PRMLゼミ

1.2　確率論

anmitsu48

1.2節のイントロ

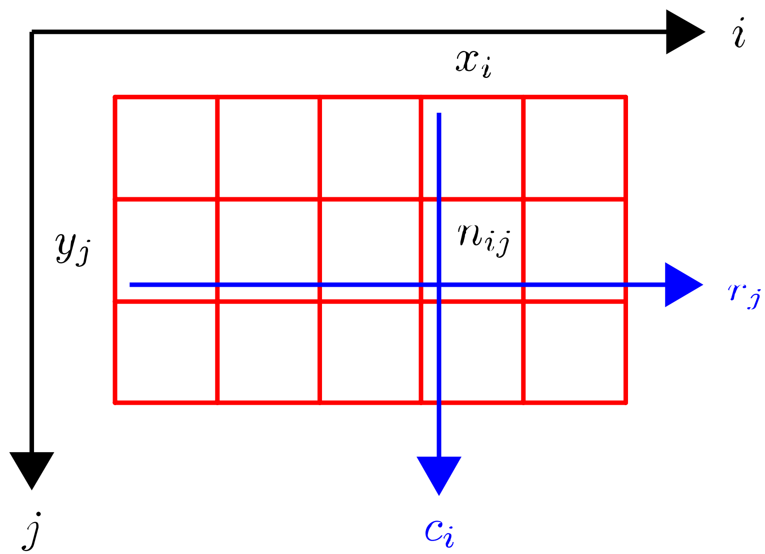
同時確率と周辺確率

- 1回の試行で確率変数 X, Y の両方について
  サンプルをとる。この試行を N 回行ったとする。
- X が x_i , Y が y_j をとる確率： $p(X=x_i, Y=y_j)$

  同時確率

- X が x_i , Y が y_j となる試行が n_{ij} 個あったとすると、

$$p(X=x_i, Y=y_j)=\frac{n_{ij}}{N}$$



- 周辺確率（Yに関する周辺化）

$$p(X=x_i)=\sum_j p(X=x_i, Y=y_j)$$

同時確率と条件付き確率

- 変数 X が x_i の下で、変数 Y が y_j となる確率。（条件付き確率）

$$p(Y = y_j | X = x_i) = \frac{p(X = x_i, Y = y_j)}{\underbrace{\sum_j p(X = x_i, Y = y_j)}}$$

分母は X が x_i となる確率

- 条件付き確率の公式

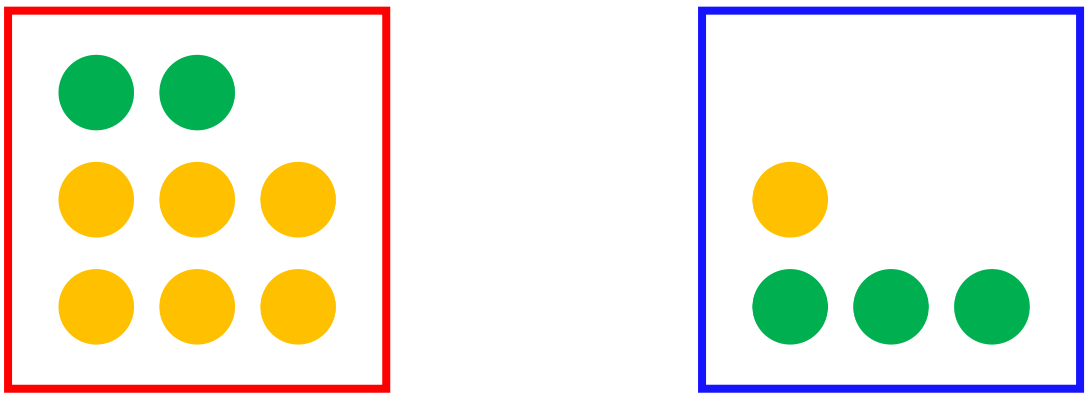
$$p(X, Y) = p(Y|X)p(X) = p(X|Y)p(Y)$$

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

ベイズの定理

1.2節のイントロで考える問題

- ２つの箱（赤・青）に、青りんごとオレンジが入っている。
- 箱を1つランダムに選び、その後、果物を1つランダムに選ぶ。どの果物かを記録したら元の箱に戻す。



- 赤色の箱を選ぶ確率:2/5, 青色の箱を選ぶ確率:3/5
- 箱の中の果物は全てを等確率で選択する。

箱とフルーツの問題の定式化 ①

- 選んだ箱が「赤」、「青」である確率

$$p(B = r) = \frac{4}{10} = \frac{2}{5}, \quad p(B = b) = \frac{6}{10} = \frac{3}{5}$$

- 箱を選んだ後に、選んだフルーツが「青りんご」、「オレンジ」である確率

$$p(F = a | B = r) = \frac{1}{4}, \quad p(F = o | B = r) = \frac{3}{4}$$

$$p(F = a | B = b) = \frac{3}{4}, \quad p(F = o | B = b) = \frac{1}{4}$$

- 選んだフルーツが「青りんご」である確率
  - 選んだ箱が赤の場合と青の場合を足し合わせる。

$$p(F = a) = p(F = a | B = r) p(B = r) + p(F = a | B = b) p(B = b)$$

$$= \frac{1}{4} \times \frac{2}{5} + \frac{3}{4} \times \frac{3}{5} = \frac{11}{20}$$

箱とフルーツの問題の定式化 ②

- 「オレンジ」を選んだとき、それが「赤」色の箱から選ばれた確率
  - 「オレンジ」を選んだということを全体としたとき、
    それが「赤」、「青」のどちらであるかを考えることができる。

$$p(B=r|F=o) = \frac{p(F=o|B=r)p(B=r)}{p(F=o)} = \frac{3/4 \times 2/5}{9/20} = \frac{2}{3}$$

ベイズの定理

「ベイズの定理」を意識しなくても、
全体（オレンジ）と部分（オレンジかつ赤）を
意識すると、6/20を9/20で割ることは明らか。

- 選ばれたフルーツがわからないとき、どの箱が選ばれたかの確かさを表す確率を「（箱に関する）事前確率」という。
- 一方で、選ばれたフルーツがわかった後に、どの箱が選ばれたかの確かさを表す確率を「（箱に関する）事後確率」という。

事前確率 $p(B=r) = \frac{2}{5}$ $p(B=r|F=o) = \frac{2}{3}$ 事後確率

確率変数と独立

- 確率変数 X, Y が独立
  - 同時確率が個々の変数の確率の積に分解できる。

$$p(X, Y) = p(X)p(Y)$$

- 条件付き確率と独立
  - 条件付き確率が条件付けている変数に依存しない。

$$p(Y|X) = p(Y)$$

※ 先ほどの箱とフルーツの例では $p(F = a) \neq p(F = a|B = r)$ となるため、
確率変数 F, B は独立ではない。

1.2.1節の紹介

- 1.2.1節　確率密度

確率密度関数

- 確率変数 x が連続で、その値が区間 (a,b) にある確率

$$p\,(x \in (a,b)) = \int_a^b \underline{p(x)}\,\mathrm{d}x$$

確率密度関数

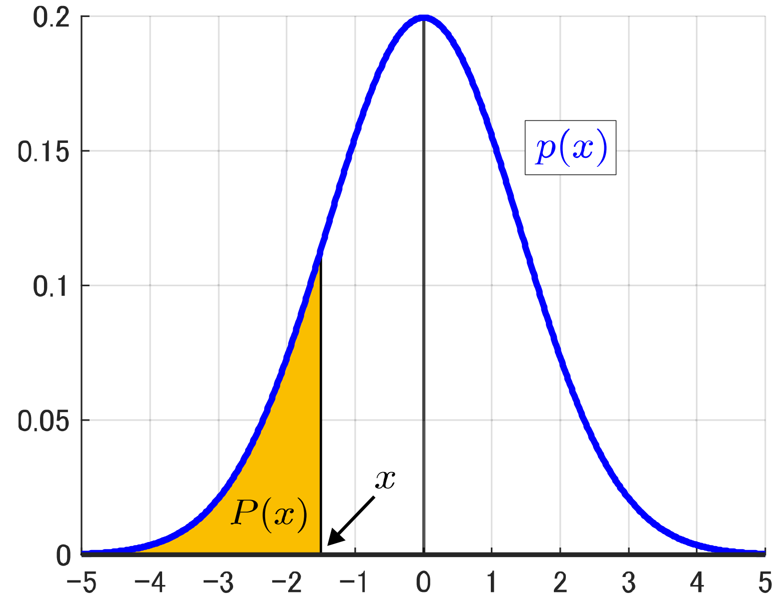
- 確率密度関数が満たす性質
  - ① 確率密度関数は非負
  - ② 全区間で積分すると1

$$p(x) \geq 0,\ \forall x \qquad \int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = 1$$

- 累積密度関数 $P(z)$:
  確率変数 x が z 以下となる確率

$$P(z) = \int_{-\infty}^{z} p(x)\,\mathrm{d}x$$

$$\frac{\mathrm{d}}{\mathrm{d}x}P(x) = p(x)$$



確率密度関数の変数変換

※　x, y の確率密度関数を $p_x(x), p_y(y)$ とここでは表すことにする。

- x, y の間に $x = g(y)$ という関係があり、x の確率密度関数 $p_x(x)$ がわかっているとき、$p_y(y)$ がどのように表されるかを考える。
- x の世界の区間 $(x, x + \delta x)$ が、y の世界の区間 $(y, y + \delta y)$ に対応するとき、$p(x)\delta x \approx p(y)\delta y$ が成立する。

$$\Rightarrow \quad p_y(y) = p_x(x)\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right| = p_x(g(y))|g'(y)|$$

【cf. 2次元の場合】

- $(x, y) = \left(h_1(z, w), h_2(z, w)\right)$ のときは、ヤコビアンを使って変数変換をする。

$$J = \begin{pmatrix} \dfrac{\partial h_1(z,w)}{\partial z} & \dfrac{\partial h_1(z,w)}{\partial w} \\ \dfrac{\partial h_2(z,w)}{\partial z} & \dfrac{\partial h_2(z,w)}{\partial w} \end{pmatrix} \quad \Rightarrow \quad \begin{aligned} p_{z,w}(z, w) &= p_{x,y}(x, y) \\ &\quad \times |\det J(z, w)| \end{aligned}$$

確率密度関数に関するその他のコメント

- 多変数の場合も1変数と同様の式が成立する。

$$p(\boldsymbol{x}) \geq 0, \forall \boldsymbol{x} \qquad \int \underline{p(\boldsymbol{x})} \, \mathrm{d}\boldsymbol{x} = 1$$

同時確率密度関数

- 周辺確率と条件付き確率の公式の連続版

$$p(x) = \int p(x, y) \, \mathrm{d}y \qquad p(x, y) = p(y|x)p(x) = p(x|y)p(y)$$

1.2.2節の紹介

- 1.2.2節　期待値と分散

期待値

- 離散分布（確率分布: $p(x)$）の場合

$$\mathbb{E}[f(x)] = \sum_x p(x)f(x)$$

- 連続分布（確率分布: $p(x)$）の場合

$$\mathbb{E}[f(x)] = \int p(x)f(x)\,\mathrm{d}x$$

- コンピュータなどで期待値の計算を行う場合
  - 真の確率分布 $p(x)$ から N 個のデータ点が得られているだけで、確率分布自体はわからない。
  - 期待値の計算は、得られたデータの平均値で近似する。

$$\mathbb{E}[f(x)] \approx \frac{1}{N}\sum_{n=1}^{N} f(x_n)$$

分散と共分散

- $f(x)$ の分散(variance)の定義

$$\mathrm{var}[f(x)] = \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])^2\right] = \mathbb{E}\left[f(x)^2\right] - \mathbb{E}[f(x)]^2$$

$$\Downarrow$$

$$\mathrm{var}[x] = \mathbb{E}\left[x^2\right] - \mathbb{E}[x]^2$$

- 2つの確率変数 x, y の共分散(covariance)の定義

$$\mathrm{cov}[x, y] = \mathbb{E}_{x,y}\left[(x - \mathbb{E}[x])\,(y - \mathbb{E}[y])\right] = \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y]$$

  - 2つの確率変数がベクトルの場合

$$\begin{aligned}\mathrm{cov}[\boldsymbol{x}, \boldsymbol{y}] &= \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\left[(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])\,(\boldsymbol{y} - \mathbb{E}[\boldsymbol{y}])^\top\right] \\ &= \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\left[\boldsymbol{x}\boldsymbol{y}^\top\right] - \mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{y}^\top]\end{aligned}$$

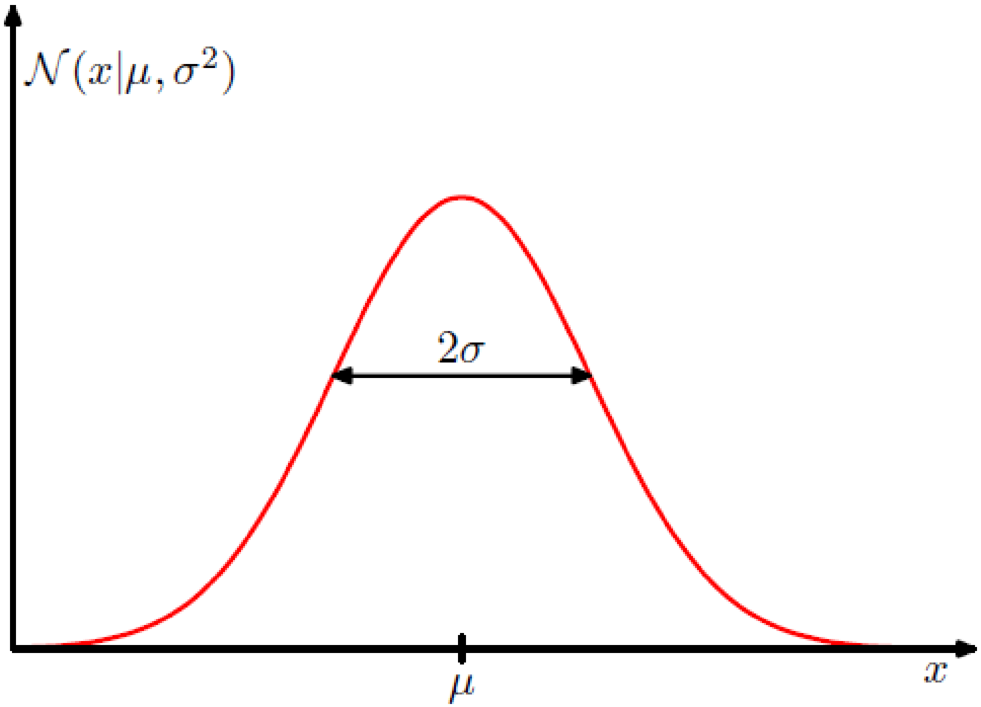
1.2.4節の紹介

- 1.2.4節　ガウス分布

ガウス分布 (正規分布)：１変数

- 平均 μ, 分散 σ^2 の1変数の正規分布（ガウス分布）

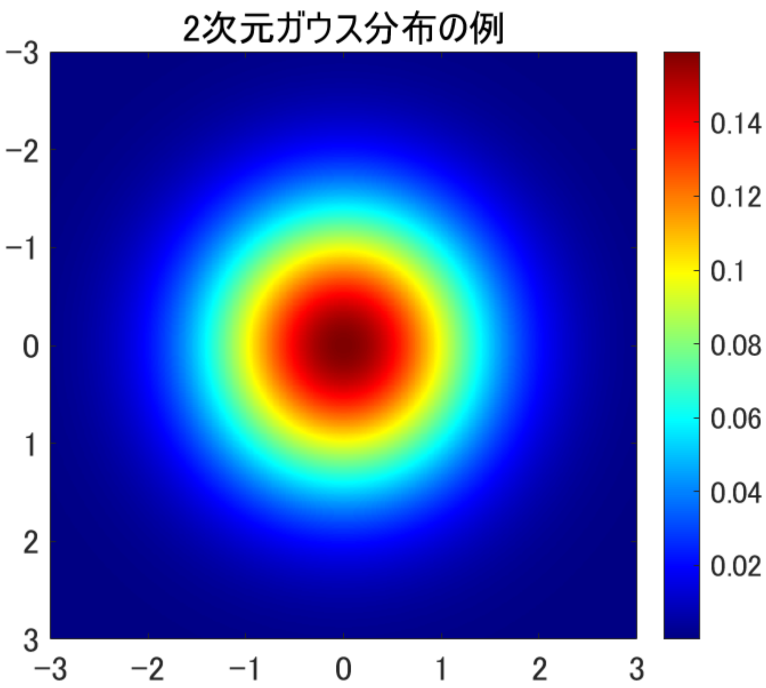
$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$



ガウス分布 (正規分布)：多変数

- D次元変数 $\boldsymbol{x} = (x_1, \dots, x_D)^\top$ の場合。
- 平均ベクトル $\boldsymbol{\mu}$, 分散共分散行列 Σ の多変数正規分布（ガウス分布）の確率密度関数

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma) = \frac{1}{\sqrt{(2\pi)^D(\det\Sigma)}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$



- 平均ベクトル：0
- 分散共分散行列：単位ベクトル

の場合のガウス分布の値（左図）

$$f(\boldsymbol{x}) = \frac{1}{2\pi}\exp\left(-\frac{1}{2}\boldsymbol{x}^\top\boldsymbol{x}\right)$$

独立同一分布

- １変数 x の観測値として N個の値が得られている。
  この値がある分布から独立に生成されたと仮定する。
  ⇒ N個のデータは「独立同一分布に従う」という。

- ある1変数ガウス分布からN個の観測値が独立に得られているとき、この分布の平均、分散を推定することを考える。

- N個の観測データをまとめたベクトル： $\mathsf{x} = (x_1, \ldots, x_N)^\top$
- N個の観測データの同時確率

$$p(\mathsf{x}|\mu, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \sigma^2) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_n-\mu)^2}{2\sigma^2}\right)$$

最尤推定

- パラメータを決めるために最尤推定を使用する。
  - 最尤推定：観測されたデータ集合の(同時)確率を最大にするパラメータを選択する。
  - そのようなパラメータが「最も尤もらしい」と考える。

- 同時確率の式をパラメータ μ, σ^2 の関数とみる。
  ⇒ パラメータの関数としてみたものを「尤度関数」という。

$$f(\mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^N \exp\left(-\frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2 \right)$$

- 尤度関数の対数(対数尤度関数)を利用すると楽になる。

$$\ln f(\mu, \sigma^2) = -\frac{N}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2 + \text{const.}$$

最尤推定の解の計算

$$\ln f(\mu,\sigma^2) = -\frac{N}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 + \mathrm{const.}$$

① μ で偏微分して0になる点を計算。

$$\frac{\partial \ln f(\mu,\sigma^2)}{\partial\mu} = \frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n-\mu) = 0$$

$$\Rightarrow \quad \mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}x_n$$

N個の観測データの平均値

② 同様に σ^2 で偏微分して0になる点を計算。

$$\frac{\partial \ln f(\mu,\sigma^2)}{\partial\sigma^2} = -\frac{N}{2}\frac{1}{\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{n=1}^{N}(x_n-\mu_{\mathrm{ML}})^2 = 0$$

$$\Rightarrow \quad \sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{\mathrm{ML}})^2$$

N個の観測データの分散

最尤推定量の期待値 (1/3)

- パラメータの最尤推定値は観測データから計算されるので、観測データによるバラツキを持つ確率変数である。
⇒「最尤推定量の期待値」について考える。

① 平均(μ)の最尤推定値の期待値

$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n \quad \Rightarrow \quad \mathbb{E}[\mu_{\mathrm{ML}}] = \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}[x_n] = \frac{1}{N}\sum_{n=1}^{N} \mu = \mu$$

ガウス分布の平均： μ

最尤推定値の期待値が真の値と一致

- 推定量の期待値と真値が一致する推定量を「不偏推定量」という。
⇒ 推定値が真値の周りに分布している。
- 不偏推定量は推定量としては望ましいものの1つといえる。

最尤推定量の期待値 (2/3)

② 分散(σ^2)の最尤推定値の期待値

$$\sigma_{\mathrm{ML}}^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{\mathrm{ML}})^2 \quad \Rightarrow \quad \mathbb{E}[\sigma_{\mathrm{ML}}^2] = \frac{1}{N}\mathbb{E}\left[\sum_{n=1}^{N}(x_n - \mu_{\mathrm{ML}})^2\right]$$

- Σの中の計算方法

$$\begin{aligned}(x_n - \mu_{\mathrm{ML}})^2 &= \{(x_n - \mu) - (\mu_{\mathrm{ML}} - \mu)\}^2 \\ &= (x_n - \mu)^2 - 2(x_n - \mu)(\mu_{\mathrm{ML}} - \mu) + (\mu_{\mathrm{ML}} - \mu)^2\end{aligned}$$

$$\therefore \sum_{n=1}^{N}(x_n - \mu_{\mathrm{ML}})^2 = \sum_{n=1}^{N}(x_n - \mu)^2 - N(\mu_{\mathrm{ML}} - \mu)^2$$

nについて和をとって整理する

$$\therefore \mathbb{E}\left[\sum_{n=1}^{N}(x_n - \mu_{\mathrm{ML}})^2\right] = N\sigma^2 - N\mathbb{E}\left[(\mu_{\mathrm{ML}} - \mu)^2\right]$$

期待値計算をして整理する

- 以上より $\mathbb{E}[\sigma_{\mathrm{ML}}^2] < \sigma^2$ が成立。
  ⇒ 分散の最尤推定量は不偏推定量ではない。

最尤推定量の期待値 (3/3)

- 分散の最尤推定量について、標本平均の分散に関する公式を利用。
  ⇒分散の最尤推定量の期待値が厳密に得られる。

$$\mathbb{E}[\sigma_{\mathrm{ML}}^2] = \frac{N-1}{N}\sigma^2$$

分散の最尤推定量の期待値は真値を過小評価している

- 上記の式から、以下の式で定義される「不偏標本分散」の期待値は、分散パラメータの真値と一致する。

$$\tilde{\sigma}^2 := \frac{N}{N-1}\sigma_{\mathrm{ML}}^2 = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \mu_{\mathrm{ML}})^2$$

1.2.5節の紹介

- 1.2.5節　曲線フィッティング再訪

1.2.6節の紹介

- 1.2.6節　ベイズ曲線フィッティング