知能機械と自然言語処理
知能機械部　第5回

ソフトウェア情報学部

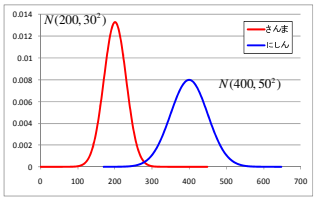
Goutam Chakraborty

正規分布の復習

さんまとにしんの二クラスのデータがあります。特色は魚の重さ(w)です。
さんまはクラス0、にしんはクラス1とします。さんまの重さの平均は200g
で標準偏差は30です。にしんの平均重さは400gで標準偏差は50gです。

- 0：さんま3000匹
- 1：にしん2000匹

$$P(C_0)=\frac{3000}{5000},\ P(C_1)=\frac{2000}{5000}$$

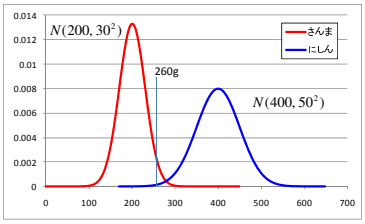


もし特色の重さが260gであれば、その条件付き確率は

$P(260 \le w < 260 + \Delta w \mid C_i)$ とする

正規分布復習

$$P(C_0)=\frac{3000}{5000},\ P(C_1)=\frac{2000}{5000}$$



if $w=260$ あたりの確率を
$P(260 \le w < 260+\Delta w \mid C_i)$ とする

正規分布の確率密度関数から確率を求める簡単な手法

$N(\mu=200, \sigma^2=30^2)$ のとき、

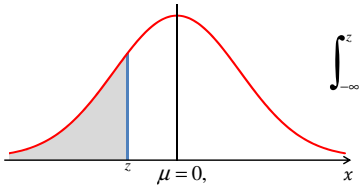
$$f(x)=\frac{1}{30\sqrt{2\pi}}\exp\left(-\frac{1}{2}\times\left(\frac{x-200}{30}\right)^2\right)$$

$$P(260 \le x < (260+\Delta w) \mid C_0)=\int_{260}^{260+\Delta w} f(x)\,dx$$

$$=\int_{260}^{260+\Delta w}\frac{1}{30\sqrt{2\pi}}\exp\left(-\frac{1}{2}\times\left(\frac{x-200}{30}\right)^2\right)dx \quad \text{.....[1]}$$

標準正規分布の確率密度関数表を参考にして確率を求める方法

前のスライドの式[1]の積分を求めるのは難しい。
しかし標準正規分布の確率密度関数表を参考にして確率を簡単に求められる。
標準正規分布の $\mu=0,\ \sigma=1$



$$\int_{-\infty}^{z}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}x^2\right)dx \quad \text{...[2]}$$

上記の図は標準分布の確率密度関数です。網掛けの面積はxの値$-\infty$から
zまでの確率です。その値は[2]の式で求められます。標準確率密度関数表
には全てのzの値に対応した網掛けの面積(確率)の値が載っています。

一般の正規分布から標準正規分布に変更方法

データ x が、平均値 μ、標準偏差 σ の一般正規分布データであるとき、
$z=(x-\mu)/\sigma$ と加工すると、データは標準正規分布のデータになる。
zの平均は0、標準偏差は1になります。

http://www.six-sigma-material.com/Normal-Distribution.html

標準正規分布表使用法

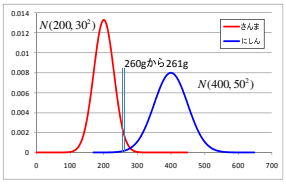
C_0(さんま)の場合赤い分布

平均200、標準偏差30の場合、正規分布の書き方は $N(200, 30^2)$

xからzを求める式は $Z=(x-200)/30$

$x=260$ の時$z=(260-200)/30$です。

260g辺り、260gから261g($\Delta w = 1g$)の間の重さのさんまの確率は下記通り求められます:



$$P\left(Z \le \frac{261-200}{30}\right) = P\left(Z \le \frac{61}{30}\right)$$
$$= P(Z \le 2.03) = 0.9788$$

さんまの重さ271gの以下の確率は0.9788です。

$$P\left(Z \le \frac{260-200}{30}\right) = P\left(Z \le \frac{60}{30}\right)$$
$$= P(Z \le 2.00) = 0.9772$$

さんまの重さ260gの以下の確率は0.9772 です。

よって、

$\therefore P(C_0 \mid 260 < w \le 261)$

$= 0.0016 \times \frac{3}{5} = 0.001$

さんまの事前確率$P(C_0) = 3/5$です。よって、事後確率

$\therefore P(C_0 \mid 260 < w \le 261)$

$= 0.0016 \times \frac{3}{5} = 0.001$

同様に、C_1についても求めると、

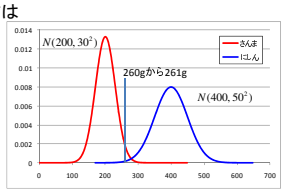
標準正規分布表使用法

C_1(にしん)の場合青い分布

平均400、標準偏差50の場合、正規分布の書き方は

$x = 400$ の時$z=(260-400)/50$です。

260g辺り、260gから261gの間の重さのにしんの確率は下記通り求められます:



$$P\left(Z \le \frac{261-400}{50}\right) = P\left(Z \le \frac{-139}{50}\right)$$
$$= P(Z \le -2.78) = 0.0027$$

にしんの重さ261gの以下の確率は0.0027 です。

$$P\left(Z \le \frac{260-400}{50}\right) = P\left(Z \le \frac{-140}{50}\right)$$
$$= P(Z \le -2.8) = 0.0026$$

にしんの重さ260gの以下の確率は0.0026 です。

よって、

$P(260 < x \le 261 \mid C_1)$

$= (0.0027 - 0.0026) = 0.0001$

にしんの事前確率$P(C_0) = 2/5$です。よって、事後確率

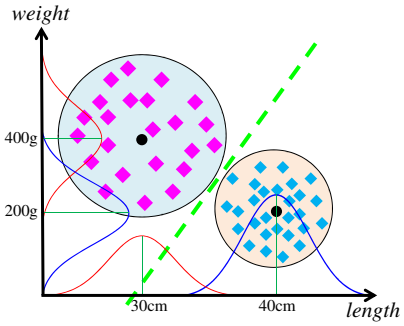
$\therefore P(C_1 \mid 260 < w \le 261)$

$= 0.0001 \times \frac{2}{5} = 0.00025$

よって、クラスC_0さんまの確率は高い。

∴この場合、 C_0 に分類される

2次元の正規分布



各特色を2次元の空間に表示させると、
重ならずに分けることができる

d次元の正規分布（特色の数dの場合）

一次元(特色は一つ)の正規分布

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

d次元の正規分布

$$f(\vec{x}) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu})^T \Sigma^{-1} (\vec{x}-\vec{\mu})\right)$$

d：次元数(dimension)　＝特色の数

Σ：相関行列(d × d)

d次元の場合ベイズの定理(公式)

$$P(C_0 \mid 200 \le weight < 220 \wedge 35 \le length < 37)$$
$$= P(C_0) \times P(200 \le weight < 220 \wedge 35 \le length < 37 \mid C_0)$$

- d次元の特徴量を表すときはx_d^n で表す

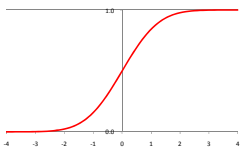
n:データの番号，d:次元の番号

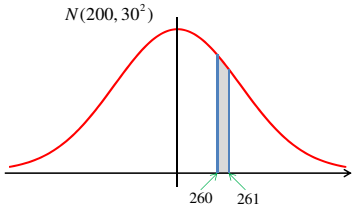
$$\begin{matrix} x_1^1 = 46.1 & x_2^1 = 180 & \cdots & x_d^1 \\ x_1^2 = 31.6 & x_2^2 = 241 & \cdots & x_d^2 \\ \vdots & & \ddots & \vdots \\ x_1^n & x_2^n & \cdots & x_d^n \end{matrix}$$

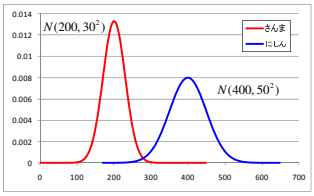
STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.9	.00005	.00005	.00004	.00004	.00004	.00004	.00004	.00004	.00003	.00003
-3.8	.00007	.00007	.00007	.00006	.00006	.00006	.00006	.00005	.00005	.00005
-3.7	.00011	.00010	.00010	.00010	.00009	.00009	.00008	.00008	.00008	.00008
-3.6	.00016	.00015	.00015	.00014	.00014	.00013	.00013	.00012	.00012	.00011
-3.5	.00023	.00022	.00022	.00021	.00020	.00019	.00019	.00018	.00017	.00017
-3.4	.00034	.00032	.00031	.00030	.00029	.00028	.00027	.00026	.00025	.00024
-3.3	.00048	.00047	.00045	.00043	.00042	.00040	.00039	.00038	.00036	.00035
-3.2	.00069	.00066	.00064	.00062	.00060	.00058	.00056	.00054	.00052	.00050
-3.1	.00097	.00094	.00090	.00087	.00084	.00082	.00079	.00076	.00074	.00071
-3.0	.00135	.00131	.00126	.00122	.00118	.00114	.00111	.00107	.00104	.00100
-2.9	.00187	.00181	.00175	.00169	.00164	.00159	.00154	.00149	.00144	.00139
-2.8	.00256	.00248	.00240	.00233	.00226	.00219	.00212	.00205	.00199	.00193
-2.7	.00347	.00336	.00326	.00317	.00307	.00298	.00289	.00280	.00272	.00264
-2.6	.00466	.00453	.00440	.00427	.00415	.00402	.00391	.00379	.00368	.00357
-2.5	.00621	.00604	.00587	.00570	.00554	.00539	.00523	.00508	.00494	.00480
-2.4	.00820	.00798	.00776	.00755	.00734	.00714	.00695	.00676	.00657	.00639
-2.3	.01072	.01044	.01017	.00990	.00964	.00939	.00914	.00889	.00866	.00842
-2.2	.01390	.01355	.01321	.01287	.01255	.01222	.01191	.01160	.01130	.01101
-2.1	.01786	.01743	.01700	.01659	.01618	.01578	.01539	.01500	.01463	.01426
-2.0	.02275	.02222	.02169	.02118	.02068	.02018	.01970	.01923	.01876	.01831
-1.9	.02872	.02807	.02743	.02680	.02619	.02559	.02500	.02442	.02385	.02330
-1.8	.03593	.03515	.03438	.03362	.03288	.03216	.03144	.03074	.03005	.02938
-1.7	.04457	.04363	.04272	.04182	.04093	.04006	.03920	.03836	.03754	.03673
-1.6	.05480	.05370	.05262	.05155	.05050	.04947	.04846	.04746	.04648	.04551
-1.5	.06681	.06552	.06426	.06301	.06178	.06057	.05938	.05821	.05705	.05592
-1.4	.08076	.07927	.07780	.07636	.07493	.07353	.07215	.07078	.06944	.06811
-1.3	.09680	.09510	.09342	.09176	.09012	.08851	.08691	.08534	.08379	.08226
-1.2	.11507	.11314	.11123	.10935	.10749	.10565	.10383	.10204	.10027	.09853
-1.1	.13567	.13330	.13136	.12924	.12714	.12507	.12302	.12100	.11900	.11702
-1.0	.15866	.15625	.15386	.15151	.14917	.14686	.14457	.14231	.14007	.13786
-0.9	.18406	.18141	.17879	.17619	.17361	.17106	.16853	.16602	.16354	.16109
-0.8	.21186	.20897	.20611	.20327	.20045	.19766	.19489	.19215	.18943	.18673
-0.7	.24196	.23885	.23576	.23270	.22965	.22663	.22363	.22065	.21770	.21476
-0.6	.27425	.27093	.26763	.26435	.26109	.25785	.25463	.25143	.24825	.24510
-0.5	.30854	.30503	.30153	.29806	.29460	.29116	.28774	.28434	.28096	.27760
-0.4	.34458	.34090	.33724	.33360	.32997	.32636	.32276	.31918	.31561	.31207
-0.3	.38209	.37828	.37448	.37070	.36693	.36317	.35942	.35569	.35197	.34827
-0.2	.42074	.41683	.41294	.40905	.40517	.40129	.39743	.39358	.38974	.38591
-0.1	.46017	.45620	.45224	.44828	.44433	.44038	.43644	.43251	.42858	.42465
-0.0	.50000	.49601	.49202	.48803	.48405	.48006	.47608	.47210	.46812	.46414

ちなみに、ExcelではNORMSDISTという関数で
上の式の計算ができる







正規分布復習

$P(C_0) = \frac{3000}{5000}$, $P(C_1) = \frac{2000}{5000}$