}{2(3.17)^2}\right)=0.116$ を計算する。

確率を求めたい特色（体重）の幅は
（40.5－39.5）＝1Kg.です。
よって確率は 0.116X1=0.116 で
す。



パラメター μ と σ の値が分かれば確率を求められます。

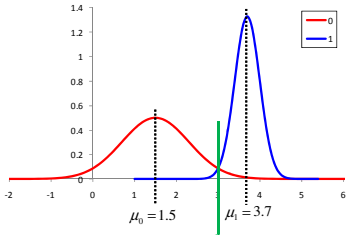
確率密度関数を用いて分類する

（特色一つ;クラス二つの分類問題）

教師データ

特徴量	クラス情報	
1.44	0	$\mu_0=1.5,\ \sigma_0=0.8$
1.34	0	
⋮	⋮	
2.52	1	$\mu_1=3.7,\ \sigma_1=0.3$
2.61	1	
⋮	⋮	

0か1を認識するための
教師データ



この特色の値のとき0と1の確率が同じ。
よってこの特色の値は境界です。

ベイズの定理(Bayes' Rule)を用いて分類する

0クラスの特色の平均1.5、標準偏差0.8； 1クラスの特色の平均3.7、標準偏差0.3

教師データ

特徴量	クラス情報	
1.44	0	60個
1.34	0	
⋮	⋮	
2.52	1	40個
2.61	1	
⋮	⋮	

事前確率（A prior probability） $P(C_0)=\frac{60}{100}$ ； $P(C_1)=\frac{40}{100}$

特色の値 t = 3.0、t の幅は Δt の場合クラスの条件付き確率は

$$P(3.0\le t<(3.0+\Delta t)\,|\,C_0)=\frac{1}{\sqrt{2\pi}\times0.8}\exp\left(-\frac{(3.0-1.5)^2}{2\times0.8^2}\right)\times\Delta t=0.35\times\Delta t$$

$$P(3.0\le t<(3.0+\Delta t)\,|\,C_1)=\frac{1}{\sqrt{2\pi}\times0.3}\exp\left(-\frac{(3.0-3.7)^2}{2\times0.3^2}\right)\times\Delta t=0.07\times\Delta t$$

特色の値 t = 3.0 のときそれぞれクラスの事後確率 (Posterior Probability) は

$$P(C_0\,|\,t=3.0)=P(3.0\le t<(3.0+\Delta t)\,|\,C_0)\times P(C_0)=0.35\times\Delta t\times0.6=0.21\times\Delta t$$

$$P(C_1\,|\,t=3.0)=P(3.0\le t<(3.0+\Delta t)\,|\,C_1)\times P(C_1)=0.07\times\Delta t\times0.4\approx0.03\times\Delta t$$

よってt=3.0のときクラス0を判断する。

ベイズの定理(Bayes' Rule)を用いて境界を求める (1)

0クラスの特色の平均1.5、標準偏差0.8； 1クラスの特色の平均3.7、標準偏差0.3

事前確率（A prior probability） $P(C_0)=\frac{60}{100}$ ； $P(C_1)=\frac{40}{100}$

教師データ

特徴量	クラス情報	
1.44	0	60個
1.34	0	
⋮	⋮	
2.52	1	40個
2.61	1	
⋮	⋮	

特色の値 t = b の時0クラスと1クラスの確率が同じであれば t=bは境界です

$$p(b\le t<(b+\Delta t)\,|\,C_0)=\frac{1}{\sqrt{2\pi}\times0.8}\exp\left(-\frac{(b-1.5)^2}{2\times0.8^2}\right)\times\Delta t$$

$$p(b\le t<(b+\Delta t)\,|\,C_0)=\frac{1}{\sqrt{2\pi}\times0.8}\exp\left(-\frac{(b-1.5)^2}{2\times0.8^2}\right)\times\Delta t \Big\} \text{[1]}$$

ベイズの定理(Bayes' Rule)を用いて境界を求める (2)

特色の値 t = b の時それぞれクラスの事後確率 (Posterior Probability) は同じですから

$$P(C_0\,|\,t=b)=P(b\le t<(b+\Delta t)\,|\,C_0)\times P(C_0)=P(C_1\,|\,t=b)=P(b\le t<(b+\Delta t)\,|\,C_1)\times P(C_1)$$

前のスライドの式[1]を用いて、クラス条件付き確率の値を代入すると

$$\frac{1}{\sqrt{2\pi}\times0.8}\exp\left(-\frac{(b-1.5)^2}{2\times0.8^2}\right)\times\Delta t\times0.6=\frac{1}{\sqrt{2\pi}\times0.3}\exp\left(-\frac{(b-3.7)^2}{2\times0.3^2}\right)\times\Delta t\times0.4$$

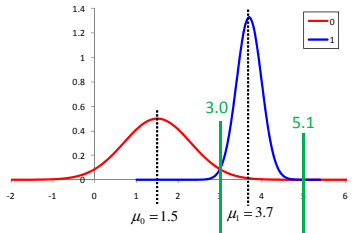
$$\frac{0.6}{2.00}\exp\left(-\frac{(b-1.5)^2}{2\times0.8^2}\right)=\frac{0.4}{0.75}\exp\left(-\frac{(b-3.7)^2}{2\times0.3^2}\right)\Rightarrow \exp\left(-\frac{(b-1.5)^2}{2\times0.8^2}+\frac{(b-3.7)^2}{2\times0.3^2}\right)=1.77 \Rightarrow$$

$$\left(-\frac{(b-1.5)^2}{2\times0.8^2}+\frac{(b-3.7)^2}{2\times0.3^2}\right)=\log_e 1.77=\ln 1.77=0.57 \Rightarrow 55b^2-447b+849=0 \Rightarrow \text{解 } b \text{ は3.0と5.1です。}$$

よって $b=3.0$ と 5.1 が境界の値である。

ベイズの定理(Bayes' Rule)を用いて境界を求める (3)

よって$b = 3.0$と5.1が境界の値である。図を見て分かるように二つの境界点がある。



この特色の値のとき0と1の確率が同じ。
よってこの特色の値は境界です。

この特色の値のとき0と1の確率が同じ。
よってこの特色の値は境界です。

次回の内容

特色増えると下記の図のように重なる
部分がなくなって分類制度が増加する。
データの特色が2以上の場合
確率密度関数はどうなりますか？