Exercise: Mutual information between class and binary features

Consider a supervised learning problem where we want to learn a mapping from $\mathbf{x}$ to y. Suppose $\mathbf{x} \in \{0,1\}^D$. Show that

$$I(X_j, Y) \quad = \quad \sum_c \left[\theta_{jc}\pi_c \log \frac{\theta_{jc}}{\theta_j} + (1-\theta_{jc})\pi_c \log \frac{1-\theta_{jc}}{1-\theta_j} \right] \tag{1}$$

where $\pi_c = p(y=c)$, $\theta_{jc} = p(x_j = 1|y=c)$, and $\theta_j = p(x_j = 1) = \sum_c \pi_c \theta_{jc}$.