تمرین تحلیل داده کاوی — تینا جزایری ۹۸۱۳۰۰۷ و

امیرحسین رجبی ۹۸۱۳۰۱۳

۱) در الگوریتم Hunt، splitting نودهای درخت تصمیم تا زمانی ادامه دارد که همه نمونه‌هایی که در آن قرار می‌گیرند متعلق به یک کلاس باشند. در واقع ناخالصی هر نود صفر باشد.

۳)

a) $1 - P_1^2(\text{train\_set}) - P_0^2(\text{train\_set})$

$= 1 - 1/4 - 1/4 = 1/2$

b) چون ID، nominal است، می‌توانیم از multiway split استفاده کنیم و I(children) را بر این اساس محاسبه کنیم.

$$I = \sum \frac{N_{child}}{N} I_{child} = \sum_{i=1}^{20} \frac{1}{20} \times (1 - 1^2) = 0$$

c)

$$I = \underbrace{\frac{10}{20} \times \left(1 - \left(\frac{4}{10}\right)^2 - \left(\frac{6}{10}\right)^2\right)}_{\text{female}} + \underbrace{\frac{10}{20} \times \left(1 - \left(\frac{6}{10}\right)^2 - \left(\frac{4}{10}\right)^2\right)}_{\text{male}}$$

$$= \frac{48}{100}$$

d)

$$\underbrace{\frac{8}{20} \times (1 - 1^2)}_{\text{sports}} + \underbrace{\frac{8}{20} \times \left(1 - \left(\frac{7}{8}\right)^2 - \left(\frac{1}{8}\right)^2\right)}_{\text{luxury}} + \underbrace{\frac{4}{20} \times \left(1 - \left(\frac{1}{4}\right)^2 - \left(\frac{3}{4}\right)^2\right)}_{\text{family}}$$

$$= \frac{\cancel{8}^{\,2}}{20} \times \frac{14}{\cancel{64}_{\,16}} + \frac{4}{20} \times \frac{6}{16} = \frac{52}{320} = 0.1625$$

e)

$$\underbrace{\frac{5}{20} \times \overbrace{\left(1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2\right)}^{12/25}}_{\text{small}} + \underbrace{\frac{4}{20} \times \overbrace{\left(1 - 1/4 - 1/4\right)}^{1/2}}_{\text{large}} + \underbrace{\frac{4}{20} \times \overbrace{\left(1 - 1/4 - 1/4\right)}^{1/2}}_{\text{extra}}$$

$$+ \underbrace{\frac{7}{20} \times \underbrace{\left(1 - \left(\frac{3}{7}\right)^2 - \left(\frac{4}{7}\right)^2\right)}_{24/49}}_{\text{medium}} = \frac{5}{20} \times \frac{12}{25} + \frac{4}{20} + \frac{7}{20} \times \frac{24}{49}$$

$$= \frac{3}{25} + \frac{4}{20} + \frac{1}{35} \simeq 0.348\ldots$$

f) بستگی به Car Type چون Impurity کمتری دارند و نود ها سریع تر خالص می شوند و عمق درخت کم می شود.

g) با وجود اینکه $I_{children}(ID) = 0$ چون عملا overfitting رخ داده و مدل چیزی یاد نمی گیرد بلکه حفظ می کند و هیچ داده جدید را نمی تواند predict کند.

a)

$$Entropy = -\sum p_i \lg p_i \longrightarrow I_{parent}$$ ۴

$$= -\left(P_+ \lg P_+ + P_- \lg P_-\right) = -\frac{4}{9}\lg\frac{4}{9} - \frac{5}{9}\lg\frac{5}{9}$$

b)

$$I(a_1) = \underbrace{\frac{4}{9} \times \left(-\frac{3}{4}\lg\frac{3}{4} - \frac{1}{4}\lg\frac{1}{4}\right)}_{a_1 = T} + \underbrace{\frac{5}{9} \times \left(-\frac{4}{5}\lg\frac{4}{5} - \frac{1}{5}\lg\frac{1}{5}\right)}_{a_1 = F}$$

$$I(a_2) = \underbrace{\frac{5}{9} \times \left(-\frac{2}{5}\lg\frac{2}{5} - \frac{3}{5}\lg\frac{3}{5}\right)}_{a_2 = T} + \underbrace{\frac{4}{9} \times \left(-\frac{1}{2}\lg\frac{1}{2} - \frac{1}{2}\lg\frac{1}{2}\right)}_{a_2 = F}$$

c)

$$\Delta_{info} = I_{parent} - I_{children}$$

پس $I_{children}$ را حساب می کنیم. ابتدا مقادیر a_3 را سورت کرده، پس:

3 instances: | 0.5 | 1+ | 2 | 3- | 3.5 | 4+ | 4.5

6 instances: | 5- 5- | 5.5 | 6+ | 6.5 | 7+ 7+ | 7.5 | $\bar{8}$ | 8.5

برای هر یک از مقادیر قرمز باید Δ_{info} محاسبه کنیم. مثلا برای $X < 4.5$, $X \geq 4.5$

$$\Delta_{info} = I_p - I_c = \underbrace{-\frac{4}{9}\lg\frac{4}{9} - \frac{5}{9}\lg\frac{5}{9}}_{I_p} - \underbrace{\left(\frac{6}{9} \times I_{\geq 4.5} + \frac{3}{9} \times I_{<4.5}\right)}_{I_c}$$

$$I_{<4.5} = -\underset{+}{\frac{2}{3}}\lg\frac{2}{3} - \underset{-}{\frac{1}{3}}\lg\frac{1}{3} \quad \text{و} \quad I_{\geq 4.5} = -\underset{+}{\frac{3}{6}}\lg\frac{3}{6} - \underset{-}{\frac{3}{6}}\lg\frac{3}{6} = -\lg\frac{1}{2}$$ که

به طور مشابه برای مابقی موارد قرمز محاسبه می شود.

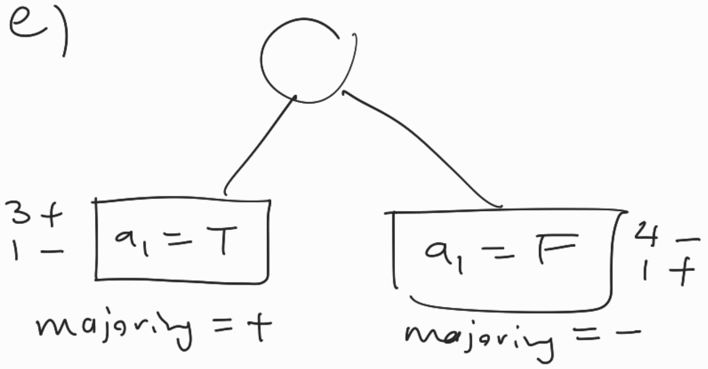
d) برای a_3 باید برای همه split ها حساب شود و سپس با a_1 و a_2 مقایسه شود. برای a_1:

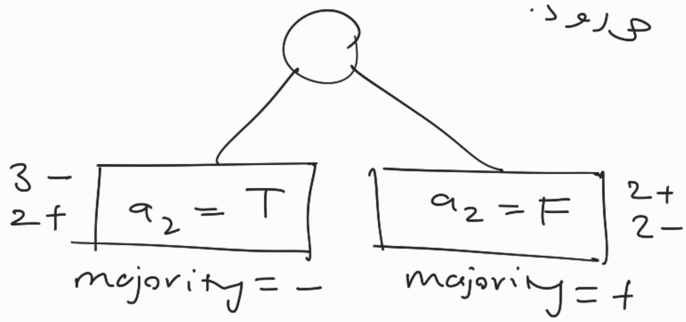
$$\Delta_{info}(a_2) = I_{parent} - \underbrace{\frac{5}{9} \times \left(-\frac{2}{5}\lg\frac{2}{5} - \frac{3}{5}\lg\frac{3}{5}\right)}_{a_2 = T} - \underbrace{\frac{4}{9} \times \left(-\frac{1}{2}\lg\frac{1}{2} - \frac{1}{2}\lg\frac{1}{2}\right)}_{a_2 = F}$$

$$\Delta(a_1) = I_{parent} - \underbrace{\frac{4}{9} \times \left(-\frac{3}{4}\lg\frac{3}{4} - \frac{1}{4}\lg\frac{1}{4}\right)}_{a_1 = T} - \underbrace{\frac{5}{9} \times \left(-\frac{1}{5}\lg\frac{1}{5} - \frac{4}{5}\lg\frac{4}{5}\right)}_{a_1 = F}$$

هرکدام که بزرگتر باشد بین $\Delta(a_1)$، $\Delta(a_2)$ و برای هر split, a_3 که در بالا یک حالت به طور کامل بحث شد، بزرگترین انتخاب شود همان برای splitting خود به کار می‌رود.

e)





$$\underset{\text{on training set}}{\text{miss}} = \frac{f_{+-} + f_{-+}}{9} = \frac{2}{9} < \text{miss} = \frac{2+2}{9} = \frac{4}{9}$$

Hence a_1 is best split.

f) $\Delta(a_1) = I_p -$

$$\Delta(a_2) = I_p - \frac{5}{9} \times \left(1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2\right) - \frac{4}{9} \times \left(1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2\right)$$

The bigger is the best split because we would obtain much purer children nodes.

10

a) optimistic approach $\rightarrow$ $err_{gen} = err_{train}$

pessimistic approach $\rightarrow$ $err_{gen} = err_{train} + 0.5 \times \frac{K}{N_{train}}$

$$err_{train} = \frac{\overset{A=0,B=1}{1} + \overset{A=1,C=0}{3} + \overset{A=1,C=1}{1}}{10} = \frac{1}{2}$$

$K = 4$, $N_{train} = 10$

Hence optimistic $= 0.5$, pessimistic $= \frac{1}{2} + \frac{1}{2} \times \frac{4}{10} = \frac{7}{10} = 0.7$

$$\text{validation} = \frac{\overset{A=0,B=1}{1}}{5} = 0.2$$

14

underfitting زمانی رخ می‌دهد که مدل الگوهای موجود در train set را یاد نگرفته باشد و خطای مدل روی هر دو train set و test set بسیار بزرگ باشد. در این موارد مدل بسیار ساده است و به اندازه کافی complex نیست که بتواند pattern ها را learn کند. مثلا در مدل درخت تصمیم ممکن است تعداد leaf ها بسیار کوچک و ناکافی باشد.