1.6 The conjugate prior for Bernoulli distribution is the Beta distribution :-

1)

$$p(\mu) = \frac{1}{B(\alpha,\beta)} \cdot \mu^{\alpha-1} (1-\mu)^{\beta-1}$$

here $\mu \in [0,1]$ and $B(\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha) \cdot \Gamma(\beta)}$

where Γ is the Gamma function.

The mode of a continuous probability distribution is the value x at which its probability density function has it maximum value.

The mode of $Beta(\alpha,\beta) = \frac{\alpha-1}{\alpha+\beta-2}$.

Now, using Bayes rule we seek value of μ, that maximizes the posterior $Pr(\mu/X)$

$$Pr(\mu/X) = \frac{Pr(X/\mu)\ Pr(\mu)}{Pr(X)}$$

We have $\hat{\mu}_{MAP} = \underset{\mu}{\text{argmax}}\ Pr(\mu/X)$

$$= \underset{\mu}{\text{argmax}}\ \frac{Pr(X/\mu)\ Pr(\mu)}{Pr(X)}$$

$$= \underset{\mu}{\text{argmax}} \; Pr(X|\mu) \; Pr(\mu)$$

$$= \underset{\mu}{\text{argmax}} \prod_{x_i \in X} Pr(x_i|\mu) \; Pr(\mu)$$

like,

For Maximum Likelihood estimation, it is easier to calculate the argmax for the logarithm.

$$\underset{\mu}{\text{argmax}} \; Pr(\mu|X) = \underset{\mu}{\text{argmax}} \; \log\left(Pr(\mu|X)\right)$$

$$= \underset{\mu}{\text{argmax}} \log \prod_{x_i \in X} Pr(x_i|\mu) \cdot Pr(\mu)$$

$$= \underset{\mu}{\text{argmax}} \sum_{x_i \in X} \{\log Pr(x_i|\mu)\} + \log Pr(\mu)$$

So now $Pr(\mu|X) \propto Pr(x_i|\mu) \cdot Pr(\mu)$

↑ Posterior ↑ Likelihood (Bernoulli) ↑ Prior (Beta)

Now we thus have

$$\rightarrow Pr(x_i/M) = \text{Bernoulli}(x_i/M) = M^{x_i}(1-M)^{1-x_i}$$

$$\rightarrow Pr(M) = \text{Beta}(M/\alpha,\beta) = \frac{1}{B(\alpha,\beta)} \cdot M^{\alpha-1}(1-M)^{\beta-1}$$

thus, $Pr(M/X) \propto Pr(X/M) \cdot Pr(M)$

is equivalent to.

$$Pr(M/X) \propto \left\{ \prod_i \text{Bernoulli}(x_i/M) \right\} \cdot \text{Beta}(M/\alpha,\beta)$$

Furthermore:-

$$L = \log Pr(M/X)$$

$$= \log \left\{ \prod_i \text{Bernoulli}(x_i/M) \right\} \cdot \text{Beta}(M/\alpha,\beta)$$

$$= \sum_i \log \text{Bernoulli}(x_i/M) + \log \text{Beta}(M/\alpha,\beta)$$

Noting that this is almost the same as the ML estimate except that we now have an additional term resulting from prior.

We now find the maximum value of μ by setting first derivative of L equal to zero and solving for μ.

$$\frac{\partial}{\partial \mu} L = \sum_i \frac{\partial}{\partial \mu} \log \text{Bernoulli}(x_i | \mu) + \frac{\partial}{\partial \mu} \log \text{Beta}(\mu | \alpha, \beta)$$

The first term is the same for mL^2 i.e.

$$\sum_i \frac{\partial}{\partial \mu} \log \text{Bernoulli}(x_i | \mu) = \frac{1}{\mu} \sum_{i=1}^{n} x_i - \frac{1}{1-\mu} \sum_{i=1}^{n} (1 - x_i)$$

To find the second term:-

$$\frac{\partial}{\partial \mu} \log \text{Beta}(\mu | \alpha, \beta)$$

$$= \frac{\partial}{\partial \mu} \log \left\{ \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha) \cdot \Gamma(\beta)} \cdot \mu^{\alpha-1} (1-\mu)^{\beta-1} \right\}$$

$$= \frac{\partial}{\partial \mu} \log \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha) \cdot \Gamma(\beta)} + \frac{\partial}{\partial \mu} \log \mu^{\alpha-1} (1-\mu)^{\beta-1}$$

$$= 0 + \frac{\partial}{\partial \mu} \log \mu^{\alpha-1} (1-\mu)^{\beta-1}$$

$$= \frac{\partial}{\partial M}(\alpha-1)\frac{\partial}{\partial M}\log(\mu) + \beta-1\frac{\partial}{\partial \mu}(1-\mu)$$

$$= \frac{\alpha-1}{M} - \frac{\beta-1}{1-M}$$

To find $\hat{M}_{MAP}$, we now set $\frac{\partial}{\partial \mu} L = 0$ and

Solve for μ.

$$0 = \frac{\partial L}{\partial M}$$

$$= \frac{1}{\mu}\sum_{i=1}^{n} x_i - \frac{1}{1-\mu}\sum_{i=1}^{n}(1-x_i) + \frac{\alpha-1}{M} - \frac{\beta-1}{1-M}$$

thus:-

$$M\left[\sum_{i=1}^{n}(1-x_i) + \beta - 1\right] = (1-M)\left[\sum_{i} x_i + \alpha - 1\right]$$

$$M\left[\sum_{i=1}^{n}(1-x_i) + \sum_{i} x_i + \beta - 1 + \alpha - 1\right] = \sum_{i} x_i + \alpha - 1$$

$$M\left[\sum_{i=1}^{n} 1 + \beta + \alpha - 2\right] = \sum_{i} x_i + \alpha - 1$$

Finally if we let our Bernoulli distribution be coded as ~~Bebol~~ Positive = 1 & Negative = 0, we have

$\sum_i x_i = n_r$ where n_r denotes number of Positive instances.

Then, $\mu\left[\sum_{i=1}^{n} 1 + \beta + \alpha - 2\right] = \sum_i x_i + \alpha - 1$

$$\mu\left[n + \beta + \alpha - 2\right] = n_R + \alpha - 1$$

Finally

$$\hat{\mu}_{MAP} = \frac{n_R + \alpha - 1}{n + \beta + \alpha - 2}$$

(1.6)(3)

Difference

For maximum likelihood

$$\hat{\mu}_{ML} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Thus the maximum likelihood is just the portion of flips that came out heads.

1.6)
2)

Dirichlet Distribution:-

There are probability distributions over multinomial parameter vectors.

Parameterized by a vector $\alpha = (\alpha_1, \cdots, \alpha_m)$ where $\alpha_j > 0$ that determines the shape of the distribution.

$$DIR(\theta|\alpha) = \frac{1}{C(\alpha)} \prod_{j=1}^{m} \theta_j^{\alpha_j - 1}$$

$$C(\alpha) = \int_{\Delta} \prod_{j=1}^{m} \theta_j^{\alpha_j - 1} \, d\theta = \frac{\prod_{j=1}^{m} \Gamma(\alpha_j)}{\Gamma\left(\sum_{j=1}^{m} \alpha_j\right)}.$$

Γ is a generalization of the factorial function

$\Gamma(k) = (k-1)!$ for positive integer k

$\Gamma(x) = (x-1)\,\Gamma(x-1)$ for all x.

→ Data $X = (X_1, \cdots, X_n)$ generated i.i.d from $DISCRETE(\theta)$.

→ Prior is $DIR(\alpha)$. By Bayes Rule, posterior is:

$$P(\theta|X) \propto P(X|\theta)\,P(\theta)$$

$$\propto \left(\prod_{j=1}^{m} \theta_j^{N_j}\right)\left(\prod_{j=1}^{m} \theta_j^{\alpha_j - 1}\right)$$

$$= \prod_{j=1}^{m} \theta_j^{N_j + \alpha_j - 1}, \text{ so}$$

$$P(\theta|X) = DIR(N+\alpha)$$

⇒ So if prior is Dirichlet with parameters α, posterior is Dirichlet with parameters $N+\alpha$.

⇒ can regard Dirichlet parameter α as "pseudo-counts" from "pseudo-data"

1.6 (3) Difference?

Prior → DIR(α)

Likelihood → DISCRETE(θ)

Posterior → DIR($N+\alpha$)

→ A class C of prior distributions $P(H)$ is conjugate to a class of likelihood functions $P(D|H)$ iff the posterior $P(H|D)$ is also a member of C.

#.

In general, conjugate priors encode "pseudo observations.

→ The difference between prior $P(H)$ and posterior $P(H/D)$ are observations in D.

→ But $P(H/D)$ belongs to same family as $P(H)$ and can serve as prior for inferences about more data D'.

⇒ must be possible to encode observations D using parameters of prior.

A true Bayesian prefers to use the full $P(H/D)$ but sometimes we have to choose a best hypothesis.

The Maximum a posteriori (MAP) or posterior mode is

$$\hat{H} = \underset{H}{\operatorname{argmax}}\ P(H/D) = \underset{H}{\operatorname{argmax}}\ P(D/H)\,P(H)$$

The expected value $E_p[X]$ of X under distribution p is:

$$E_p[X] = \int x\ P(X = x)\, dx.$$

The expected value is a kind of average weighted by $P(X)$. The expected value $E[\theta]$ of θ is an estimate of θ.

The MAP is

$$\hat{H} = \underset{H}{\operatorname{argmax}}\, P(H|D) = \underset{H}{\operatorname{argmax}}\, P(D|H)\, P(H)$$

For Dirichlets with parameters α the MAP estimate is

$$\theta_j = \frac{\alpha_j - 1}{\sum_{j'=1}^{m} (\alpha_{j'} - 1)}$$

So if the posterior is $DIR(N + \alpha)$, the MAP estimate for θ is

$$\hat{\theta}_j = \frac{N_j + \alpha_j - 1}{n + \sum_{j'=1}^{m} (\alpha_{j'} - 1)}$$

(1.6) (3) → Difference

If $\alpha = 1$ then $\hat{\theta}_j = N_j / n$ which is also the maximum likelihood estimate (MLE) for θ.

2.4

Sigmoid function:- $\sigma(a) = \frac{1}{1+e^{-a}}$; $a = w^T x$

1) ~~d~~ $\therefore \sigma(w^T x) = \frac{1}{1+e^{-w^T x}}$

$$\frac{d\sigma(w^T x)}{dx} = \left(\frac{1}{1+e^{-w^T x}}\right)^2 \frac{d}{dx}\left(1+e^{-w^T x}\right)$$

$$= \left(\frac{1}{1+e^{-w^T x}}\right)^2 e^{-w^T x} \cdot (-w^T)$$

$$= \left(\frac{1}{1+e^{-w^T x}}\right)\left(\frac{1}{1+e^{-w^T x}}\right)\left(-e^{-w^T x}\right)(w^T)$$

$$= \left(\frac{1}{1+e^{-w^T x}}\right)\left(\frac{-e^{-w^T x}}{1+e^{-w^T x}}\right)(w^T)$$

$$= \sigma(w^T x)\left(1-\sigma(w^T x)\right) \cdot (w^T)$$

$$= w^T \cdot \sigma(w^T x)\left(1-\sigma(w^T x)\right).$$

2) $$p(y=1 \mid x, w) = \sigma(w^T x) = \frac{1}{1+e^{-w^T x}} \quad \text{-Eq 1.}$$

By Bayes rule we get.

$$p(y=1 \mid x) = \frac{p(x \mid y=1)\, p(y=1)}{p(x)}$$

$$= \frac{\alpha\, p(x \mid y=1)}{\alpha\, p(x \mid y=1) + (1-\alpha)\, p(x \mid y=0)}$$

... By Product Rule.

$$= \frac{1}{1 + \frac{1-\alpha}{\alpha} \frac{p(x \mid y=0)}{p(x \mid y=1)}}$$

Given $\{y=i\}$ for $i=[0,1]$, the sequence of random variables $(x_i)_i$ are independent, we know that.

$$p(x \mid y=i) = \prod_{j=1}^{n} p(x_j \mid y=i)$$

$p(x_j \mid y=i)$ uses a Gaussian density w.r.t mean and variance.

Hence,

$$P(x_j | y=i) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(-\frac{1}{2\sigma_j^2}(x_j - \mu_{j,i})^2\right)$$

Also we have

$$P(y=-1|x) = 1 - \sigma(w^T x)$$

$$= 1 - \frac{1}{1+e^{-w^T x}}$$

$$= \frac{e^{-w^T x}}{1+e^{-w^T x}} \quad \text{— Eq 2}$$

∴ To prove $P(y=\pm 1 | x, w) = \sigma(w^T x)$

$$= \frac{1}{1+e^{-y w^T x}}$$

put $P(y=1|x,w) = \frac{1}{1+e^{-w^T x}}$ ... given

put $P(y=-1|x,w) = \frac{1}{1+e^{w^T x}}$

Now multiplying and dividing by $e^{-w^T x}$

$$= \frac{1}{1+e^{w^T x}} \times \frac{e^{-w^T x}}{e^{-w^T x}}$$

$$= \frac{e^{-w^T x}}{e^{-w^T x} + e^{w^T x} \cdot e^{-w^T x}}$$

$$\Rightarrow 1$$

$$= \frac{e^{-w^T x}}{1+e^{-w^T x}} \quad \text{Same as eq 2.}$$

Hence we can express posterior for both classes as

$$P(y=\pm 1 \mid w, x) = \sigma(w^T x) = \frac{1}{1+e^{-y w^T x}}$$

3) Extra credit

Log-likelihood (X, Y, w) of $P(y = \pm 1 \mid x, w)$

$$= \sum_i \log \frac{1}{1 + e^{-y_i w^T x_i}}$$

$$= \sum_i -\log\left(1 + e^{-y^{(i)} w^T x^i}\right)$$

Therefore maximizing LL is equivalent to minimizing $\sum_i^N \log\left(1 + e^{-y^{(i)} w^T x^i}\right)$

... This is the logistic loss.