1. (10%) Suppose that we have three colored boxes R (red), B (blue), and G (green). Box R contains 3 apples, 4 oranges, and 3 guavas, box B contains 2 apples, 0 orange, and 2 guavas, and box G contains 12 apples, 4 oranges, and 4 guavas. If a box is chosen at random with probabilities p(R)=0.2, p(B)=0.4, p(G)=0.4, and a piece of fruit is removed from the box (with equal probability of selecting any of the items in the box), then what is the probability of selecting guava? If we observe that the selected fruit is in fact an apple, what is the probability that it came from the blue box?

		apples	oranges	guavas	choosen
BOX	R	3	4	3	0.2
	B	2	0	2	0.4
	G	12	4	4	0.4

Let the event of selecting guavas be g

, the event of choosing box R, box G, box B be R, G, B

① $P(g \cap R) + P(g \cap B) + P(g \cap G)$

$= P(g|R)P(R) + P(g|B)P(B) + P(g|G)P(G)$

$= \frac{3}{10} \times 0.2 + \frac{2}{4} \times 0.4 + \frac{4}{20} \times 0.4$

$= 0.06 + 0.2 + 0.08 = 0.34$

② Let the event of selecting apples be a

$$P(B|a) = \frac{P(a \cap B)}{P(a \cap R) + P(a \cap B) + P(a \cap G)}$$

$$= \frac{\frac{2}{4} \times 0.4}{\frac{3}{10} \times 0.2 + \frac{2}{4} \times 0.4 + \frac{12}{20} \times 0.4} = \frac{0.2}{0.06 + 0.2 + 0.24}$$

$$= \frac{0.2}{0.5} = 0.4$$

※

2. (15%) Consider two nonnegative numbers a and b, and show that, if $a \leq b$, then $a \leq (ab)^{1/2}$. Use this result to show that, if the decision regions of a two-class classification problem are chosen to minimize the probability of misclassification, this probability will satisfy

$$p(mistake) \leq \int \{p(x, C_1)\, p(x, C_2)\}^{1/2}\, dx.$$

(Hint: Please refer to the textbook 1.5. Decision Theory)

Let R_1 be the distribution area of class C_1, and R_2 be the area of class C_2

$$p(\text{mistake}) = \int_{R_1} p(x, C_2)\, dx + \int_{R_2} p(x, C_1)\, dx \quad \cdots ①$$

In the error made in R_1 we always have $p(C_1|x) \geq p(C_2|x)$.

$$\Rightarrow p(C_2|x) \leq [p(C_1|x)\, p(C_2|x)]^{1/2}$$

$$\Rightarrow \int_{R_1} p(x, C_2)\, dx = \int_{R_1} p(C_2|x)\, p(x)\, dx$$

$$\leq \int_{R_1} [p(C_1|x)\, p(C_2|x)]^{1/2} p(x)\, dx = \int_{R_1} [p(x|C_1)\, p(x|C_2)]^{1/2}\, dx \quad \cdots ②$$

and similar situations apply for errors in R_2, $p(C_2|x) \geq p(C_1|x)$ when in R_2

$$\Rightarrow p(C_1|x) \leq [p(C_1|x)\, p(C_2|x)]^{1/2}$$

$$\Rightarrow \int_{R_2} p(x, C_1)\, dx = \int_{R_2} p(C_1|x)\, p(x)\, dx$$

$$\leq \int_{R_2} [p(C_1|x)\, p(C_2|x)]^{1/2} p(x)\, dx = \int_{R_2} [p(x|C_1)\, p(x|C_2)]^{1/2}\, dx \quad \cdots ③$$

Substitute ②. ③ back to ① :

$$p(\text{mistake}) = \int_{R_1} p(x, C_2)\, dx + \int_{R_2} p(x, C_1)\, dx$$

$$\leq \int_{R_1} [p(x, C_1)\, p(x, C_2)]^{1/2}\, dx + \int_{R_2} [p(x, C_1)\, p(x, C_2)]^{1/2}\, dx$$

$$= \int [p(x, C_1)\, p(x, C_2)]^{1/2}\, dx$$

※

3. (15%) Consider two variables x and y with joint distribution $p(x, y)$. Prove the following two results

$$E[x] = E_y\left[E_x[x|y]\right]$$

$$var[x] = E_y\left[var_x[x|y]\right] + var_y\left[E_x[x|y]\right].$$

Here $E_x[x|y]$ denotes the expectation of x under the conditional distribution $p(x|y)$, with a similar notation for the conditional variance.

① If x, y are continuous variables

$E_y[E_x[x|y]]$

$= \int_{-\infty}^{\infty} E_x[X|Y=y] f_Y(y)\, dy$

$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{X|Y}(x|y)\, dx\, f_Y(y)\, dy$

$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{X|Y}(x|y) f_Y(y)\, dx\, dy$

$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{X,Y}(x,y)\, dy\, dx$

$= \int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dy\, dx$

$= \int_{-\infty}^{\infty} x f_X(x)\, dx = E[X]$

If x, y are discrete variables

$E_y[E_x[x|y]]$

$= \sum_{y \in Y} E_x[x|Y=y]\, p(y)$

$= \sum_{y \in Y} \left(\sum_{x \in X} x\, p(x|y) \right) p(y)$

$= \sum_{y \in Y} \sum_{x \in X} x\, p(x|y)\, p(y)$

$= \sum_{x \in X} \sum_{y \in Y} x\, p(x,y)$

$= \sum_{x \in X} x \sum_{y \in Y} p(x,y)$

$= \sum_{x \in X} x\, p(x) = E[X]$

② $E_y[var_x[x|y]]$

$= E_y\left[E_x[x^2|y] - (E_x[x|y])^2 \right]$

$= E_y\left[E_x[x^2|y] \right] - E_y\left[E_x[x|y]^2 \right]$ ... ⓐ

$var_y[E_x[x|y]]$

$= E_y\left[E_x[x|y]^2 \right] - \{ E_y[E_x[x|y]] \}^2$ ... ⓑ

ⓐ + ⓑ : $E_y[var_x[x|y]] - var_y[E_x[x|y]]$

$= E_y\left[E_x[x^2|y] \right] - \{ E_y[E_x[x|y]] \}^2$

due to ① : $= E[x^2] - (E[x])^2 = var[x]$

※