3.2) with rule of chain in probability

$$p(x_{1:N}) = p(x_1)\,p(x_2/x_1)\,p(x_3/x_{1:2}) \cdots$$

After seeing dataset D = H, T, T, H, H and if we have prior probabilities as $P(T) = \alpha_0$, $P(H) = \alpha_1$ and $\alpha = \alpha_0 + \alpha_1$. Then we can ~~update~~ derive likelihood by

$$P(D) = P(H)\,P(T/H)\,P(T/_{H,T})\,P(H/_{H,T,T})\,P(H/_{H,T,T,H})$$

$$= \frac{\alpha_1}{\alpha} \cdot \frac{\alpha_0}{\alpha+1} \cdot \frac{\alpha_0+1}{\alpha+2} \cdot \frac{\alpha_1+1}{\alpha+3} \cdot \frac{\alpha_1+2}{\alpha+4}.$$

$$= \frac{[(\alpha_0)(\alpha_0+1)]\,[(\alpha_1)(\alpha_1+1)(\alpha_1+2)]}{[\alpha\cdot(\alpha+1)\cdot(\alpha+2)(\alpha+3)(\alpha+4)]}$$

To generalize

$$= \frac{[(\alpha_0)\,(\alpha_0+1) \cdots (\alpha_0+N_0-1)]\,[(\alpha_1)\,(\alpha_1+1)\cdots(\alpha_1+N_1-1)]}{[(\alpha)\,(\alpha+1) \cdots (\alpha+N-1)]}$$

We can rewrite ~~this~~ as $(\because \Gamma(\alpha) = (\alpha-1)!)$

$$= \frac{\Gamma(\alpha_0+N_0)}{\Gamma(\alpha_0)} \cdot \frac{\Gamma(\alpha_1+N_1)}{\Gamma(\alpha_1)} \Bigg/ \frac{\Gamma(\alpha+N)}{\Gamma(\alpha)}$$

$$= \frac{\Gamma(\alpha_0+N_0)\,\Gamma(\alpha_1+N_1)}{\Gamma(\alpha_1+\alpha_0+N)} \cdot \frac{\Gamma(\alpha_1+\alpha_0)}{\Gamma(\alpha_0)\,\Gamma(\alpha_1)}$$

//

3.3) Beta-binomial posterior pledictive

~~$P(x/n, D) = \int Bin_\theta(x/\theta, n) \cdot Beta(\theta/a_D, b_D) \cdot d\theta$~~

~~$= {}^nC_x \frac{1}{B(a_D, b_D)} \int \theta$~~

$$P_n(x/D) = \int Bin_{n,\theta}(x) \cdot Beta_{a_D, b_D}(\theta) \cdot d\theta$$

$$= {}^nC_x \cdot \frac{1}{B(a_D, b_D)} \int \theta^x \cdot (1-\theta)^{n-x} \cdot \theta^{a_D - 1} \cdot (1-\theta)^{b_D - 1} \cdot d\theta$$

$$= {}^nC_x \cdot \frac{1}{B(a_D \cdot b_D)} \cdot B(x + a_D, n - x + b_D).$$

when $n = 1$ then

$$P(x=1/D) = \frac{B(a_D + 1, b_D)}{B(a_D, b_D)} = \frac{\Gamma(a_D+1)\Gamma(b_D)}{\Gamma(a_D+1+b_D)} \cdot \frac{\Gamma(a_D+b_D)}{\Gamma(a_D)\Gamma(b_D)}$$

$$= \frac{a_D}{a_D + b_D} \quad \left(\because \Gamma(a) = (a-1)\Gamma(a-1)\right)$$

3.4) $X \sim$ Number of Heads $\sim Bin(x), n=5$

$$P(X<3/\theta) = P(X=0/\theta) + P(X=1/\theta) + P(X=2/\theta)$$

$$= {}^5C_0 \cdot (1-\theta)^5 + {}^5C_1 \cdot \theta(1-\theta)^4 + {}^5C_2 \cdot \theta^2 (1-\theta)^3.$$

Posterior: $P(\theta/X<3) \propto p(X<3/\theta)\, P(\theta).$

$$P(\theta) = Beta_{(1,1)}(\theta) = \frac{\theta^{1-1} \cdot (1-\theta)^{1-1}}{B(1,1)} = 1 = U(0,1)$$

As prior is uniform, then

$$P(\theta/X<3) \propto P(X<3/\theta)$$

$$\propto {}^5C_0 \cdot (1-\theta)^5 + {}^5C_1 \theta(1-\theta)^4 + {}^5C_2 \cdot \theta^2 (1-\theta)^3$$

3.7) a) Non-Bayesian model is

$$Pois_{\lambda}(x) = e^{-\lambda} \cdot \frac{\lambda^{x}}{x!} \qquad \theta = (\lambda)$$

Likelihood: $P(D/\theta) = \prod_{i=1}^{N} \cdot e^{-\theta} \cdot \frac{\theta^{x_i}}{x_i!}$

prior: $p(\theta) = Ga_{(a,b)}(\theta) = \frac{b^{a}}{\Gamma(a)} \cdot \theta^{a-1} \cdot e^{-\theta b}.$

a, b – hyper params

posterior $p(\theta/D) \propto P(D/\theta)\, p(\theta)$

$$\propto \prod_{i=1}^{N} \cdot e^{-\theta} \cdot \frac{\theta^{x_i}}{x_i!} \cdot (\theta^{a-1} \cdot e^{-\theta b})$$

$$\propto \frac{e^{-N\theta} \cdot \theta^{\sum_i x_i}}{\prod_i x_i!} \cdot \theta^{a-1} \cdot e^{-\theta b}$$

$$\propto \theta^{\sum_i x_i + a - 1} \cdot e^{-\theta(N+b)}$$

posterior is also in the form of prior

$Ga(\theta) \propto \theta^{a-1} \cdot e^{-\theta b}$ with paremeters.

$P(\theta/D) = Ga(a_D, b_D) =$ $a_D = \sum_i x_i + a$, $b_D = N+b.$

b) $E(p(\theta/D)) = E[Ga_{(a_D,b_D)}(\theta)]$

$$= \frac{a_D}{b_D} = \frac{\sum x_i + a}{N+b}$$

when a → 0, b → 0 this becomes

$$= \frac{\sum x_i}{N}$$

5.10) non-Baysian model

$$p_\theta(x) = U(0,\theta)$$

prior, $p(\theta) = K m^K \cdot \theta^{-(K+1)} \; \mathbb{1}_{\{\theta \geq m\}}$

hyper-params (K, m)

with non-informative prior $(K, m) = (0, 0)$

$$p(\theta) \propto 1/\theta$$

In first problem, we're given $\mathcal{D} = \{100\}$, so

likelihood: $p(\mathcal{D}/\theta) = \prod_{i=1}^{N} 1/\theta \; \mathbb{1}_{\{x_i \in [0,\theta]\}}$

posterior: $p(\theta/\mathcal{D}) \propto p(\mathcal{D}/\theta)\, p(\theta)$

$$\propto \prod_{i=1}^{N} \frac{1}{\theta} \cdot \mathbb{1}_{\{x_i \in [0,\theta]\}} \cdot K . m^K . \theta^{-(K+1)} . \mathbb{1}_{\{\theta \geq m\}}$$

$$\propto \frac{K . m^K}{\theta^{N+K+1}} \underbrace{\left(\prod_{i=1}^{N} \mathbb{1}_{\{x_i \in [0,\theta]\}}\right) \mathbb{1}_{\{\theta \geq m\}}}$$

$$\propto \frac{K . m^K}{\theta^{N+K+1}} . \mathbb{1}_{\{\theta \geq max(\mathcal{D})\}}$$

if $max(\mathcal{D}) \leq m$:

$$\propto \frac{(N+K)\, m^{N+K}}{\theta^{N+K+1}} . \mathbb{1}_{\{\theta \geq m\}} = paleto(\theta)_{(N+K, m)}$$

if $max(\mathcal{D}) \geq m$:

$$\propto \frac{(N+K)\, max(\mathcal{D})^{N+K}}{\theta^{N+K+1}} . \mathbb{1}_{\{\theta \geq max(\mathcal{D})\}} = paleto(\theta)_{(N+K,\, max(\mathcal{D}))}$$

Scanned by CamScanner

a) In this problem $\mathcal{D} = \{100\}$

$$\max(\mathcal{D}) = 100 \;, \; N = 1$$

posterior prior $p(\theta) = Pa(0,0)$ $k=0, m=0$

posterior $p(\theta/\mathcal{D}) =$ as $\max(\mathcal{D}) \geq m$

$$\propto \frac{(N+K).\max(\mathcal{D})^{N+K}}{\theta^{N+K+1}} \cdot \mathbb{1}_{\{\theta \geq \max(\mathcal{D})\}}$$

$$\propto \frac{100}{\theta^2} \cdot \mathbb{1}_{\{\theta \geq 100\}}$$

$$\propto \text{pareto}(\theta/1, 100).$$

b) Mean of $\text{pareto}(\theta/k,m)$ if $k>1$:

$$\frac{km}{k-1}$$

But here our $k_D = 1$, so mean doesn't exi

mode of $\text{pareto}(\theta/k,m) = m = 100$

Median of $\text{pareto}(\theta/k,m) = m \sqrt[k]{2}$

$$= 100 . \sqrt[1]{2} = 200$$

c) posterior predictive

$$p(x') \propto \int p(x'/\theta) \cdot p(\theta/\mathcal{D})\, p(\theta)$$

$$\propto \int p\cancel{(x'/\theta) . \text{Pareto}(\theta/k}_{D}, m_D$$

As we have two types of posterior for $p(\theta/\mathcal{D})$

we get $p(\mathcal{D}') = \begin{cases} \dfrac{K_D}{(N'+K_D)\, m_D^{N'}} & \text{if } \max(\mathcal{D}') \le m_D \\ \dfrac{K_D \cdot m_D^{K_D}}{(N'+K_D)\, \max(\mathcal{D}')^{N'+K_D}} & \text{if } \max(\mathcal{D}') > m_D \end{cases}$

we are asked to predikt for

$K_D = 1$, $m_D = 100$
$N' = 1$, $\max(\mathcal{D}') = x$

$$p(x) = \frac{1}{2m} \mathbb{1}_{\{x \le 100\}}$$

$$+ \frac{100}{2x^2} \cdot \mathbb{1}_{\{x > 100\}}$$

d) $p(x=100) = \frac{1}{200} \cdot \mathbb{1}_{\{100 \le 200\}} + \frac{100}{2\times 10^4} \cdot \mathbb{1}_{\{100 > 100\}}$

$= \frac{1}{200}$

$p(x=50)$

$= \frac{1}{200} \cdot \mathbb{1}_{\{50 \le 100\}} + 0 = \frac{1}{200}$

$p(x=150) = \frac{100}{2\times(150)^2} = \frac{1}{445}$

e) i) As our data is discrete, we can choose discrete probability distribution rather than (uniform) continous distribution

2) we took an un-informative prior, ~~but~~ which, didn't effect much on posterior. So, informative prior may help to get better results.

3.11) Bayesian Analysis of Exp

$p_\theta(x) = \theta e^{-\theta x}$ for $\mathcal{X} = \{x \geq 0\}$ $\forall x_i$

$\theta > 0$

a) MLE: log-liklihood

$$l(\theta) = \sum_{i=1}^{N} \log\theta \cdot e^{-\theta x_i}$$

$$= \sum_{i=1}^{N} (\log\theta - \theta x_i) = N\log\theta - \theta\sum_{i=1}^{N}$$

Set, $\frac{\partial l}{\partial \theta} = 0$ to get the $\arg\max_\theta l(\theta)$

$$\hat{\theta}_{MLE} = \arg\max_\theta l(\theta)$$

$$\Rightarrow \frac{\partial}{\partial\theta}\left(N\log\theta - \theta\sum_{i=1}^{N} x_i\right) = 0$$

$\frac{N}{\theta} - \sum_{i=1}^{N} x_i = 0 \Rightarrow$ ~~$\hat{\theta}_{MLE} = 1$~~

$$\hat{\theta}_{MLE} = \frac{N}{\sum_{i=1}^{N} x_i}$$

b) After observing dataset $\mathcal{D} = \{5, 6, 4\}$

$$\hat{\theta}_{MLE} = \frac{3}{5+6+4} = \frac{3}{15} = 1/5$$

c) we assume prior $p(\theta)$ also follows exponential

$$P(\theta) = Exp_\lambda(\theta) = \lambda \cdot e^{-\lambda\theta}.$$

and with our knowledge $E[\theta] = 1/3$.

we already know, $E_\lambda[\theta] = 1/\lambda = 1/3 \Rightarrow \lambda = 3$

d) posterior $p(\theta/D) \propto p(D/\theta)\,p(\theta)$

$$\propto \prod_{i=1}^{N} (\theta . e^{-\theta x_i}) . \lambda . e^{-\lambda\theta}$$

$$\propto \theta^{N} . e^{-\theta(\sum_{i=1}^{N} x_i + \lambda)}$$

So, posterior has form of Gamma with params $(N+1, \sum_{i=1}^{N} x_i + \lambda)$ $\left(\because Ga_{(a,b)}(\theta) \propto \theta^{a-1} . e^{-\theta b}\right)$

e) If we write exponential has Gamma distribution then both liklihood and prior look similar

$\left(\because Exp_{\lambda}(\theta) = Ga_{(1,\lambda)}(\theta)\right)$.

$$p(D/\theta) \cancel{\propto} = \prod_{i=1}^{N} \theta . e^{-\theta x_i} = \theta^{N} . e^{-\theta \sum_{i=1}^{N} x_i}$$

$$p(\theta) = \lambda . e^{-\lambda\theta} \propto \theta^{1-1} . e^{-\lambda\theta}$$

Both are in Gamma distribution form, so ~~[illegible]~~

prior is conjugate to liklihood.

f) $E[\theta/D]$ of Gamma distribution with params $(N+1, \sum_{i=1}^{N} x_i + \lambda)$ is simply $\left(\because E[Ga_{(a,b)}(\theta)] = a/b\right)$

$$= \frac{N+1}{\sum_{i=1}^{N} x_i + \lambda}$$

g) $$\hat{\theta}_{MLE} = \frac{1}{\cancel{\sum_{i=1}}} = \frac{1}{\sum_{i=1}^{N} x_i / N}$$

$$E[\theta/D] = \frac{1}{\frac{\sum_{i=1}^{N} x_i + \lambda}{N+1}}$$

In denominator of posterior mean, its weighted combination of prior mean and ~~liklihood~~ ~~mean~~ MLE.

As our dataset is small and prior is informative, we have to use posterior mean for θ. in this example.

3.15) prior $p(\theta) = \text{Beta}(a, b)$.

and we are given $E[\theta] = m$ and $\text{Var}(\theta) = v$

As we already know expectation & variance for beta as

$$E[\theta] = \frac{a}{a+b} = m$$

$$\text{Var}(\theta) = \frac{ab}{(a+b)^2(a+b+1)} = v.$$

We can solve a, b using these two equations. by setting $m = 0.7$ and $v = 0.2^2 = 0.04$

$$\frac{a}{a+b} = 0.7 \Rightarrow a = 0.7a + 0.7b$$

$$a = \tfrac{7}{3} b.$$

$$\frac{a.b}{(a+b)^2(a+b+1)} = 0.04 \Rightarrow \frac{\tfrac{7}{3} b \cdot b}{(\tfrac{7}{3}b+b)^2(\tfrac{7}{3}b+b+1)} = 0.04$$

$$\Rightarrow \frac{\tfrac{7}{3}\cancel{b^2}}{\tfrac{100}{9}\cdot\cancel{b^2}\cdot(\tfrac{7}{3}b+b+1)} = 0.04$$

$$21 = 4(\tfrac{7}{3}b+b+1) \Rightarrow \frac{10b+3}{3}\cdot 4 = 21$$

$$10b+3 = 63/4$$

$$10b = 63/4 - 3 = 51/4 = 12.75$$

$$b = 1.275$$

So $a = 7/3 \cdot 1.275 = 2.975$

4.14) $p(x) = N_{\mu}(\mu, \sigma^2)$ σ^2 - is given

and prior for μ is assumed as

$$p(\mu) = N(m, s^2) \quad \text{hyper-params } (m, s^2).$$

From the derivation in class, posterior of μ is gaussian in form

$$p(\mu/D) \propto e^{-\left(\mu - \frac{\frac{\sum x_i}{2\sigma^2} + \frac{m}{2s^2}}{n/2\sigma^2 + 1/2s^2}\right) \Big/ \frac{1}{n/2\sigma^2 + 1/2s^2}}$$

a) The MAP estimate for this is

$$\mu_{MAP} = \frac{\frac{\sum x_i}{2\sigma^2} + \frac{m}{2s^2}}{n/2\sigma^2 + 1/2s^2} = \left(\frac{n\bar{x}}{\sigma^2} + \frac{m}{s^2}\right)\left(n/\sigma^2 + 1/s^2\right)^{-1}$$

So, μ_{MAP} is weighted average of sample average $(\bar{x})$ and prior mean (m)

b) when n increases then m/s^2 in numerator and $1/s^2$ in denominator values diminishes

$$\text{so } \mu_{MAP} \triangleq \frac{\sum x_i}{2\sigma^2} \Big/ n/2\sigma^2 = \frac{\sum x_i}{n} = \mu_{MLE}$$

c) When prior variance (s^2) increases then $1/s^2 \to 0$

So, the expression becomes again

$$\mu_{MAP} \triangleq \frac{\sum x_i/2\sigma^2}{n/2\sigma^2} = \frac{\sum x_i}{n} = \mu_{MLE}$$

d) when prior variance decreases then $1/s^2$ becomes bigger. So $\mu_{MAP} = m/2s^2 \big/ 1/2s^2 = m$

* Discriminative model based on exponential family with sufficient statistics $\psi(\cdot)$ and parameters for $\psi_i(y)$ is $w_i^T \phi(x)$

$$p(y/x) = e^{\sum_i w_i^T \varphi(x) \cdot \psi_i(y) - A(\sum_i w_i^T \varphi(x))}$$

$$= e^{\sum_i (\sum_j w_{ij}^T \varphi_j(x)) \cdot \psi_i(y) - A(\sum_i w_i^T \varphi(x))}$$

Lets call ~~$\sum$~~ $w^T \varphi(x) = W$, so

$$A(W) = \log\left(\int e^{\sum_i (\sum_j w_{ij}^T \varphi_j(x)) \psi_i(y)} dy\right)$$

In Generative model, we have sufficient Statistics as $\varphi(x) \otimes \psi(y)$, so

$$p(y/x) = p(x,y)/p(x)$$

$$\propto p(x,y)$$

$$\propto e^{W^T(\varphi(x) \otimes \psi(y)) - A(V)}$$

$$\propto e^{\sum_{i,j} v_{ij} \varphi_j(x) \psi_i(y) - A(V)}$$

$$\propto e^{\sum_i (\sum_j v_{ij} \varphi_j(x)) \psi_i(y) - A(V)}$$

Here $A(V) = \log\left(\left(\int e^{\sum_i (\sum_j v_{ij} \varphi_j(x)) \psi_i(y)} dy\right)\right)$

As we can see both of them are in same form with only parameters differ. So, we can say even in ~~generati~~ discriminative model $\varphi(\cdot)$ are

fficient statistics for inputs though they are
cting as ~~a~~ variables in Canonical parameters of
tputs.