Problem 1.

(i) Bayes classifier is

$$f^*(x) = \begin{cases} 1 & \text{if } r(x) > 1/2 \quad (\text{i.e. } x > c) \\ 0 & \text{otherwise} \end{cases}$$

(ii) Bayes risk is

$$\begin{aligned}
\mathcal{R}^* = \mathcal{R}(y^*) &= \mathbf{P}(Y \neq f^*(X)) \\
&= \mathbf{P}(Y = 1,\, X < c) + \mathbf{P}(Y = 0,\, X \geq c) \\
&= \int_0^c \mathbf{P}(Y = 1 | X = x)\mathbf{P}(X \in dx) + \int_c^\infty \mathbf{P}(Y = 0 | X = x)\mathbf{P}(X \in dx) \\
&= \int_0^c r(x)\mathbf{P}(X \in dx) + \int_c^\infty (1 - r(x))\mathbf{P}(X \in dx) \\
&= \int_0^\infty \min(r(x), 1 - r(x))\mathbf{P}(X \in dx) \\
&= \mathbf{E}_X \left(\frac{\min(c, X)}{c + X} \right),
\end{aligned}$$

where the second-last equality follows from the fact that $r(x) < 1 - r(x) \Leftrightarrow r(x) < 1/2 \Leftrightarrow x < c$.

(iii) If $X = c$ with probability one, we have that $\mathcal{R}^* = 1/2$.

If X is uniformly distributed on the interval $[0,\, 4c]$,

$$\mathcal{R}^* = \frac{1}{4c}\int_0^{4c} \frac{\min(c, x)}{c + x} dx = \frac{1}{4c}\int_0^c \frac{x}{c + x} dx + \frac{1}{4c}\int_c^{4c} \frac{c}{c + x} dx = \frac{1}{4}\log\left(\frac{5e}{4}\right) \approx 0.306\,.$$

On $[0,\, 2c]$, Bayes risk becomes 0.356. The further away we stand from c (= discrimination point), the easier the classification (Bayes risk smaller).

Problem 2.
Note that

$$\mathbf{E}|Y - f(X)| = \int\int |y - f(x)| d\mathbf{P}_{X,Y}(x, y) = \int \left(\int |y - f(x)| d\mathbf{P}_{Y|X}(y|x) \right) d\mathbf{P}_X(x)\,.$$

For each x, we need to minimize the inner integral with respect to the conditional probability measure of Y given $X = x$. The minimizer is known to be the median (prove it - see lecture notes in PT). The resulting optimal regression function is then the conditional median of Y given X.

Problem 3.

(i) Expanding the risk,

$$\begin{aligned}\mathcal{R}(f) &= \mathbf{E}\{(Y - \mathbf{E}(Y|X) + \mathbf{E}(Y|X) - f(X))^2\} \\ &= \mathbf{E}\{(Y - \mathbf{E}(Y|X))^2\} + \mathbf{E}\{(f(X) - \mathbf{E}(Y|X))^2\} + 2 \times \text{cross-product}\end{aligned}$$

The cross-product term can be seen to be equal to 0, since

$$\mathbf{E}_X \mathbf{E}_{Y|X}(Y - \mathbf{E}(Y|X))(\mathbf{E}(Y|X) - f(X)) = \mathbf{E}_X\{(\mathbf{E}(Y|X) - f(X))\mathbf{E}_{Y|X}(Y - \mathbf{E}(Y|X))\} = 0\,.$$

We get the decomposition

$$\mathcal{R}(f) = \mathbf{E}\{\mathrm{var}(Y \mid X)\} + \mathbf{E}\left\{(f(X) - \mathbf{E}(Y \mid X))^2\right\}\,,$$

as required. The first term represent the irreducible error (or noise term). This term is beyond our control and cannot be reduced. The second term represents what we want to minimise. Ideally, the optimal solution is the conditional mean. In practice however, this term cannot be made exactly zero.

(ii) Making use of the decomposition derived in *(i)*, the expected risk can be written

$$\mathbf{E}_{\mathcal{L}_n} \mathcal{R}(\hat{f}_n) = \mathbf{E}\{\mathrm{var}(Y \mid X)\} + \mathbf{E}_{\mathcal{L}_n}\mathbf{E}\left\{(\hat{f}_n(X) - \mathbf{E}(Y \mid X))^2 \mid \mathcal{L}_n\right\}\,.$$

Introducing the variable $\mathbf{E}_{\mathcal{L}_n}(\hat{f}_n(X)|X)$ in the last term,

$$\mathbf{E}_{\mathcal{L}_n}\mathbf{E}\left\{(\hat{f}_n(X) - \mathbf{E}_{\mathcal{L}_n}(\hat{f}_n(X)|X) + \mathbf{E}_{\mathcal{L}_n}(\hat{f}_n(X)|X) - \mathbf{E}(Y \mid X))^2 \mid \mathcal{L}_n\right\}\,,$$

and expanding the square,

$$\begin{aligned}&\mathbf{E}_{\mathcal{L}_n}\mathbf{E}\left\{[\hat{f}_n(X) - \mathbf{E}_{\mathcal{L}_n}(\hat{f}_n(X)|X)]^2 \mid \mathcal{L}_n\right\} \\ &\quad + \mathbf{E}_{\mathcal{L}_n}\mathbf{E}\left\{[\mathbf{E}_{\mathcal{L}_n}(\hat{f}_n(X)|X) - \mathbf{E}(Y \mid X)]^2 \mid \mathcal{L}_n\right\} \\ &\quad\quad + 2\mathbf{E}_{\mathcal{L}_n}\mathbf{E}\left\{\left(\hat{f}_n(X) - \mathbf{E}_{\mathcal{L}_n}(\hat{f}_n(X)|X)\right)\left(\mathbf{E}_{\mathcal{L}_n}(\hat{f}_n(X)|X) - \mathbf{E}(Y \mid X)\right) \mid \mathcal{L}_n\right\}\,.\end{aligned}$$

Again, the cross-product term vanished since it is equal to

$$\begin{aligned}&\mathbf{E}\,\mathbf{E}_{\mathcal{L}_n}\left\{\left(\hat{f}_n(X) - \mathbf{E}_{\mathcal{L}_n}(\hat{f}_n(X)|X)\right)\left(\mathbf{E}_{\mathcal{L}_n}(\hat{f}_n(X)|X) - \mathbf{E}(Y \mid X)\right) \mid X\right\} \\ &\quad = \mathbf{E}\left([\mathbf{E}_{\mathcal{L}_n}(\hat{f}_n(X)|X) - \mathbf{E}(Y \mid X)]\mathbf{E}_{\mathcal{L}_n}\left\{\left(\hat{f}_n(X) - \mathbf{E}_{\mathcal{L}_n}(\hat{f}_n(X)|X)\right) \mid X\right\}\right) = 0\,.\end{aligned}$$

The decomposition

$$\mathbf{E}\{\mathrm{var}(Y \mid X)\} + \mathbf{E}\left\{[\mathbf{E}_{\mathcal{L}_n}(\hat{f}_n(X) \mid X) - \mathbf{E}(Y \mid X)]^2\right\} + \mathbf{E}\left\{[\hat{f}_n(X) - \mathbf{E}_{\mathcal{L}_n}(\hat{f}_n(X) \mid X)]^2\right\}$$

follows. The first term represents the irreducible error, the second term the square bias, and the last term the variance.

Problem 4.

(i) Bayes classifier assigns label 1 if $\mathbf{P}(Y=1|X=x) > \mathbf{P}(Y=0|X=x)$, or equivalently, if $\mathbf{P}(Y=1,\, X=x) > \mathbf{P}(Y=0\,, X=x)$. This in turns implies that $\mathbf{P}(X=x|Y=1)\mathbf{P}(Y=1) > \mathbf{P}(X=x|Y=0)\mathbf{P}(Y=0)$. Since $\mathbf{P}(Y=1)=2/3$, we get that Bayes classifier assigns label 1 if $2\,\mathbf{P}(X=x|Y=1) > \mathbf{P}(X=x|Y=0)$. We get:

x	1	2	3	4	5	6	7	8	9	10
$\mathbf{P}(X=x\,\vert\,Y=1)$	0.04	0.07	0.06	0.03	0.24	0	0.02	0.09	0.25	0.2
$\mathbf{P}(X=x\,\vert\,Y=0)$	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Bayes classifier	0	1	1	0	1	0	0	1	1	1

The error rate is

$$\sum_{x|\text{ obs classified as }0} \mathbf{P}(X=x|Y=1)\mathbf{P}(Y=1) + \sum_{x|\text{ obs classified as }1} \mathbf{P}(X=x|Y=0)\mathbf{P}(Y=0)\,,$$

which is 2/3(0.04+0.03+0+0.02)+1/3(0.1+0.1+0.1+0.1+0.1+0.1)=0.26.

(ii) We have

x	2	4	6	8	10
$\mathbf{P}(X=x\,\vert\,Y=1)$	0.11	0.09	0.24	0.11	0.45
$\mathbf{P}(X=x\,\vert\,Y=0)$	0.2	0.2	0.2	0.2	0.2
Bayes classifier	1	0	1	1	1

The error rate is 2/3(0.09)+1/3(0.2+0.2+0.2+0.2+0.2)=0.326.