Chapter 2. Probability Distributions

> [p67] One role for the distributions discussed in this chapter is to model the probability distribution $p(x)$ of a random variable x

この感覚は大事な気がする．なんというか「正規分布やポアソン分布からデータが発生している」とはあんまり思っていなくて，完全に把握することは不可能な未知の(複雑な)分布 $p(x)$ からデータが発生していて，それをモデル化(あくまで近似)するための分布として，正規分布やポアソン分布を使う，と思っている感じ．

> [p68] In a frequentist treatment, we choose specific values for the parameters by optimizing some criterion, such as the likelihood function. By contrast, in a Bayesian treatment we introduce prior distributions over the parameters and then use Bayes' theorem to compute the corresponding posterior distribution given the observed data.

この頻度主義とベイズ主義の対比，分かりやすい．ベイズ主義では事前情報(主観,知見)を事前分布を通して使うが，頻度主義では一切使わずデータだけを頼る．

> [p68] We shall see that an important role is played by conjugate priors, that lead to posterior distributions having the same functional form as the prior, and that therefore lead to a greatly simplified Bayesian analysis.

共役事前分布はあまりやってこなかったから，重視して読む．

2.1. Binary Variables

ベルヌーイ分布と二項分布については，メモ省略．

2.1.1. The beta distribution

話の流れは...

1. データ集合が小さいと，最尤法では非常に過学習してしまう (偶然 3 枚のコインが表だったら今

後全て表が出ると未来永劫表が出ると予測されてしまう) ことがある.
2. これを避けるため，事前情報(常識)を使ってベイズ主義的に推定したい.
3. そのためにはまず事前情報をどういう事前分布で表せば良いか決めないといけない.
4. 事前分布を決める指針(望ましい性質)として，共役性つまり「事後分布と事前分布の関数型が同じになる」ってのがある．(なぜ嬉しいかは例えば後述の逐次学習を参照)
5. ここで，ベイズの定理より「パラメータ事後分布 $\propto$ 尤度 $\times$ パラメータ事前分布」である.
6. 尤度関数(今回の例だと $\prod \mu_n^x (1-\mu)^{1-x_n}$) は既知なので，こういう事前分布なら事後分布も同じ形になるな，ってやつを探そう.
7. 今回の例(ベルヌーイ分布の μ の推定)では，ベータ分布がまさにそれ.

という感じ.

事後分布 (2.17) の導出を丁寧めに書くと,

$$
\begin{aligned}
p(\mu|m,l,a,b) &= \frac{p(\mu,m,l|a,b)}{p(m,l)} \\
&= \frac{p(m,l|\mu)\, p(\mu|a,b)}{p(m,l)} \\
&\propto p(m,l|\mu)\, p(\mu|a,b) \\
&= {}_N C_m\ \mu^m\ (1-\mu)^l\ \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1}(1-\mu)^{b-1} \\
&\propto \mu^{m+a-1}(1-\mu)^{l+b-1}
\end{aligned}
$$

となり,

> [p72] We see that (2.17) has the same functional dependence on μ as the prior distribution, reflecting the conjugacy properties of the prior with respect to the likelihood function. Indeed, **it is simply another beta distribution**, and its normalization coefficient can therefore be obtained by comparison with (2.13) to give

$$
p(\mu|m,l,a,b) = \frac{\Gamma(m+a+l+b)}{\Gamma(m+a)\Gamma(l+b)} \mu^{m+a-1}(1-\mu)^{l+b-1} \qquad (2.18)
$$

という感じで簡単に事後分布が求まる．さらに嬉しいのは,

> [p72] We see that the effect of observing a data set of m observations of $x = 1$ and l observations of $x = 0$ has been to increase the value of a by m, and the value of b by l, in going from the prior distribution to the posterior distribution.

ということ．つまり，新たに得たデータを使って事後分布に更新するには，クラス1の個数とクラス0の個数の分だけベータ分布のパラメータを変えれば良いだけ．(なので最初のベータ分布のパラメータは，事前に何回ずつ表裏が出ているか，みたいな意味合いになる．未知なら$a = b = 1$が無難．)

この利点を活かし「データを観測して事後分布更新 (ベータ分布パラメータ更新)」を繰り返すことを **逐次学習** (図2.2を見るとイメージつきやすい) と呼び,

> [p73] They can be used, for example, in real-time learning scenarios where a steady stream of data is arriving, and predictions must be made before all of the data is seen. Because they do not require the whole data set to be stored or loaded into memory, sequential methods are also useful for large data sets.

などの現実的な大きいメリットがある.

次に，クラスの事後予測密度 (2.19) を丁寧めに導出してみると，

$$
\begin{aligned}
p(x|\mathcal{D}) &= \int p(x, \mu|\mathcal{D})\, d\mu \\
&= \int p(x|\mu, \mathcal{D})\, p(\mu|\mathcal{D})\, d\mu \\
&= \int p(x|\mu)\, p(\mu|\mathcal{D})\, d\mu \\
&= \int \mu^x\, (1-\mu)^{1-x}\, p(\mu|\mathcal{D})\, d\mu \\
&= \begin{cases} \int \mu\, p(\mu|D)\, d\mu & [x=1] \\ \int (1-\mu)\, p(\mu|D)\, d\mu & [x=0] \end{cases} \\
&= \begin{cases} E[\mu \sim p(\mu|D)] & [x=1] \\ 1 - E[\mu \sim p(\mu|D)] & [x=0] \end{cases} \\
&= \begin{cases} \frac{m+a}{m+a+l+b} & [x=1] \\ 1 - \frac{m+a}{m+a+l+b} & [x=0] \end{cases}
\end{aligned}
$$

という感じになる．最後の等号では，事後分布(2.18)がベータ分布であることとベータ分布の期待値(2.15)を使っている．さらにこの結果について

> [p73] Note that in the limit of an infinitely large data set $m, l \to \infty$ the result (2.20) reduces to the maximum likelihood result (2.8).

とあるが，これは $m, l \to \infty$ だと a, b が相対的ににゴミみたいに小さくなって，

$$
p(x|\mathcal{D}) \approx \begin{cases} \frac{m}{m+l} & [x=1] \\ 1 - \frac{m}{m+l} & [x=0] \end{cases}
$$

となり頻度主義で最尤推定した時のクラス予測分布と一致する，てことを言ってる．
別の捉え方としては「パラメータ事後分布(ベータ分布)の分散が $\to 0$ となって最尤推定量(ベータ分布期待値)のとこをピークとした超鋭い分布になる．これってパラメータを最尤推定してるのとほぼ同じだよね」という感じ．これについて，

> [p73] it is a very general property that the Bayesian and maximum likelihood results will agree in the limit of an infinitely large data set

と言ってる．直感的に，観測されたデータが増えればパラメータの不確実性(分散)は減る感じがするし，まあ納得できる．

2.2. Multinomial Variables

ノーテーションが若干違ってくるだけで，２クラスのときと同じことが成り立つ．

ざっくりした対応表としては，

クラス数	単一の試行の結果	複数回の試行の集計	共役事前分布
2クラス	ベルヌーイ分布	二項分布	ベータ分布
多クラス	カテゴリカル分布	多項分布	ディリクレ分布

という感じ.