1. The risk is: $E[1(A \neq Y)] = E[1(f(X) \neq Y)]$

$$= P(f(X) \neq Y)$$

The Bayes decision function is: $f^* = \underset{f}{\arg\min} P(f(X) \neq Y)$

$$= \underset{f}{\arg\max} P(f(X) = Y)$$

$$= \underset{f}{\arg\max} \sum_x P(Y = f(x) \mid X = x) \cdot P(X = x)$$

It apparently holds if, for every x, f^* satisfies with: $P(Y = f^*(x) \mid X = x) = \max_y P\{Y = y \mid X = x\}$.

So if $f^*(x) = \underset{y}{\arg\max} P(Y = y \mid X = x)$, f^* can be Bayes decision function of 0-1 loss function

3. $P(Y = y, X = x) = P(Y = y \mid X = x) \cdot P(X = x)$

$$= \frac{1}{10x}, \quad x \in \{1, \cdots, 10\} \wedge y \in \{1 \cdots x\}.$$

(a) $\ell(a, y)$ is square loss function,

$$f^*(x) = E(Y \mid X)$$

$$= \sum_y y \cdot P(Y = y \mid X = x)$$

$$= \sum_{y=1}^{x} \frac{y}{x}$$

$$= \frac{x+1}{2}$$

$A = Y$, so. $f^*(x) = \lfloor \frac{x+1}{2} \rfloor$ or $\lceil \frac{x+1}{2} \rceil$

(b) $R(f) = E[\ell(a, y)]$

$$= \sum_{x,y} |f(x) - y| P(X = x, Y = y)$$

$$= \sum_{x=1}^{10} \sum_{y=1}^{x} |f(x) - y| \cdot \frac{1}{10x}$$

$$= \frac{1}{10} \sum_{x=1}^{10} \frac{1}{x} \sum_{y=1}^{x} |f(x) - y|$$

If for every x, f satisfies with:

$$f(x) = \underset{K_x}{\arg\min} \sum_{y=1}^{x} |K_x - y|,$$

(3) For 0-1 loss function.

$f^*(x) = \arg\max_y P(Y=y \mid X=x)$

Because $Y|X$ is uniform distribution,

$f^*(x)$ can be arbitrary number

belonging to $\{1, \cdots, x\}$

then $f = f^*$.

If x is even, $\sum_{y=1}^{x} |K_x - y| = \sum_{n=1}^{x/2} |K_x - n| + |x - K_x - n|$

$$\geq \sum_{n=1}^{x/2} |x - 2n|$$

else, $\sum_{y=1}^{x} |K_x - y| = \sum_{n=1}^{(x-1)/2} |K_x - n| + |x - K_x - n| + |\frac{x+1}{2} - K_x|$

$$\geq \sum_{y=1}^{(x-1)/2} |x - 2n|$$

In both cases, the equal condition can be achieved if K_x is the median of y sequence.

Therefore, $f^*(x) = \lfloor \frac{x+1}{2} \rfloor$ or $\lceil \frac{x+1}{2} \rceil$

2. We have: $E(Y) = E[E(Y|X)]$, $f^*(X) = E(Y|X)$

$R(f^*) = E\{[Y - E(Y|X)]^2\}$

$= E\{E\{[Y - E(Y|X)]^2 | X\}\}$

$= E[Var(Y|X)]$

$Var(Y) = E\{[Y - E(Y)]^2\}$

$= E[Var(Y|X)] + 2E\{[Y - E(Y|X)][E(Y|X) - E(Y)]\} + E\{[E(Y|X) - E(Y)]^2\}$

$= E[Var(Y|X)] + 2E\{[Y - E(Y|X)] \cdot E(Y|X)\} - 2E(Y) \cdot \{E(Y) - E[E(Y|X)]\}$

$+ E\{\{E(Y|X) - E[E(Y|X)]\}^2\}$

$= E[Var(Y|X)] + Var[E(Y|X)] + 2E[Y \cdot E(Y|X)] - 2E[E^2(Y|X)]$

$= E[Var(Y|X)] + Var[E(Y|X)] + 2E\{E\{[Y \cdot E(Y|X)] | X\}\} - 2E[E^2(Y|X)]$

$= E[Var(Y|X)] + Var[E(Y|X)]$

So $Var(Y) - R(f^*) = Var[E(Y|X)]$

4. $\hat{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(X_i), Y_i)$

则 $E[\hat{R}_n(f)] = \frac{1}{n} \cdot \sum_{i=1}^{n} E(\ell(f(X_i), Y_i))$

$= \frac{1}{n} \cdot n \cdot E(\ell(f(X), Y))$

$= E(\ell(f(X), Y))$

$= R(f)$.

$Var[\hat{R}_n(f)] = Var[\frac{1}{n}\sum_{i=1}^{n} \ell(f(X_i), Y_i)]$

$= \frac{1}{n^2}\sum_{i=1}^{n} Var[\ell(f(X_i), Y_i)]$

$= \frac{1}{n} Var[\ell(f(X), Y)]$

$\lim_{n\to\infty} Var[\hat{R}_n(f)] = 0$. So $\hat{R}_n(f)$ is consistent.

So $\hat{R}_n(f)$ is the unbiased estimation of $R(f)$.

5. (a) Because $\mathcal{F}_1$ is the hypothesis space of constant function.

The ER should be: $\hat{R}(f) = \frac{1}{n}\sum_{i=1}^{n} 1(C \neq Y_i)$

In this data set, C should be 3 or 5.

So EMR is: $\hat{f}(x) = 3$ or $\hat{f}(x) = 5$ and it's not unique.

EM is 3/5.

(b) One choice of $\hat{f}$ is: $\hat{f}(x) = \begin{cases} 5, & 0 \le x < 0.15 \\ 3, & 0.15 \le x \le 1 \end{cases}$ with EM = 2/5.

It is not unique, $\hat{f}(x)$ can also be like: $\hat{f}(x) = \begin{cases} 3, & 0 \le x < 0.95 \\ 5, & 0.95 \le x \le 1. \end{cases}$

6. (a) For 0-1 loss function. $1(f^*(X) \neq Y)$ equals to 1 almost everywhere.

So $R(f^*) = E(1) = 1$

(b) For square loss function, $f^*(x) = E(Y|X) = a+bx$.

$R(f^*) = E[Var(Y|X)] = E(1) = 1$.

(c). EM is 0 for full hypothesis space.

For example, if $f^*(x)$ is a stair function which is properly set for these data points. The difference between $f^*(x)$ and y can be 0. So the EM is 0

(d) For liner function, the known close form solution for w is:

$$\hat{\omega} = (X^T \cdot X)^{-1} \cdot X^T \cdot y$$

$$X = \begin{pmatrix} 0 & 1 \\ 0 & 1 \\ 1 & 1 \\ 2.5 & 1 \\ -4 & 1 \end{pmatrix}, \quad y = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 3.1 \\ -2.1 \end{bmatrix}, \quad \hat{\omega} = \begin{pmatrix} 0.856 \\ 1.468 \end{pmatrix}$$

So. $\hat{R}_r(\hat{f}) = \frac{1}{5} \| X \cdot \hat{\omega} - y \|_2^2 = 0.247$.

(e). Let $X = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \\ 2.5 & 6.25 & 1 \\ -4 & 16 & 1 \end{bmatrix}$, $y = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 3.1 \\ -2.1 \end{bmatrix}$.

then, $\hat{w} = (X^T \cdot X)^{-1} \cdot X \cdot y = \begin{pmatrix} 0.755 \\ -0.052 \\ 1.718 \end{pmatrix}$.

$\hat{R}_5(\hat{f}) = \frac{1}{5} \| X \cdot \hat{w} - y \|_2^2 = 0.193$