Homework 1

Adheesh Juvekar, Arul Thileeban

September 18, 2019

Writter Problems

1. $\hat{\pi}_0 = \frac{1}{|\mathcal{D}|}\sum_{i\in\mathcal{D}} \mathbb{I}(y_i = 0)$
   $\hat{\pi}_0 = \frac{1}{10} \times 5$
   $\hat{\pi}_0 = 0.5$

   $\hat{\pi}_1 = \frac{1}{|\mathcal{D}|}\sum_{i\in\mathcal{D}} \mathbb{I}(y_i = 1)$
   $\hat{\pi}_1 = \frac{1}{10} \times 5$
   $\hat{\pi}_1 = 0.5$

   $H(\hat{\pi}) = -\left(\sum_{c=0}^{1} \hat{\pi}_c log_e \hat{\pi}_c\right)$
   $H(\hat{\pi}) = -(0.5 \times log_e 0.5 + 0.5 \times log_e 0.5)$
   $H(\hat{\pi}) = -(-0.6931)$
   $H(\hat{\pi}) = 0.6931$

   $InfoGain(X_1) = H(\hat{\pi}) - H(\ \hat{\pi}|X_1)$
   $InfoGain(X_1) = 0.6931 - \left[\left(-\frac{8}{10} \times \left(\frac{5}{8} \times log_e(\frac{5}{8}) + \frac{3}{8} \times log_e(\frac{3}{8})\right)\right) + \left(-\frac{2}{10} \times (0 + 1 \times log_e(1))\right)\right]$
   $InfoGain(X_1) = 0.6931 - [0.5292 + 0]$
   $InfoGain(X_1) = 0.1639$

   $InfoGain(X_2) = H(\hat{\pi}) - H(\ \hat{\pi}|X_2)$
   $InfoGain(X_2) = 0.6931 - \left[\left(-\frac{6}{10} \times \left(\frac{4}{6} \times log_e(\frac{4}{6}) + \frac{2}{6} \times log_e(\frac{2}{6})\right)\right) + \left(-\frac{4}{10} \times \left(\frac{1}{4} \times log_e(\frac{1}{4}) + \frac{3}{4} \times log_e(\frac{3}{4})\right)\right)\right]$
   $InfoGain(X_2) = 0.6931 - [0.3819 + 0.2249]$
   $InfoGain(X_2) = 0.0863$

   $InfoGain(X_3) = H(\hat{\pi}) - H(\ \hat{\pi}|X_3)$
   $InfoGain(X_3) = 0.6931 - \left[\left(-\frac{6}{10} \times \left(\frac{3}{6} \times log_e(\frac{3}{6}) + \frac{3}{6} \times log_e(\frac{3}{6})\right)\right) + \left(-\frac{4}{10} \times \left(\frac{2}{4} \times log_e(\frac{2}{4}) + \frac{2}{4} \times log_e(\frac{2}{4})\right)\right)\right]$
   $InfoGain(X_3) = 0.6931 - [0.6931]$
   $InfoGain(X_3) = 0$

   $InfoGain(X_4) = H(\hat{\pi}) - H(\ \hat{\pi}|X_4)$
   $InfoGain(X_4) = 0.6931 - \left[\left(-\frac{5}{10} \times \left(\frac{3}{5} \times log_e(\frac{3}{5}) + \frac{2}{5} \times log_e(\frac{2}{5})\right)\right) + \left(-\frac{5}{10} \times \left(\frac{2}{5} \times log_e(\frac{2}{5}) + \frac{3}{5} \times log_e(\frac{3}{5})\right)\right)\right]$
   $InfoGain(X_4) = 0.6931 - [0.3365 + 0.3365]$
   $InfoGain(X_4) = 0.0201$

Therefore, $InfoGain(X_1) > InfoGain(X_2) > InfoGain(X_4) > InfoGain(X_3)$
Hence, **Decision tree algorithm will choose X_1 for the first split of the following dataset.**

2. (a) $p(D/\theta) = \theta^{N_1}(1-\theta)^{N_0}$

   Take log on both sides

   $logp(D/\theta) = N_1 log\theta + N_0 log(1-\theta)$

   Partially differentiate and equation to 0 to find maxima

   $\mathbf{\frac{N_1}{\widehat{\theta}} + \frac{N_0}{1-\widehat{\theta}} = 0}$

   $\mathbf{\frac{N_1(1-\widehat{\theta})+N_0\widehat{\theta}}{\widehat{\theta}(1-\widehat{\theta})} = 0}$

   Rearrange the terms to have theta in one side

   $\widehat{\theta} = \mathbf{\frac{N_1}{N_1+N_0}}$

   (b) $P(\theta)\ \alpha\ \theta^{\alpha}(1-\theta)^{\alpha}$

   $P(\theta) = K\theta^{\alpha}(1-\theta)^{\alpha} - -(1)$

   $P(\theta|D) = \frac{P(D|\theta).P(\theta)}{P(D)}$

   Take log on both sides

   $logP(\theta|D) = logP(D|\theta) + logP(\theta) - logP(D)$

   Put (1) in the previous equation

   $= N_1 log\theta + N_0 log(1-\theta) + \alpha log\theta + \alpha log(1-\theta)$

   Partially differentiate the equation to 0 to find maxima

   $= \frac{N_1}{\theta} + \frac{N_0}{(1-\theta)} + \frac{\alpha}{\theta} + \frac{\alpha}{1-\theta}$

   Rearrange with theta at LHS

   $\theta = \frac{N_1+\alpha}{N_1+N_0+2\alpha}$