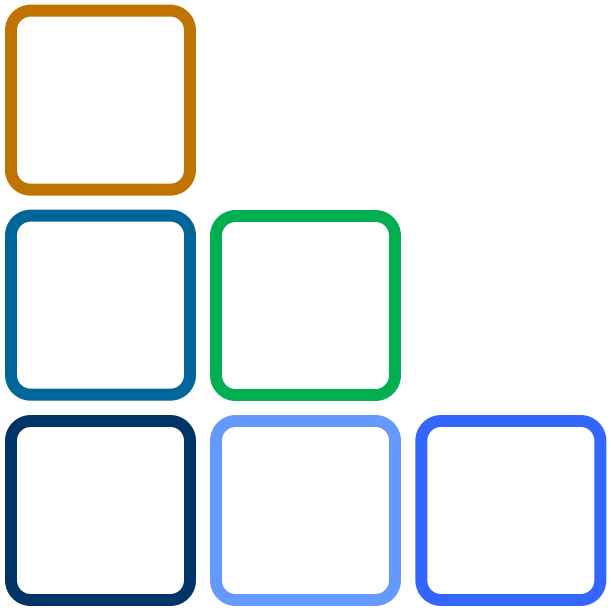
Pattern Recognition and Machine Learning
2. Probability Distribution



2.1 Binary Variables

北海道大学 大学院情報科学研究科 CS専攻 情報知識ネットワーク研究室 鈴木康広

6月15日

イントロダクション

ベータ分布(Beta Distribution)

- ベータ分布の定義
- 過学習
- ベータ分布の期待値と分散
- ベイズ的ベータ分布
- 超パラメータ
- 逐次学習

6月15日

ベータ分布(1)

仮に、コインの表裏（表の時x＝1,裏の時、x＝0）を選ぶ問題を考える。3回コインを投げ、3回とも表だった時、最尤推定を考えると4回以降も全て表が出続けると予測される。

この方法では「過学習」してしまう可能性がある。

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

6月15日

ベータ分布(2)

2値変数をベイズ的に扱うために、パラメータμに関する事前分布p(μ)を考える。

事後確率 $\propto$ 尤度関数 × 事前確率

$$Bin(m \mid N, \mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m} \qquad Bern(x \mid \mu) = \mu^x (1-\mu)^{1-x}$$

もし事前確率が$\mu^x(1-\mu)^{1-x}$に比例するように選ぶ時、事後確率は尤度関数と事前確率の積に比例するので、事後確率も同じ形になる(conjugacy , 共役性) 。

6月15日

ベータ分布の定義

$$Beta(\mu \mid a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1}(1-\mu)^{b-1}$$

で定義され、また

$$\int_0^1 Beta(\mu \mid a,b) d\mu = 1$$

を満たしており、平均と分散はそれぞれ、

$$E[\mu] = \frac{a}{a+b} \qquad \mathrm{var}[\mu] = \frac{ab}{(a+b)^2(a+b+1)}$$

6月15日

ガンマ関数

$$\Gamma(a) = \int_0^\infty x^{a-1} e^{-x} dx$$

$$\Gamma(a) = \int_0^\infty x^{a-1} e^{-x} dx$$

$$= \int_0^\infty x^{a-1} (-e^{-x})' dx$$

$$= \left[-x^{a-1} e^{-x} \right]_0^\infty + (a-1) \int_0^\infty x^{a-2} e^{-x} dx$$

$$= 0 + (a-1) \int_0^\infty x^{a-2} e^{-x} dx$$

$$= (a-1)\Gamma(a-1)$$

$$\Gamma(a) = (a-1)!$$

$$\Gamma(1) = 1$$

6月15日

ベータ分布について詳説

$$\int_0^1 Beta\,(\mu \mid a,b)d\,\mu = 1$$

$$\int_0^1 Beta\,(\mu \mid a,b)d\,\mu = \int_0^1 \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}d\,\mu = 1$$

$$\int_0^1 \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \times \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}d\,\mu = 1$$

$$\begin{aligned}\Gamma(a)\Gamma(b) &= \int_0^\infty e^{-x}x^{a-1}dx\int_0^\infty e^{-y}y^{b-1}dy \\ &= \int_0^\infty x^{a-1}\left\{\int_0^\infty e^{-(x+y)}y^{b-1}dy\right\}dx \\ &= \int_0^\infty x^{a-1}\left\{\int_x^\infty e^{-t}(t-x)^{b-1}dt\right\}dx\end{aligned}$$

定義から
自然対数をまとめる
t=x+yに置換する，つまりy=t−x

$$\Gamma(a)\Gamma(b) = \int_0^\infty x^{a-1}\left\{\int_x^\infty e^{-t}(t-x)^{b-1}dt\right\}dx$$

$$= \int_0^\infty \int_0^t x^{a-1}e^{-t}(t-x)^{b-1}dxdt$$

二重積分の形

$$= \int_0^\infty \int_0^1 (t\mu)^{a-1}e^{-t}(t-t\mu)^{b-1}t\,d\mu\,dt$$

x=tμに置換する

$$= \int_0^\infty \int_0^1 t^{a-1}\times\mu^{a-1}\times e^{-t}\times t^{b-1}\times(1-\mu)^{b-1}\times t\,d\mu\,dt$$

$$= \int_0^\infty e^{-t}t^{a+b-1}dt\int_0^1 \mu^{a-1}(1-\mu)^{b-1}d\mu$$

t,μについての積分に分解する

$$\Gamma(a)\Gamma(b) = \Gamma(a+b)\int_0^1 \mu^{a-1}(1-\mu)^{b-1}d\mu$$

$\Gamma(a+b) = \int_0^\infty x^{a+b-1}e^{-x}dx$ を代入

$$\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} = \int_0^1 \mu^{a-1}(1-\mu)^{b-1}d\mu$$

Γ(a+b)で両辺で割る

$$\int_0^1 Beta(\mu\,|\,a,b)d\mu = \int_0^1 \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}d\mu = 1$$

中辺に上記の式を代入

$$\int_0^1 \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\times\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}d\mu = 1$$

以上

6月15日

期待値の求め方

$$E(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 \mu^{a-1}(1-\mu)^{b-1} \times \mu \, d\mu$$

$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 \mu^{a}(1-\mu)^{b-1} d\mu$$

$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \left(\left[\mu^a - \frac{1}{b}(1-\mu)^b \right]_0^1 + \int_0^1 a\,\mu^{a-1} \times \frac{1}{b}(1-\mu)^b \, d\mu \right)$$

$$= \frac{a\Gamma(a+b)}{b\Gamma(a)\Gamma(b)} \int_0^1 \mu^{a-1}(1-\mu)^{b} d\mu$$

$$= \frac{a\Gamma(a+b)}{b\Gamma(a)\Gamma(b)} \int_0^1 \mu^{a-1}(1-\mu)^{b-1}(1-\mu) d\mu$$

$$= \frac{a\Gamma(a+b)}{b\Gamma(a)\Gamma(b)} \left(\int_0^1 \mu^{a-1}(1-\mu)^{b-1} d\mu - \int_0^1 \mu^{a}(1-\mu)^{b-1} d\mu \right)$$

$$= \frac{a}{b} \left(1 - \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 \mu^{a}(1-\mu)^{b-1} d\mu \right)$$

$$= \frac{a}{b}(1 - E(x))$$

$$E(x) = \frac{a}{b}(1 - E(x))$$

$$bE(x) = a(1 - E(x))$$

$$(a+b)E(x) = a$$

$$E(x) = \frac{a}{a+b}$$

6月15日

分散の求め方

$$E(x^2) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 \mu^{a-1}(1-\mu)^{b-1} \times \mu^2 d\mu$$

$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 \mu^{a+1}(1-\mu)^{b-1} d\mu$$

$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \left(\left[(\mu^{a+1} \times \frac{(-1)}{b}(1-\mu)^b \right]_0^1 + \int_0^1 (a+1)\mu^a \times \frac{1}{b}(1-\mu)^b d\mu \right)$$

$$= \frac{(a+1)\Gamma(a+b)}{b\Gamma(a)\Gamma(b)} \int_0^1 \mu^a (1-\mu)^b d\mu$$

$$= \frac{(a+1)\Gamma(a+b)}{b\Gamma(a)\Gamma(b)} \int_0^1 \mu^{a-1}(1-\mu)^{b-1} \mu(1-\mu) d\mu$$

$$= \frac{(a+1)\Gamma(a+b)}{b\Gamma(a)\Gamma(b)} \left(\int_0^1 \mu^a (1-\mu)^{b-1} d\mu - \int_0^1 \mu^{a+1}(1-\mu)^{b-1} d\mu \right)$$

$$= \frac{(a+1)}{b} \left(\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 \mu^a (1-\mu)^{b-1} d\mu - \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 \mu^{a+1}(1-\mu)^{b-1} d\mu \right)$$

$$= \frac{a+1}{b} \left(E(x) - E(x^2) \right)$$

6月15日

分散の求め方

$E(x)=\frac{a}{a+b}$ だから、

$$E(x^2)=\frac{a+1}{b}\left(E(x)-E(x^2)\right)$$

$$E(x^2)=\frac{a+1}{b}\left(\frac{a}{a+b}-E(x^2)\right)$$

$$E(x^2)+\frac{a+1}{b}E(x^2)=\frac{a(a+1)}{b(a+b)}$$

$$\left(\frac{a+b+1}{b}\right)E(x^2)=\frac{a(a+1)}{b(a+b)}$$

$$E(x^2)=\frac{a(a+1)}{(a+b)(a+b+1)}$$

$\mathrm{var}(x)=E(x^2)-\left(E(x)\right)^2$ だから、

$$\mathrm{var}(x\ )=\frac{a(a+1)}{(a+b)(a+b+1)}-\left(\frac{a}{a+b}\right)^2$$

$$=\frac{a(a+1)(a+b)-a^2(a+b+1)}{(a+b)^2(a+b+1)}$$

$$=\frac{(a^2+a)(a+b)-a^2(a+b+1)}{(a+b)^2(a+b+1)}$$

$$=\frac{a^3+(1+b)a^2+ab-a^3-a^2b-a^2}{(a+b)^2(a+b+1)}$$

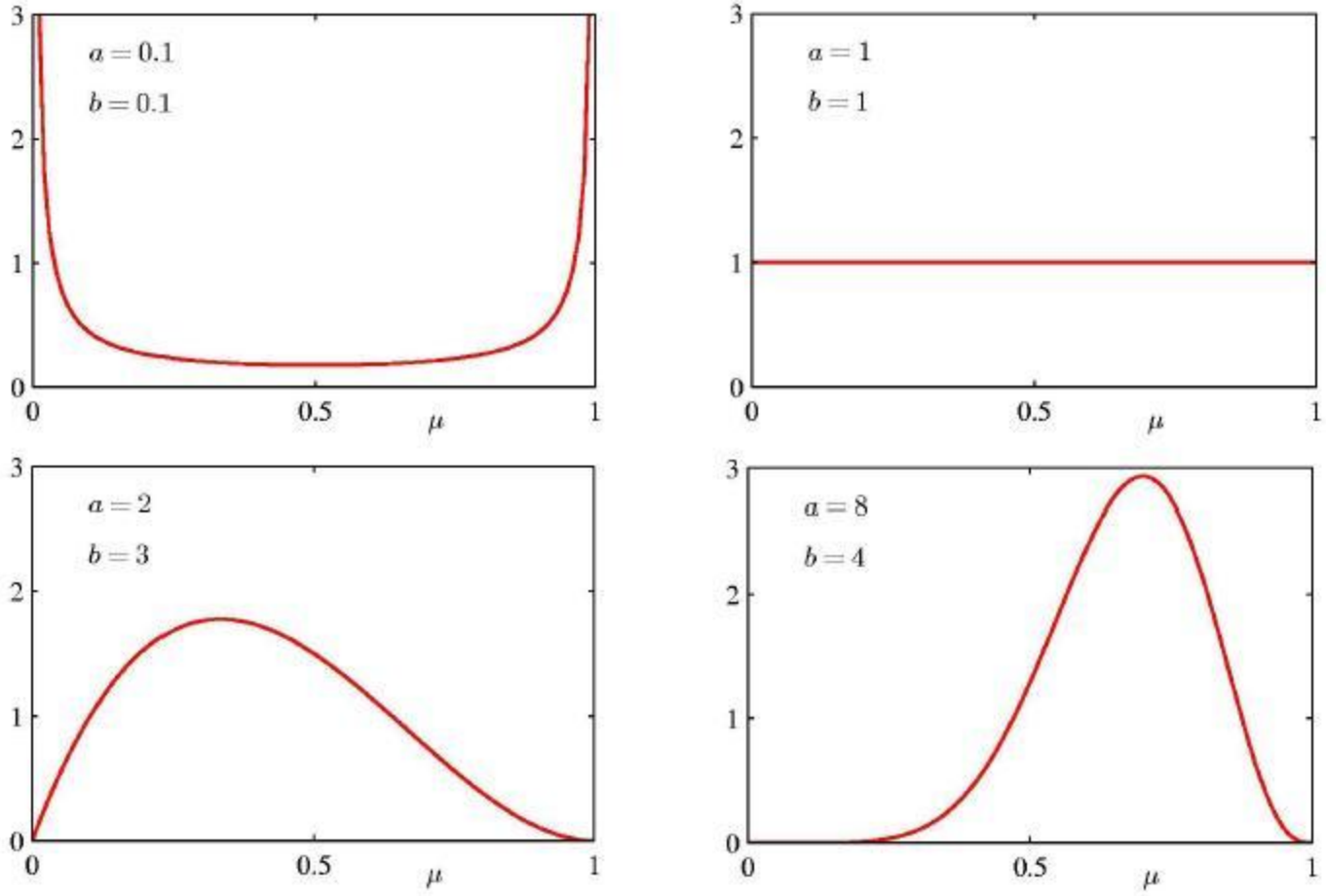
$$=\frac{ab}{(a+b)^2(a+b+1)}$$

6月15日

11

超パラメータ

今示した通り、パラメータμの分布はパラメータa,bによって決まり、このようなパラメータを超パラメータ(hyperparameter)と呼ぶ。



6月15日

μの事後分布

μの事後分布は、ベータの事前分布と二項尤度関数の積で得られる。

$$Beta(\mu \mid a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \quad Bin(m \mid N,\mu) = \binom{N}{m}\mu^{m}(1-\mu)^{N-m}$$

μに依存する要素のみを取り出すと以下のようになる。

$$p(\mu \mid m,l,a,b) \propto \mu^{m+a-1}(1-\mu)^{l+b-1}$$

ただしl=N-mとする。共役性から尤度関数が事前分布と同じ関数形式を取る時事後分布の形式も同様の関数形式を取るため、事後分布ベータも同様のベータ分布の形で表すことができる。

6月15日

ベイズ的ベータ分布

$$p(\mu \mid m,l,a,b) = \frac{\Gamma(m+a+l+b)}{\Gamma(m+a)\Gamma(l+b)} \mu^{m+a-1}(1-\mu)^{l+b-1}$$

観測値から事後分布を求めるためには、aの値をmだけ、bの値をlだけ増やすことで求められる。

「mもlも自然数であるがa,bは自然数である必要はない。」

次に、観測したデータを追加して観測を続ける事を考える。つまり各時刻にデータを1つずつ観測されるとし、各観測後新しい観測値に対する尤度関数を掛け、正規化を行う。結果新しい事後分布に更新する。

6月15日

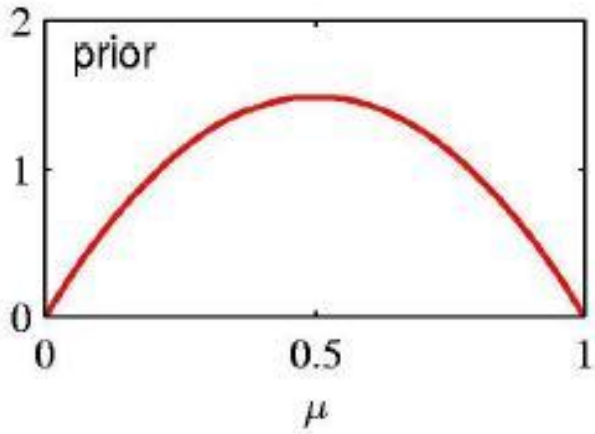
逐次学習(sequential learning)

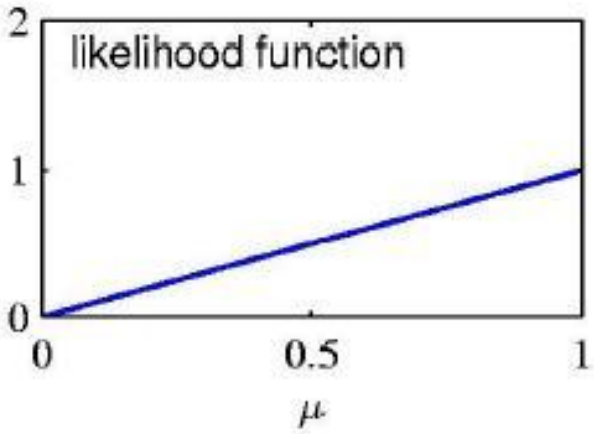
$$p(\mu \mid m,l,a,b) = \frac{\Gamma(m+a+l+b)}{\Gamma(m+a)\Gamma(l+b)} \mu^{m+a-1}(1-\mu)^{l+b-1}$$

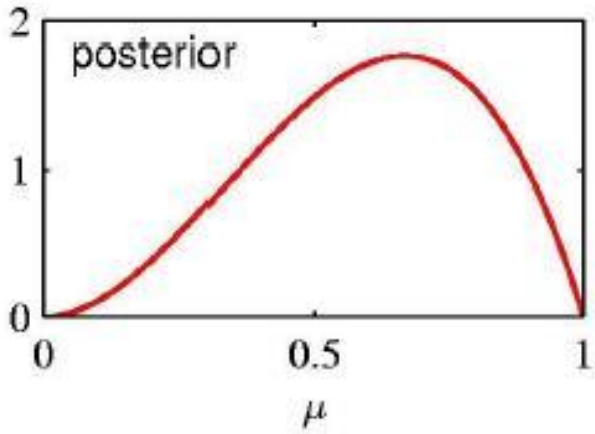
観測データがx＝1の時、aに1を追加し、x＝0の時、bに1を追加する。

変更したaとbに更新した事後確率を用いてデータを観測する。

データを観測する。







- 事前分布：a＝b＝2の時のベータ分布
- 尤度関数：x＝1を観測した時の尤度関数(Bin(m|N,μ)＝μ)。
- 事後分布：x＝1を観測したのでaを1加算した時、a＝3,b＝2の時の事後分布

- 逐次学習のアプローチは事前分布と尤度関数には無関係である。
- データがそれぞれ独立であれば成り立つ。
- また、少数の観測値から、事後分布を更新し、そのデータを捨てるようなメモリが少量である必要がある場合に有用である。

6月15日

二値変数の予測分布(1)

正確に次の出力を予測するために、p(x=1|D)の分布を考える。

$$p(x=1|D)=\int_0^1 p(x=1|\mu)p(\mu|D)d\mu=\int_0^1 \mu\, p(\mu|D)d\mu=E[\mu|D]$$

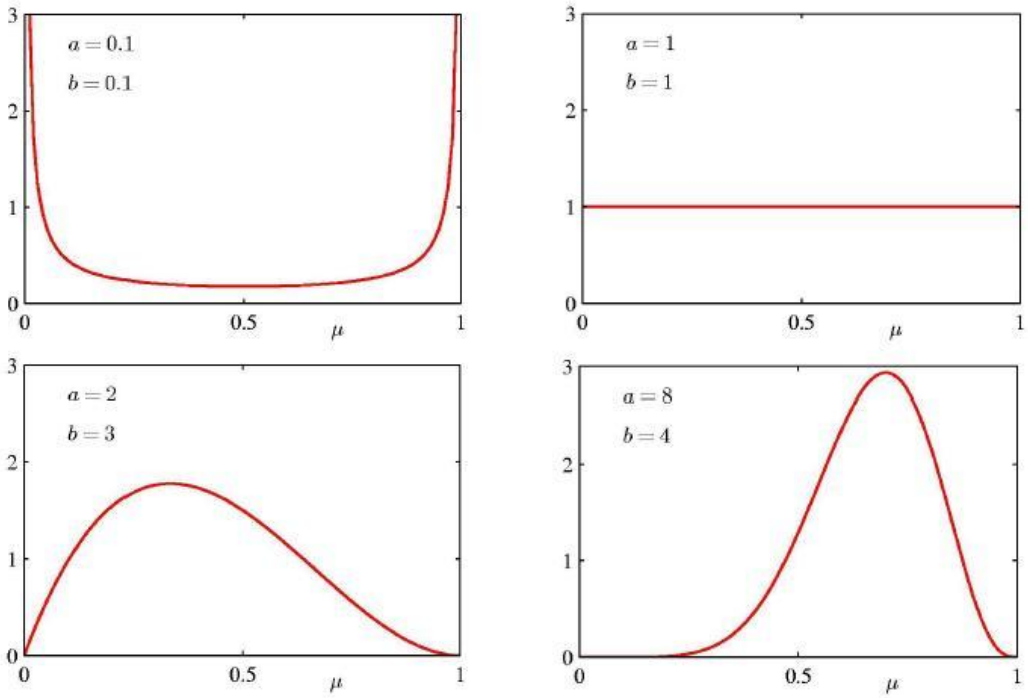
ベータ分布の平均 $E(x)=\frac{a}{a+b}$ から

$$p(x=1|D)=\frac{m+a}{m+a+l+b}$$

観測値の総数に対する、観測値がx=1の割合と考えることができる。

6月15日

二値変数の予測分布(1)



aやbが増えれば増えるほど、最大値(ピーク)は鋭くなる。

仮にa→∞、b→∞の時、分散は0になる。

データ数が多ければ多いほど不確実性は弱まる。