1. Binary, Beta, Multinomial, Dirichlet Distribution

(1) Binary Variable.

for a R.V. $x \in \{0, 1\}$ $p(x=1) = \mu$. we can have

$$\text{Bern}(x) = \mu^x (1-\mu)^{1-x} \iff \text{Bern}(x) = \begin{cases} \mu & x=1 \\ 1-\mu & x=0 \\ 0 & \text{else} \end{cases}$$

Bernulli Distribution

$$\mathbb{E}(X) = \mu \cdot 1 + (1-\mu) \cdot 0 = \mu$$

$$\text{Var}(X) = \mu(1-\mu)^2 + (1-\mu)(0-\mu)^2$$

$$= \mu(1-\mu)(1-\mu+\mu)$$

$$= \mu(1-\mu)$$

If we repeat the experiment N times and collect $D = [x_1, x_2 \dots x_N]$. the joint pmf would be

$$p(D) = \prod_{i=1}^{N} p(x_i) = \prod_{i=1}^{N} \mu^{x_i} (1-\mu)^{1-x_n} = \mu^{\sum x_i} (1-\mu)^{N-\sum x_i}$$

Use the traditional logorithm trick

$$\ln p(D) = \sum x_i \ln \mu + (N - \sum x_i) \ln (1-\mu)$$

$$\frac{\partial}{\partial \mu} \ln p(D) = \sum x_i \cdot \frac{1}{\mu} + (N - \sum x_i) \frac{-1}{1-\mu} = 0 \quad (\text{MLE})$$

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

Problem for MLE.

We have derived the estimation formula for μ that

$$\hat{\mu}_{ML} = \frac{1}{N}\sum_i x_i$$

But if our experiment can not repeat for sufficiently large times, MLE may have very large bias.

E.g. We flip a coin, 1 - head, 0 - tail. We see 3 head among 3 experiment. According to $\hat{\mu}_{ML} = \frac{\#\text{ of head}}{\#\text{ of exp}} = 1$, which is far away from our common sense. This is an instance of over-fitting.

We can use another model to describe the experiment when we repeat the flipping coin experiment N times, which is **Binomial Distribution**

for $x \in \{0, 1\}$. $P(x=1) = \mu$. we repeat the exp N times, we see $x=1$ for m times.

$$P(m \mid N, \mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m}$$

$$E(m) = \sum_{i=1}^{N} P(x_i=1)\, x_i = N\mu$$

$$Var(m) = \sum_{i=1}^{N} Var(x_i=1) = N\mu(1-\mu)$$

(2). Beta Distribution. (Conjugate Prior)

In part 1, we know that MLE can have over-fit problem when data size is small. We can introduce Bayesian's method to improve this method. We want to turn MLE to MAP

From the perspective of Bayesian, parameter μ also satisfies a distrib-ution, which is called prior. For each experiment, μ may change. All we can have is the posterior and likelihood.

$$P(\mu|x) = \frac{P(\mu)\cdot P(x|\mu)}{P(x)}$$

First, we need to choose a prior. For the likelihood function

$$P(m|N,\mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m}$$

We want to find a form/model for prior which can make likelihood of the same form as posterior. In this case, likelihood/posterior is proportional to $\mu^m(1-\mu)^{N-m}$.

We choose Beta distribution

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

remind you of $\binom{a+b}{a}$?

This is a continuous function though.

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t}\, dt$$

$\Gamma(x)$ has very interesting properties.

① $\Gamma(x+1) = x\Gamma(x)$

② $\Gamma(1) = 1$

③ $\Gamma(x) = (x-1)!$ if x is integer

Proof

1. $\Gamma(x+1) = \int_0^\infty t^x e^{-t} dt = -\int_0^\infty t^x de^{-t} = \underbrace{-t^x e^{-t}\Big|_0^\infty}_{0} + \int_0^\infty e^{-t} dt^x$

$= \int_0^\infty e^{-t} dt^x = \int_0^\infty e^{-t} x \cdot t^{x-1} dt = x\int_0^\infty t^{x-1} e^{-t} dt$

$= x \cdot \Gamma(x)$

2. $\Gamma(1) = \int_0^\infty t^0 e^{-t} dt = -e^{-t}\Big|_0^\infty = 1$

3. $\Gamma(x) = (x-1)\Gamma(x-1) = \prod_{i=1}^{x-1}(x-i) = (x-1)!$

Let's prove that Beta distribution is a valid distribution.

Target: $\int_0^1 \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1}(1-\mu)^{b-1} d\mu = 1$ for $\forall a, b \in (0, \infty)$

Instead of proving brutally, we turn to prove.

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

Proof:

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

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

$$= \int_0^\infty \int_0^1 (\gamma\mu)^{a-1} [\gamma(1-\mu)]^{b-1} \gamma e^{-\gamma} d\gamma d\mu$$

Now, use variable substitution

let $x = \gamma\mu$. $y = \gamma(1-\mu)$

we have:

$x \in (0, \infty)$ $y \in (0, \infty)$ $\gamma = x+y$ $\mu = \frac{x}{x+y}$

Jacobian matrice is

$$\begin{bmatrix} \frac{\partial \gamma}{\partial x} & \frac{\partial \mu}{\partial x} \\ \frac{\partial \gamma}{\partial y} & \frac{\partial \mu}{\partial y} \end{bmatrix} = \begin{bmatrix} 1 & \frac{y}{(x+y)^2} \\ 1 & \frac{-x}{(x+y)^2} \end{bmatrix},$$

The absolute value of determinant is $\frac{1}{x+y}$

So, our transformed integration turns to be

$$\int_0^\infty \int_0^\infty x^{a-1} y^{b-1} (x+y) e^{-(x+y)} \cdot \frac{1}{x+y} \, dx\, dy$$

$$= \int_0^\infty \int_0^\infty x^{a-1} y^{b-1} e^{-x-y} \, dx\, dy$$

$$= \int_0^\infty x^{a-1} e^{-x} \, dx \int_0^\infty y^{b-1} e^{-y} \, dy$$

$$= \Gamma(a)\,\Gamma(b)$$

proof end.

Let's compute the expectation and variance of Beta. R.V.

The result is

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

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

Proof.

We will prove a much more general formula, the moment of μ.

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

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

Turn to page 4

$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \cdot \frac{\Gamma(a+k)\Gamma(b)}{\Gamma(a+b+k)} = \frac{(a+b-1)!\,(a+k-1)!}{(a-1)!\,(a+b+k-1)!}$$

$$= \frac{\text{Beta}(\mu \mid a+k, b)}{\text{Beta}(\mu \mid a, b)}$$

So.

$$\mathbb{E}(\mu) = \frac{(a+b-1)!\,a!}{(a-1)!\,(a+b)!} = \frac{a}{a+b}$$

$$\text{Var}(\mu) = \mathbb{E}(\mu^2) - \mathbb{E}(\mu)^2 = \frac{(a+b-1)!\,(a+1)!}{(a-1)!\,(a+b+1)} - \frac{a^2}{(a+b)^2}$$

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

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

---

a, b in Beta distribution is called hyperparam, because it controls the distribution of μ. We can randomly or manually choose at first. As the data accumulates, the effects of a & b will decrease to $\to 0$.

Now posterior.

$$p(\mu|X) = \frac{1}{\Gamma} \; p(X|\mu) \cdot \text{Beta}(\mu|a,b)$$

Normalization Term

$$= \frac{1}{\Gamma} \binom{N}{m} \mu^{m} (1-\mu)^{N-m} \cdot \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1} (1-\mu)^{b-1}$$

Remain only terms involving μ.

$$\propto \mu^{m+a-1} (1-\mu)^{N-m+b-1}$$

$$\Rightarrow \frac{\Gamma(N+a+b)}{\Gamma(m+a)\Gamma(N-m+b)} \mu^{m+a-1} (1-\mu)^{N-m+b-1}$$

$$= \text{Beta}(\mu|a+m, b+N-m)$$

So, if we have a dataset with m, "$x=1$" cases and $N-m$, "$x=0$", cases we will modify/improve our $\hat{\mu}$ by $\text{Beta}(\mu|a,b) \to \text{Beta}(\mu|a+m, b+N-m)$

Another important property is that no matter what prior we choose (say, a & b), as our experiment data go to ∞, we will always converge to a satisfying result. This is called **Sequential learning**. We can therefore introduce an **online training algorithm** based on **MLE**.

But, if we are not given infinite number of data, all we have is dataset D. the best we can do is MAP rather than ML.

MAP.

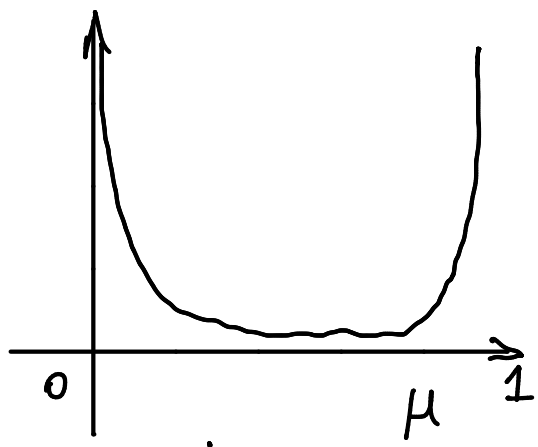
$$p(x=1|D) = \int_0^1 p(x=1|\mu)\, p(\mu|D)\, d\mu = \int_0^1 \mu\, p(\mu|D)\, d\mu = E(\mu|D)$$

if there are m samples of $x=1$, N samples in all.

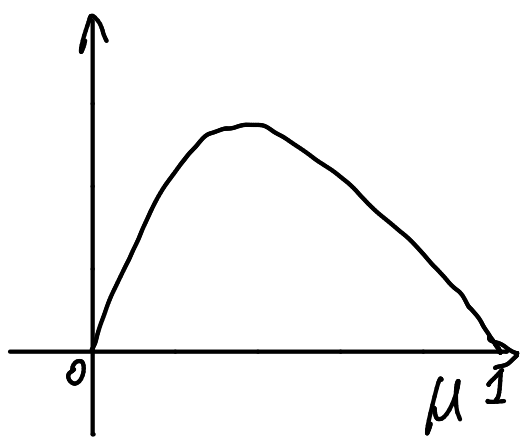
$$p(x=1|D) = \frac{m+a}{N+a+b}$$

Also, if we have infinite number of samples. MAP will have same result as MLE.

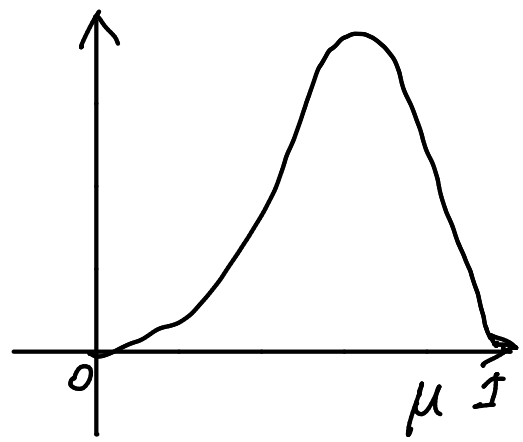
Beta$(\mu|a,b)$



$a=b=0.1$



$a=2$
$b=3$



$a=8$
$b=4$

as experiment number keeps increasing, the distribution will become sharper and narrower.

Idea:

As we get more and more data, uncertainty represented by posterior distribution decrease. steadily.

This is true.

We know that

$$\mathbb{E}_\theta[\theta] = \mathbb{E}_D[\mathbb{E}(\theta|D)]$$

Because

$$\mathbb{E}_D[\mathbb{E}(\theta|D)] = \int_D \int_\theta \theta \cdot P(\theta|D)\, d\theta \cdot P(D)\, d_D$$

$$= \int \theta P(\theta) d\theta = \mathbb{E}_\theta[\theta]$$

and

$$\underbrace{Var_\theta[\theta]}_{\text{Prior's Variance}} = \underbrace{\mathbb{E}_D[Var_\theta[\theta|D]]}_{\text{average of posterior's variance } (\geq 0)} + \underbrace{Var_D[\mathbb{E}_\theta(\theta|D)]}_{\text{Variance of MAP estimator}}$$

So, on average MAP's variance $\leq$ Prior's variance after averaging over D.

(3) Multinomial Variable.

Multinomial Variable is a generalization of Binary Variable. We now have K possible outputs.

E.g. $K=6$. and we see $x_3=1$, then we have

$\underline{x} = (0,0,1,0,0,0)^T$ — one-bit hot code

$$p(\underline{x}|\underline{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k}$$

$$\sum_x p(x|\underline{\mu}) = \sum_{k=1}^{K} \mu_k = 1$$

$$E(x|\underline{\mu}) = \sum_x p(x|\underline{\mu}) \cdot x = (\mu_1, \mu_2 \cdots \mu_K) = \underline{\mu}$$

The likelihood function is also similar to Binary Variable's

$$p(D|\underline{\mu}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mu_k^{x_{nk}} = \prod_{k=1}^{K} \mu_k^{\sum_n x_{nk}} = \prod_{k=1}^{K} \mu_k^{m_k}$$

$m_k = \sum_n x_{nk}$ ⋯ # of state where $x=k$ shows up.

Also, we get start from MLE.

$$\ln p(D|\mu) = \sum_{k=1}^{K} m_k \ln \mu_k$$

Because we should have another constrain $\sum_k \mu_k = 1$, so we use Lagrangian.

Solve: $\max \ln p(D|\mu) + \lambda\left(\sum_k \mu_k - 1\right) = L(\underline{\mu}, \lambda)$

$$\frac{\partial}{\partial \mu_k} L(\underline{\mu}, \lambda) = \frac{m_k}{\mu_k} + \lambda = 0 \Rightarrow \mu_k = -\frac{m_k}{\lambda}$$

So $\sum_k \mu_k = \frac{-\sum m_k}{\lambda} = 1 \Rightarrow \lambda = -\sum_k m_k = -N$

$$\boxed{\mu_k^{ML} = \frac{m_k}{N}}$$

If we similarly consider **multinomial distribution**

$$\text{Multi}(m_1, m_2 \dots m_K \mid \underset{\sim}{\mu}, N) = \binom{N}{m_1, m_2 \dots m_K} \prod_k \mu_k^{m_k}$$

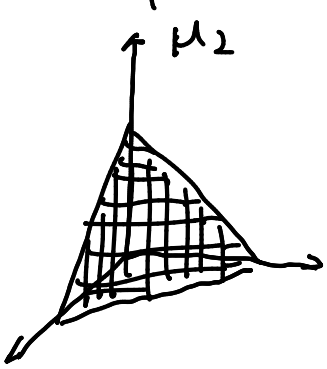
$$\binom{N}{m_1, m_2 \dots m_K} = \frac{N!}{m_1! \dots m_K!}$$

(4) Dirichlet Distribution

Now, we will find the prior of the Multinomial Distribution, which is Dirichlet Distribution, which is the conjugate prior to multinomial.

$$\mathrm{Dir}(\underline{\mu}|\underline{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\cdots\Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$$

if $K=3$, say $\mu_1+\mu_2+\mu_3=1$, $\mu_1, \mu_2, \mu_3 \in [0,1]$ it's distribution is called simplex.



also, the posterior is another Dirichlet dist.

$$p(\underline{\mu}|D,\underline{\alpha}) = \mathrm{Dir}(\underline{\mu}|\underline{\alpha}+\underline{m})$$

$$\underline{m} = [m_1, m_2 \cdots m_K]^T$$