(1) Suppose there is a single feature, denoted by $X$ which takes values in the set $\{1, 2, 3\}$ and a binary label $Y \in \{0, 1\}$. The distribution $D$ is described as follows:

$$\Pr(X = i) = \frac{1}{3}, \ i \in \{1, 2, 3\}$$

$$\Pr(Y = 1 | X = i) = 1, \ i \in \{1, 3\}$$

$$\Pr(Y = 0 | X = i) = 1, \ i \in \{2\}$$

Let $h$ be the classifier: $h(x) = 1$ if $x > 1.5$ and 0 otherwise. Calculate the error of $h$ with respect to $D$.

**Solution:**

$$\text{err}(h) = \Pr(h(X) \neq Y)$$

$$= \Pr(h(X) \neq Y, X = 1) + \Pr(h(X) \neq Y, X = 2) + \Pr(h(X) \neq Y, X = 3)$$

$$= \Pr(h(X) \neq Y | X = 1) \Pr(X = 1) + \Pr(h(X) \neq Y | X = 2) \Pr(X = 2)$$

$$+ \Pr(h(X) \neq Y | X = 3) \Pr(X = 3)$$

$$= \Pr(Y = 1 | X = 1) \Pr(X = 1) + \Pr(Y = 0 | X = 2) \Pr(X = 2) + \Pr(Y = 0 | X = 3) \Pr(X = 3)$$

$$= 1 \times \frac{1}{3} + 1 \times \frac{1}{3} + 0 \times \frac{1}{3}$$

$$= \frac{2}{3}$$

(2) Suppose you have a dataset of images of digits, and you use it to build a linear classifier for classifying the images into the corresponding digits. Which of the following actions may reduce the bias of your classifier? Justify your answer.

- (a) Collect more training data.
- (b) Switch to a kernel classifier with a quadratic kernel (that is, where $K(x, z) = \langle x, z \rangle^2$).

**Solution:**

- (a) cannot reduce the bias, since the bias only depends on the concept class and data distribution, and is not affected by the size of training data.
- (b) may reduce the bias since using kernels enables it to model more complex data distribution.