入門ベイズ統計　by 松原　望

レジュメ作成者・朝倉利晃

1.7補足 ～確率の表記の違いによる意味の違いについて～

前回の勉強会では MLE(maximum likelihood estimation)について、概念と簡単な具体例についてみた。ここでは、P(data), P(θ), P(data|θ), P(θ |data)と Likelihood function の違いを明確にする。ここで data は観測した事象、例えば生徒の点数$x = (x_1, x_2, x_3, \dots, x_n)$、Θ はモデルのパラメータである。例えば正規分布ならば θ = (μ,σ) となる。

P(data), すなわち data が起きる確率については、このままでは求めることが出来ない。我々がある事象の確率を議論するときにはその背景になんらかのモデルやパラメーターを仮定しなければいけない。よく使われる方法とはしては周辺分布(marginal distribution)を用いる方法がある。

$$P(data) = \int_{\theta} P(data|\theta)d\theta \qquad (1)$$

これは、あるモデル上で取りうる全てのパラメーターの値に対して、data が起きる確率を評価し均一に足し合わせたものである。また、周辺分布はベイズの定理の分母に相当する。

P(θ)に関しては、ベイズの定理の下では事前分布として取り扱われる。P(data)における議論と同様に我々は θ に対して何らかのモデルやパラメーターの仮定を置かない限り、θ に関して確率を議論することは出来ない。ベイズの定理の下では往々にして P(θ)は知ることが出来ないため恣意的な選択をせざるを得ない。そこで理由不十分の原則に則り、離散分布の場合、取りうるパラメーターの確率が全て等しくなるように事前分布を与える。すなわち、θが取りうる値が$(\theta_1, \theta_2, \theta_3, .., \theta_n)$であるとき任意の$i, (i = 1,2, \dots, n)$に関して$P(\theta_i) = \frac{1}{n}$とする。連続分布の場合は、P(θ) = 1 としたり（これは improper prior distribution と呼ぶ。パラメーターの全区間を積分すると 1 にならないという意味で improper）、分散が非常に大きい正規分布を仮定したりする。

P(data| θ)はあるモデル、パラメータの仮定のもとでの data が生じる確率である。一般的な確率分布が表しているのは全てこの形である。確率の基礎的な教科書では”X が正規分布に従うとすれば P(X=x)は～”と書いたりすることがあるが、これは P(data | θ)の内、パラメーターの表記を省略している。

僕が MLE(maximum likelihood estimation)を最初に学んだときに非常に戸惑ったのがデータが生じる確率に関する関数と Likelihood function の違いについてだ。前者は P(X| θ)と表され、後者も同様に P(X| θ)と書かれることも多いが、ここでは明示的に L(data| θ)と書く。前者の例としては、data が平均 μ、分散 σ^2 の正規分布に従う変 数 X(X ~ N(μ, σ^2)) だとすると、

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

となり、横軸に確率変数 X、縦軸に P(data| θ)のグラフが書かれる。X に関して積分するとその値は 1 になる。($\int_{-\infty}^{\infty} f(x|\mu, \sigma)\mathrm{d}x = 1$)。
また、data が平均 μ、分散 σ^2 の正規分布から独立に取った n 個の確率変 数 X だとすると、確率は$P(X = x) = \prod_i f(x_i|\mu, \sigma)$ となる。また取りうる値に関して積分すればその値は 1 になる。

一方、Likelihood function の場合は、data が例えば 1 つの観測値で値が 2 だとする。これが分散 σ^2 の正規分布に従うと仮定すると、Likelihood function は

$$\mathrm{L}(\mathrm{data}\mid \mu, \sigma) = \mathrm{f}(2\,|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{2 - \mu}{\sigma}\right)^2\right)$$

となり、横軸に μ、縦軸にL(data | μ, σ)のグラフが書かれる。今回の場合は、μ に関して積分を行っても値は 1 になるが、likelihood function は取りうる値について積分を行った結果が 1 になることは保証されていない。また、data が複数の観測値$x = (x_1, x_2, \dots, x_n)$だとすると、Likelihood function は

$$\mathrm{L}(\mathrm{data}\mid \mu, \sigma) = \mathrm{f}(\mathrm{x}\,|\mu, \sigma) = \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right)$$

となる。この場合でもグラフにすれば横軸に μ、縦軸に L(data|μ,σ)が来る。

以上の違いをまとめると、P(X| θ)は、パラメーターが既知、かつ、データの値が変化したときに確率がどう変化するかを表す。また、変数に関しての積分値は常に 1 である。L(data| θ)は、データが既知、かつ、データに対してパラメーターが変化したときに likelihood(尤度、確からしさ)がどう変化するかを表す。また、変数に関しての積分値は 1 になるとは限らない。

最後に P(θ |data)について考える。我々は θ に関して確率を考えたいが、残念ながら data は θ に関する確率を考えるためのモデルやパラメーターに関して何の知見も与えない。そのため、何らかのモデルやパラメーターを仮定しなければいけないがここでベイズの定理を用いると、θ 自身が θ に関する確率のモデルやパラメーターに関して知見を与えてくれることがわかる。

$$P(\theta|data) = \frac{L(data|\theta)P(\theta)}{P(data)} \tag{2}$$

左辺を事後分布、右辺の分子の第一項目は Likelihood、第二項目は事前分布、分母は周辺分布と呼ばれる。分母が周辺分布と呼ばれる所以は(1)に示したように計算するためにはなんらかの仮定をおいた上で積分操作を行わない限り確率は求められないからだ。

*一言述べて置くと、ベイズの定理の式を用いるとデータがある下でパラメーターの分布を求める際には Likelihood function が必要となるが、これはたまたま Likelihood function がベイズの定理と親和性があるのであって理論としては別物である。

1.7 事後分布

ベイズの定理の確認

-離散値の場合、

原因を θ、結果を z (多くの場合 data)とおいて事前確率を$w(\theta_i)$、事後確率を$w'(\theta_i|z)$、likelihood を$P(z|\theta_i)$と書けば

$$w'(\theta_i|z) = \frac{w(\theta_i)p(z|\theta_i)}{\sum \mathrm{w}(\theta_\mathrm{j})p(z|\theta_j)} \tag{1.7.1}$$

補足の表記に従えば

$$P(\theta_i|data) = \frac{L(data|\theta_i)P(\theta_i)}{P(data)} = \frac{L(data|\theta_i)P(\theta_i)}{\sum L(data|\theta_j)\mathrm{P}(\theta_\mathrm{j})}$$

-連続値の場合、

$$w'(\theta|z) = \frac{w(\theta)p(z|\theta)}{\int_\theta \mathrm{w}(\theta)p(z|\theta)d\theta} \tag{1.7.2}$$

補足の表記に従えば

$$P(\theta|data) = \frac{L(data|\theta)P(\theta)}{P(data)} = \frac{L(data|\theta)P(\theta)}{\int_\theta L(data|\theta)P(\theta)d\theta}$$

分母に関しては定数となるから *MLE を行う際には*分子のみを考えればよい。正確な分布を求める際には分母の積分を行わなければならない。代表的なものの場合、分母の計算が楽な場合がある。

$$w'(\theta|z) \propto w(\theta)p(z|\theta)$$

これらの形式でデータを取り扱う統計学を”ベイズ統計学”
これを用いた決定を”ベイズ統計”という。

1.8 事前分布

Likelihood funcition p(z|θ)に対して“自然な共役事前分布” とは事前分布にこの分布を選べば事後分布も同じ分布族になるような分布を指す。例えば二項分布に対しては、ベータ分布を事前分布とすると事後分布もベータ分布となる。

この章では、Likelihood function とその自然な共役事前分布としてとして次のものが紹介されている。Likelihood function (共役事前分布) として書く。

・二項分布(ベータ分布)
・正規分布(正規分布)
・ポアソン分布(ガンマ分布)

1.8.1 二項分布(ベータ分布)

今、$w_i(i = 1, \dots, n)$は独立な確率変数で θの確率で$x_i = 1,\ 1 - \theta$の確率で$x_i = 0$ であるとする。このとき data が$z = (x_1, \dots, x_n)$のときの likelihood は

$$p(z|\theta) = \ \theta^{\Sigma x_i}(1-\theta)^{n-\Sigma x_i}$$

となる。事前分布, w(θ),としてパラメーターα,β のベータ分布をとる。ベータ関数を

$$B(\alpha, \beta) = \int_0^1 u^{\alpha-1}(1-u)^{\beta-1}du$$

と書けば、

$$w(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}$$

となる。したがって(1.7.2)の計算は

$$w'(\theta|z) = \frac{w(\theta)p(z|\theta)}{\int_\theta w(\theta)p(z|\theta)d\theta}$$

$$= \frac{\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}\theta^{\Sigma x_i}(1-\theta)^{n-\Sigma x_i}}{\int_0^1 \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}\theta^{\Sigma x_i}(1-\theta)^{n-\Sigma x_i}d\theta}$$

$$= \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}\theta^{\Sigma x_i}(1-\theta)^{n-\Sigma x_i}}{\int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1}\theta^{\Sigma x_i}(1-\theta)^{n-\Sigma x_i}d\theta}$$

$$= \frac{\theta^{\Sigma x_i+\alpha-1}(1-\theta)^{n-\Sigma x_i+\beta-1}}{\int_0^1 \theta^{\Sigma x_i+\alpha-1}(1-\theta)^{n-\Sigma x_i+\beta-1}d\theta}$$

$$= \frac{\theta^{\alpha'-1}(1-\theta)^{\beta'-1}}{\int_0^1 \theta^{\alpha'-1}(1-\theta)^{\beta'-1}d\theta}$$

$$= \frac{\theta^{\alpha\prime-1}(1-\theta)^{\beta'-1}}{B(\alpha',\beta')}$$

ここで　$\alpha' = \sum x_i + \alpha,\ \beta' = n - \sum x_i + \beta$

したがってこれはまたベータ分布となる。したがって実際に分布の更新を行う際にはパラメーターの更新だけ行えばよい。

--------例：女の子が生まれる確率----------

ある両親から、連続して男の子が 3 人生まれた。次の子が女の子である確率はどれほどか。

Ans1. 通常の統計学としては男:女=1.05:1.00 の比で生まれてくる。

Ans2. ベイズ統計学としては、次のとおりになる。

事前分布: $\alpha = \beta = 1$のベータ分布、

n = 3 , $\sum x_i = 3$より

事後分布: $\alpha = 4, \beta = 1$のベータ分布

となる。

*注：最初においた 1 については何の客観的な妥当性はない。

1.8.2 正規分布(正規分布)

今、標本$z = (x_1, \dots, x_n)$のそれぞれの x_i が独立な確率変数で平均 θ、分散 σ ^2 の正規分布に従うとすると、標本 z の likelihood p(・|θ)は、

$$p(z|\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{x_i-\theta}{\sigma}\right)^2\right)$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2\right)$$

exp カッコ内の和について

$$\sum_{i=1}^{n}(x_i-\theta)^2 = \sum_{i=1}^{n}(x_i^2 - 2x_i\theta + \theta^2)$$

$$= n\overline{x^2} - 2n\bar{x}\theta + n\theta^2$$

$$= n(\theta-\bar{x})^2 + n\left(\overline{x^2} - \bar{x}^2\right)$$

$$= n(\theta-\bar{x})^2 + \sum(x_i-\bar{x})^2$$

よって

$$p(z|\theta) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\left(n(\theta-\bar{x})^2 + \sum(x_i-\bar{x})\right)\right)$$

$$= \exp\left(-\frac{n(\theta-\bar{x})}{2\sigma^2}\right)\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{\sum(x_i-\bar{x})^2}{2\sigma^2}\right) \qquad (1.8.7)$$

興味あるのは θ に関してであるので第一因子のみが情報として寄与する。

θ の事前分布が N(μ,τ^2)に従うとする。

$$w(\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{\theta-\mu}{\tau}\right)^2\right) \qquad (1.8.8)$$

これらを用いれば事後分布は次のように表される。ただし、正規分布の更新の際には標本の分散は一定であると仮定している。

$$w'(\theta|z)=\frac{w(\theta)p(z|\theta)}{\int_\theta \mathrm{w}(\theta)p(z|\theta)d\theta}$$
$$=\frac{\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2}\left(\frac{\theta-\mu}{\tau}\right)^2\right)\exp\left(-\frac{n(\theta-\bar{x})}{2\sigma^2}\right)\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n\exp\left(-\frac{\sum(x_i-\bar{x})^2}{2\sigma^2}\right)}{\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2}\left(\frac{\theta-\mu}{\tau}\right)^2\right)\exp\left(-\frac{n(\theta-\bar{x})}{2\sigma^2}\right)\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n\exp\left(-\frac{\sum(x_i-\bar{x})^2}{2\sigma^2}\right)d\theta}$$
$$=\frac{\exp\left(-\frac{1}{2}\left(\frac{\theta-\mu}{\tau}\right)^2\right)\exp\left(-\frac{n(\theta-\bar{x})}{2\sigma^2}\right)}{\int_{-\infty}^{\infty}\exp\left(-\frac{1}{2}\left(\frac{\theta-\mu}{\tau}\right)^2\right)\exp\left(-\frac{n(\theta-\bar{x})}{2\sigma^2}\right)d\theta} \qquad (1.8.9.5)$$

分子に関して計算してあげると

$$\exp\left(-\frac{1}{2}\left(\frac{\theta-\mu}{\tau}\right)^2\right)\exp\left(-\frac{n(\theta-\bar{x})}{2\sigma^2}\right)=\exp\left(-\frac{(\theta-\mu')}{2\tau'^2}\right)\exp\left(-\frac{n(\mu-\bar{x})}{2(n\tau^2+\sigma^2)}\right)$$

ここで

$$\frac{1}{\tau'^2}=\frac{1}{\tau^2}-\frac{n}{\sigma^2}$$
$$\mu'=\frac{\left(\frac{1}{\tau^2}\right)\mu+\left(\frac{n}{\sigma^2}\right)\bar{x}}{\frac{1}{\tau^2}+\frac{n}{\sigma^2}}$$

である。これを(1.8.9.5)の中に叩き込んであげると

$$w'(\theta|z)=\frac{\exp\left(-\frac{(\theta-\mu')}{2\tau'^2}\right)\exp\left(-\frac{n(\mu-\bar{x})}{2(n\tau^2+\sigma^2)}\right)}{\int_{-\infty}^{\infty}\exp\left(-\frac{(\theta-\mu')}{2\tau'^2}\right)\exp\left(-\frac{n(\mu-\bar{x})}{2(n\tau^2+\sigma^2)}\right)d\theta}$$
$$=\frac{\exp\left(-\frac{(\theta-\mu')}{2\tau'^2}\right)}{\int_{-\infty}^{\infty}\exp\left(-\frac{(\theta-\mu')}{2\tau'^2}\right)d\theta}$$
$$=\frac{1}{\sqrt{2\pi\tau'^2}}\exp\left(-\frac{(\theta-\mu')}{2\tau'^2}\right)$$

となるので、結局、$w'(\theta|z)\propto\exp\left(-\frac{(\theta-\mu')}{2\tau'^2}\right)$ (1.8.10) が言えるのである。

1.8.3 ポアソン分布 (ガンマ分布)

今までと同様な手順で$x_i(i=1,..,n)$を独立な確率変数として、平均 θ のポアソン分布に従うとする。$z=(x_1,\dots,x_n)$とすれば likelihood は

$$p(z|\theta)=\prod_{i=1}^{n}\frac{e^{-\theta}\theta^{x_i}}{x_i!}$$

事前分布を parameter λ,α のガンマ分布とする。

$$w(\theta)=e^{-\alpha\theta}\theta^{\lambda-1}\frac{\alpha^\lambda}{\Gamma(\lambda)}$$

ただし、$\Gamma(s)=\int_0^\infty e^{-u}u^{s-1}\,du$ である。
これらを用いて事後分布を計算すると

$$w'(\theta|z)=\frac{w(\theta)p(z|\theta)}{\int_\theta \mathrm{w}(\theta)p(z|\theta)d\theta}$$
$$=\frac{e^{-\alpha\theta}\theta^{\lambda-1}\frac{\alpha^\lambda}{\Gamma(\lambda)}\prod_{i=1}^{n}\frac{e^{-\theta}\theta^{x_i}}{x_i!}}{\int_0^\infty e^{-\alpha\theta}\theta^{\lambda-1}\frac{\alpha^\lambda}{\Gamma(\lambda)}\prod_{i=1}^{n}\frac{e^{-\theta}\theta^{x_i}}{x_i!}d\theta}$$
$$=\frac{e^{-\alpha\theta}\theta^{\lambda-1}\prod_{i=1}^{n}e^{-\theta}\theta^{x_i}}{\int_0^\infty e^{-\alpha\theta}\theta^{\lambda-1}\prod_{i=1}^{n}e^{-\theta}\theta^{x_i}\,d\theta}$$
$$=\frac{e^{-\alpha\theta-n\theta}\theta^{\lambda-1+\sum x_i}}{\int_0^\infty e^{-\alpha\theta-n\theta}\theta^{\lambda-1+\sum x_i}d\theta}$$
$$=\frac{e^{-\alpha'\theta}\theta^{\lambda'-1}}{\int_0^\infty e^{-\alpha'\theta}\theta^{\lambda'-1}d\theta}$$

したがって、$w'(\theta|z)\propto e^{-\alpha'\theta}\theta^{\lambda'-1}$ (1.8.15) が得られる。
ただし、$\alpha'=\alpha+n,\lambda'=\lambda+\sum x_i$　である。

--------例：平均火災件数----------
新人消防署長は管内の 1 ヶ月あたりの火災件数の平均を 10 土√10 (土は標準偏差)とみた。実際は、最初の四半期に 23 件であった。彼の予想はどう変わったか。

Ans.
まずモデルとして、事前分布にガンマ分布、それぞれの独立変数がポアソン分布に従うとする。(1.8.3 の設定と同じ)。
事前分布の α,β を定めるために、

$$E(\theta) = \frac{\lambda}{\alpha} = 10\,, V(\theta) = \frac{\lambda}{\alpha^2} = 10$$

を解くと$\lambda = 10, \alpha = 1$ となる。

最初の四半期に 23 件は換言すれば、一ヶ月に起きる火災件数がポアソン分布に従うならば、ポアソン分布に従う 3 つの確率変数の実現値の和が 23 件と言える。
よって $n = 3,\ \sum x_i = x_1 + x_2 + x_3 = 23$ より$\lambda' = 10 + 23 = 33\,, \alpha' = 1 + 3 = 4$ である。これは平均 33/4 = 5.75, 分散 23/16=1.44 のガンマ分布である。