HOME WORK 1

PROBLEM 1

20/20

1 (a)

$I(X,Y) = H(X) - H(X|Y)$.

We need to prove that

$I(X,Y) = KL\left(p(x,y) \,||\, p(x)\,p(y)\right)$

$$= -\sum_x \sum_y p(x,y) \log \frac{p(x)\,p(y)}{p(x,y)} \quad \text{(from definition)}$$

First, let us write

$$H(X|Y=v) = -\sum_{i=1}^{n} P(X=i|Y=v) \log P(X=i|Y=v)$$

$$H(X|Y) = \sum_v P(y=v)\, H(X|Y=v)$$

$$= -\sum_v \sum_{i=1}^{n} P(y=v)\, P(x=i|Y=v) \log P(x=i|Y=v)$$

$$= -\sum_y \sum_x p(y)\, p(x|y) \log P(x|y)$$

$$= -\sum_y \sum_x p(x,y) \log p(x|y)$$

$$= -\sum_y \sum_x p(x,y) \log \frac{p(x,y)}{p(y)}.$$

Now, we can write

$$H(X) = -\sum_x p(x) \log p(x)$$

$$= -\sum_x \sum_y p(x,y) \log p(x).$$

Therefore,

$$I(X,Y) = H(X) - H(X/Y)$$

$$= -\sum_x \sum_y p(x,y) \log p(x) + \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(y)}.$$

$$= -\sum_x \sum_y p(x,y) \log \frac{p(x)\,p(y)}{p(x,y)}$$

$$= KL\left(p(x,y) \,||\, p(x)p(y)\right)$$

Hence, proved.

(b) The definition is

$$I(X,Y) = -\sum_x \sum_y p(x,y) \log \frac{p(x)p(y)}{p(x,y)}.$$

We have $I(X,Y) = 0$ when

~~$p(x,y) = 0$ i.e. X and Y are disjoint.~~

1> $\log \frac{p(x)\,p(y)}{p(x,y)} = 0 \Rightarrow \frac{p(x)\,p(y)}{p(x,y} = 1$

$\therefore\ p(x,y) = p(x)\,p(y)$

This is the condition for independence of x and y. Thus, $I(X,Y)$ is zero when X and Y are independent. Given Y, there is no new information about X and thus $I(X,Y) = 0$.

2> $p(x,y] = 0$ or X and Y are disjoint event.

?

2) $H(X) = -\int p(x) \ln p(x) \, dx$

We can write $\ln p(x)$ as:

$$\ln p(x) = \ln\left[\frac{1}{\sqrt{2\pi}\sigma} \exp - \frac{(x-\mu)^2}{2\sigma^2}\right]$$

$$= \ln\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \frac{(x-\mu)^2}{2\sigma^2}$$

$$= -\ln(\sqrt{2\pi}\sigma) - \frac{(x-\mu)^2}{2\sigma^2}$$

$$H(X) = -\int_{-\infty}^{\infty} \left[-\ln\sqrt{2\pi}\sigma - \frac{(x-\mu)^2}{2\sigma^2}\right] p(x) \, dx$$

$$= \int_{-\infty}^{\infty} \ln\sqrt{2\pi}\sigma \, p(x) \, dx + \int_{-\infty}^{\infty} \frac{(x-\mu)^2}{2\sigma^2} p(x) \, dx$$

$$= \ln\sqrt{2\pi}\sigma + \int_{-\infty}^{\infty} \frac{(x-\mu)^2}{2\sigma^2} p(x) \, dx \qquad \left(\text{since} \int_{-\infty}^{\infty} p(x) \, dx = 1\right)$$

To compute the second term,

$$\int_{-\infty}^{\infty} \frac{(x-\mu)^2}{2\sigma^2} p(x) \, dx = \int_{-\infty}^{\infty} \frac{(x-\mu)^2}{2\sigma^2} \frac{1}{\sqrt{2\pi}\sigma} \exp - \frac{(x-\mu)^2}{2\sigma^2} \, dx$$

Write $x_1 = \frac{x-\mu}{\sqrt{2}\sigma}$. $dx_1 = \frac{dx}{\sqrt{2}\sigma}$. when $x \to \infty$, $x_1 \to \infty$
" $x \to -\infty$, $x_1 \to -\infty$

Then we can write this term as:

$$\int_{-\infty}^{\infty} x_1^2 \frac{1}{\sqrt{2\pi}\sigma} \exp(-x_1^2) \, dx = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} x_1^2 \exp(-x_1^2) \frac{dx}{\sqrt{2}\sigma}$$

$$= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} x_1^2 \exp(-x_1^2) \, dx_1$$

~~Integrating by parts:~~

Integrating by parts

$$= \frac{1}{\sqrt{\pi}} x_1 \int_{-\infty}^{\infty} x_1 e^{-x_1^2} dx_1 - \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} 1 \cdot \frac{1}{2} (-e^{-x_1^2}) dx_1$$

$$= \frac{x_1}{\sqrt{\pi}} \left[\frac{e^{-x_1^2}}{2} \right]_{-\infty}^{\infty} + \frac{1}{\sqrt{\pi}} \int \frac{1}{2} e^{-x_1^2} dx_1$$

$$= 0 + \frac{1}{\sqrt{\pi}} \frac{1}{2} \int_{-\infty}^{\infty} e^{-x_1^2} dx_1$$

$$= \frac{1}{2} \cdot \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\sigma}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

$$= \frac{1}{2} \cdot \cancel{\sqrt{\pi}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma}} \exp \frac{-(x-\mu)^2}{2\sigma^2}$$

$$= \frac{1}{2} \cdot 1 = \cancel{0}\, \frac{1}{2}.$$

Hence,

$$\boxed{H(X) = \ln(\sqrt{2\pi\sigma}) + \frac{1}{2}}$$

This is same as

$$H(X) = \frac{1}{2} \ln\left(\sqrt{2\pi\sigma}\right)^2 + \frac{1}{2}$$

$$= \frac{1}{2}\left(\ln(2\pi\sigma^2) + 1\right).$$

4

PROBLEM 2

$\frac{30}{30}$

1. $p(\text{disease}) = 0.01$

$p(\sim\text{disease}) = 0.99$

$p(+ve \mid \text{disease}) = 0.95$

$p(+ve \mid \sim\text{disease}) = 0.05$

(a) $P(+ve) = p(+ve, \text{disease}) + p(+ve, \sim\text{disease})$

$= p(+ve \mid \text{disease})\, p(\text{disease})$
$+ p(+ve \mid \sim\text{disease})\, p(\sim\text{disease})$

$= 0.95 \times 0.01 + 0.05 \times 0.99$

$= 0.059$ ✓

(b) $P(\text{disease} \mid +ve) = \dfrac{P(\text{disease}, +ve)}{P(+ve)}$

$= \dfrac{0.95 \times 0.01}{0.059}$

$= 0.161.$ ✓

2> $\lambda_{MLE} = \underset{\lambda}{\operatorname{argmax}}\; P(x_1 x_2 \cdots x_n \mid \lambda).$

Since the random variables are drawn independently,

$P(x_1 x_2 \cdots x_n \mid \lambda) = P(x_1 \mid \lambda)\, P(x_2 \mid \lambda) \cdots P(x_n \mid \lambda).$

$\ln P(x_1 x_2 \cdots x_n \mid \lambda) = \sum_{i=1}^{n} P(x_i \mid \lambda)$

$$\ln P(x_i|\lambda)$$

$$= x_i \ln\lambda - \lambda - \ln x_i!$$

$$\frac{d}{d\lambda} \ln P(x_i|\lambda) = x_i \cdot \frac{1}{\lambda} - 1.$$

$$\therefore \frac{d}{d\lambda} \ln P(x_1 x_2 \cdots x_n|\lambda)$$

$$= \frac{d}{d\lambda} \sum_{i=1}^{n} \ln P(x_i|\lambda)$$

$$= \sum_{i=1}^{n} \frac{x_i}{\lambda} - \sum_{i=1}^{n} 1$$

$$= \frac{\sum x_i}{\lambda} - n.$$

for maximum λ, $\frac{d}{d\lambda} \ln P(x_1 x_2 \cdots x_n|\lambda) = 0$

$$\therefore \frac{\sum x_i}{\lambda_{MLE}} - n = 0$$

$$\underline{\lambda_{MLE} = \hat{\lambda} = \frac{\sum x_i}{n}}$$

To show that it is an unbiased estimate, we write

$$E(\hat{\lambda}) = E\left(\frac{\sum x_i}{n}\right) = \frac{1}{n} E\left(\sum x_i\right)$$

$$= \frac{1}{n} \sum E(x_i)$$ (from linearity of expectation)

$$= \frac{1}{\cancel{n}} \cdot \cancel{n} \lambda$$

$$= \lambda.$$

(b)

~~$p(\lambda/x) \propto p(x/\lambda$~~

$$p(\lambda / x_1 x_2 \cdots x_n) \propto p(x_1 x_2 \cdots x_n / \lambda)\, p(\lambda)$$

$$p(x_1 x_2 \cdots x_n / \lambda) = \prod_{i=1}^{n} p(x_i / \lambda)$$

$$= \frac{\lambda^{\sum x_i} e^{-n\lambda}}{\prod_{i=1} x_i!}$$

$$p(\lambda) = p(\lambda / \alpha, \beta) = \frac{\beta^{\alpha} \lambda^{\alpha-1} e^{-\beta\lambda}}{\Gamma(\alpha)}$$

$$\therefore p(x_1 x_2 \cdots x_n / \lambda) \propto \frac{\lambda^{\sum x_i} e^{-n\lambda}}{\pi(x_i!)} \cdot \frac{\beta^{\alpha} \lambda^{\alpha-1} e^{-\beta\lambda}}{\Gamma(\alpha)}$$

$$= \frac{\lambda^{\sum x_i + \alpha - 1} e^{-(\beta\lambda + n\lambda)}}{\pi(x_i!)\,\Gamma(\alpha)} \cdot \checkmark$$

(c) To compute MAP, we take

$\underset{\lambda}{\text{argmax}}\; p(x_1 x_2 \cdots x_n / \lambda)$.

$$\ln p(x_1 x_2 \cdots x_n / \lambda) \propto \left(\sum x_i + \alpha - 1\right) \ln \lambda - \lambda(\beta + n) - \sum \ln x_i! - \ln \Gamma(\alpha)$$

$$\frac{d}{d\lambda} P(x_1, x_2 \ldots x_n \mid \lambda) = \frac{\left(\sum_{i=1}^{n} x_i + \alpha - 1\right)}{\lambda} - (\beta + n) = 0$$

$$\therefore \lambda_{MAP} = \frac{\sum_{i=1}^{n} x_i + \alpha - 1}{\beta + n}$$

PROBLEM 3

3.1> Attached the code added to entropy.c and prune-dt.c

6/20

3.2> 1) For the fully-grown tree:

0/20

Tree size: depth = 9
Nodes = 768

Accuracy on training set: 90.3%

Accuracy on testing set: 87.1%

2) For post-pruning with top-down approach:

Tree size: depth = 7
Nodes = 184

Accuracy on training set: 89.3%

Accuracy on testing set: 88.3%

3) For post-pruning with bottom-up approac

Tree size: depth = 9
Nodes = 512

Accuracy on training set: 89.9%

Accuracy on test set: 89.0%

→ For the case where the tree is not pruned, the size is highest. This also has the highest training accuracy, and lowest test accuracy, because it overfits the training data.

→ Both the prunings lead to an increase in test accuracy.

→ With the top down approach, if the accuracy increases by making the current node a leaf, we remove the subtree belonging the node.

In the bottom up approach, we first check the subtree of the current node. Thus, in the bottom up approach, we might decide it to retain the node, if, after pruning its subtree, we have already attained a high accuracy, (and hence pruning current node does not increase accuracy).

→ Thus, probability of current node being pruned in bottom up apprach is lower. Hence tree size is higher and tree depth is higher for bottom up apprach.

→ also, in bottom up approach, instead of removing the whole subtree of current node, we look for best possible subtree. (Top down approach is a special case of this, where removing the whole subtree gives best accuracy).

Thus, bottom up approach takes into account more combinations of nodes, and gives better training and test accuracies.

3.3>

Epsilon	No. of nodes Bottom up	top down
0.001 →	635	90
0.005 →	512	184
0.01 →	752	752
0.03 →	768	768

→ In general, as epsilon increases, probability that node will get pruned decreases. Hence less nodes are pruned, and tree size gets higher.

→ Only exception to this is $\epsilon = 0.001$ for bottom-up approach. Here, subtree is checked first. Although it is more probable for a subtree node to get pruned, that leads to the ~~roster~~ a accuracy changes, and then, removing the current node, does not increase accuracy in many cases. So, nodes removed is not very high.

??