2.13 Mutual Information $I(X_1, X_2)$

(3 point)

$$I(X_1, X_2) = H(X_1) + H(X_2) - H(X_1, X_2)$$

where H is entropy of a random variable.

Hence as per given values for entropies in book

$$= \frac{1}{2}\log_2[2\pi e \sigma^2] + \frac{1}{2}\log_2[2\pi e \sigma^2] - \frac{1}{2}\log[(2\pi e)^2(\sigma^4 - \sigma^4\rho^2)]$$

$$= \log_2[2\pi e \sigma^2] - \frac{1}{2}\log[(2\pi e)^2\sigma^4(1-\rho^2)]$$

$$= \log_2 \frac{2\pi e \sigma^2}{[(2\pi e)^2\sigma^4(1-\rho^2)]^{1/2}}$$

(take cofficient of second log back into power and using $\log\frac{a}{b} = \log a - \log b$

$$= \log_2 \frac{\cancel{2\pi e \sigma^2}}{\cancel{2\pi e \sigma^2}(1-\rho^2)^{1/2}}$$

$$= \log \frac{1}{(1-\rho^2)^{1/2}}$$

If $\rho = +1$ or -1 ($X_1 X_2$ highely correlated) then $I(X_1, X_2) = \infty$ justifies this

If $\rho = 0$ i.e $X_1 X_2$ are statistically not correlated at all then $I(X_1, X_2) = \log 1 = 0$ justifies this.

3·6 MLE For the Poisson distribution

(Points 4) Poisson pmf $Poi(x|\lambda) = e^{-\lambda} \frac{\lambda^x}{x!}$ [Note the support in 0, 1, 2 ⋯ all natural number]

Hence let $D = \{(x_i)\}_{i=1}^N$ be our I.I.D observations from this distribution

then log liklihood of D is

$$= \log P(D) = \log \prod_{i=1}^N Poi(x_i|\lambda) = \sum_{i=1}^N \log Poi(x_i|\lambda)$$

$$= \sum_{i=1}^N \log e^{-\lambda} \frac{\lambda^{x_i}}{x_i!}$$

$$= \sum_{i=1}^N (-\lambda) + \sum_{i=1}^N x_i \log\lambda - \sum_{i=1}^N \log(x_i!)$$

$$= -N\lambda + \sum_{i=1}^N x_i \log\lambda - \sum_{i=1}^N \log(x_i!)$$

Taking derivative with respect to λ and equating to zero yield

$$-N + \frac{\sum_{i=1}^N x_i}{\lambda} = 0 \Rightarrow \lambda = \frac{\sum_{i=1}^N x_i}{N}$$

3·7 (a) (2 points each (a, b, ⋯))

$$P(\lambda|D) \propto P(D|\lambda) P(\lambda)$$

$$\propto \left(\prod_{i=1}^N Poi(x_i|\lambda)\right) Ga(\lambda|a,b)$$ (using Given conjugate prior $Ga(\lambda|a,b)$ for λ)

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

$$\propto \frac{\lambda^{\sum_{i=1}^N x_i + a - 1}}{\prod_{i=1}^N x_i!} e^{-N\lambda - b\lambda}$$

$$= Ga\left(\lambda \,\Big|\, a + \sum_{i=1}^N x_i,\ b + N\right)$$

(b) Posterior mean is $E[\lambda|D] = \frac{a + \sum_i^N x_i}{b + N}$

as $a \to 0$, $b \to 0$, it tends to MLE estimate $\frac{\sum_i^N x_i}{N}$ (without prior on λ)

3.11 (only for ENCE 4630) (1 point each) For log likelihood proceed as in previous problem ②

(a) it should be

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

For MLE estimate of θ, take derivative and equate to zero

$$\frac{\partial L(\theta)}{\partial\theta} = \frac{N}{\theta} - \sum_i x_i \Rightarrow \hat{\theta} = \frac{N}{\sum_i x_i}$$

(b) $\hat{\Theta}_{MLE}(D) = \hat{\theta}_{MLE}(\{x_1=5, x_2=6, x_3=4\})$ $(\because D = \{5, 6, 4\})$

$$= \frac{3}{5+6+4} = \frac{3}{15} = \frac{1}{5}$$

(c) $\mathbb{E}(\theta) = 1/\lambda = \frac{1}{3} \Rightarrow \hat{\lambda} = 3$ $\left(\because \underbrace{\begin{aligned} P(\theta) &= \lambda e^{-\theta\lambda} = \mathrm{Expon}(\theta|\lambda) \\ &\propto \theta^{1-1} e^{-\theta\lambda} = \mathrm{Ga}(\theta|1,\lambda) \end{aligned}}_{\text{Hence exponential is a special case of gamma}}\right)$

(d) posterior $P(\theta|D,\hat{\lambda}) \propto P(D|\theta)\,P(\theta|\hat{\lambda})$

$$\propto \underline{\theta e^{-\theta 5}\,\theta e^{-\theta 6}\,\theta e^{-\theta 4}}\;\underline{3e^{-3\theta}}$$

$$\propto \theta^{3} e^{-18\theta} \propto \theta^{4-1} e^{-18\theta} = \mathrm{Ga}(\theta|4,18)$$

$(\because \mathrm{G}(\theta|a,b) \propto \theta^{a-1} e^{-\theta b})$

(e) yes, exponential is a special case of Gamma

(f) posterior $P(\theta|D,\hat{\lambda})$ is $\mathrm{Ga}(\theta|4,18)$

Hence mean $= 4/18$

(g) Posterior mean ($4/18$) is a probabilistic adjustement between prior mean ($1/3$, expert choice) and data driven mean ($MLE = 1/5$). In small smaple size we should take expert advice. Hence posterior mean is reasonable.

3.20 (1 point each)

(a) since feature are not conditionally indepent we need to specify probability for each configuration of bit in $X \in \{0,1\}^D$

Hence we need $2^D - 1$ parameters | class

or $C(2^D - 1)$ Parameter for classes

we would need C different probability distribution $P(X|C)$ or Histogram. [One choice can be binomial $(K; D, \theta)$?
[probability of K success in D trial of coin with parameter θ]

Note: we cannot use multivariate Gaussian as data is binary vector.

(b) Naive Bayes based model will work better as it has less number of parameters. with less sample size it will not overfit and parameters estimation will be more reliable.

(c) with large sample size, use full model as it is more accurate and there is enough data to reliably estimate parameters.

(d) Both take $O(ND)$ time. As estimation of parameters is mostly about counts and we can do this by iterating over N examples and updating counts. (4)

(e) Both model take $O(D)$ time

For NB
$$P(y=c|x,\theta) = \prod_{j=1}^{D} \theta_{jc}^{\mathbb{I}(x_j=1)}$$

we know θ_{jc} after estimation and indexing $x_{test} \in \{0,1\}^D$ ~~can be done in~~ to get $x_{test}[j] = x_j$ can be done in $O(D)$ time.

For Full model we need to pick the parameter corresponding to bit pattern in $x_{test} \in \{0,1\}^D$. This is equivalent to converting x_{test} to an integer and can be done in $O(D)$ time. If we use hash it can be done in $O(1)$ constant time.

(f) we can play the marginalizing trick from probability. we need to compute
$$P(y|x_v,\hat{\theta}) \propto P(x_v|y,\hat{\theta})P(y) \propto \sum_{x_h} P(x_v, x_h|y,\hat{\theta})$$

for Naive Bayes $\sum_{x_h} P(x_v, x_h|y,\hat{\theta}) = \sum_{x_h} P(x_v|y,\hat{\theta})\, P(x_h|y,\hat{\theta})$ conditional independent

$$= P(x_v|y,\hat{\theta}) \left(\sum_{x_h} P(x_h|y,\hat{\theta})\right)$$

If we have pre-computed $\sum_{x_h} (x_h | y, \hat{\theta})$ (5)

then in naive Bayes we need $O(v)$ time

For Full model we need to enumerate over

all 2^h values of $x_h \in \{0,1\}^h$

for marginalising $\sum_{x_h} P(x_v, x_h | y, \hat{\theta})$.

Hence it takes $O(D 2^h) \sim O(2^h)$ if $D << 2^h$

computational time