#2 a). $S_1, S_2, S_3, \ldots, S_k$ are disjoint dividents of X_j, therefore,

$$\frac{P}{P+n} = \frac{P_k}{P_k+n_k} \quad \text{for all } k=1,\ k=2,\ \ldots,\ k=k$$

$$\therefore H(S_k) = B\left(\frac{P}{P+n}\right) \text{ for all } k$$

$$H(S|X_j) = H(S_1)\frac{P_1+n_1}{P+n} + H(S_2)\frac{P_2+n_2}{P+n} + \cdots + H(S_k)\frac{P_k+n_k}{P+n}$$

$$= H(S) \cdot \frac{P_1+P_2+\cdots+P_k+n_1+n_2+\cdots+n_k}{P+n}$$

$$= B\left(\frac{P}{P+n}\right)$$

Therefore, $H(S) - H(S|X_j) = B\left(\frac{P}{P+n}\right) - B\left(\frac{P}{P+n}\right) = 0$
Thus, the information gain of this attribute is 0.

#3 a) Since a point can be it's own neighbor, so $k=0$ minimizes the training set error, the resulting training error is 0.

b) Too big k may lead to misclassify on datapoints, too small k may leads datapoints not fit in graph.

c). $k=5$ or $k=7$ minimizes the LOO-CV error for this dataset. The resulting error is $\frac{4}{14}$.