۱) مسئله اول

چهار ویژگی (Tenure) داریم:

X_1) نرخ شکایت از مدرسه

- بالا: H:1 S:6
- پایین: H:4 S:3

X_2) طول بدنه ایمیل

- کوتاه: H:2 S:3
- متوسط: H:3 S:2
- طولانی: H:0 S:4

X_3) مزیت ایمیل

- مشکوک: H:2 S:6
- نامشکوک: A:3 S:3

X_4) دامنه ایمیل

- gmail: H:2 S:4
- academic: H:1 S:3
- کنایی: H:2 S:2

باید ویژگی information gain را محاسبه کنیم.

طبق اسلایدها داریم:

$$\underset{i}{\operatorname{argmax}}\, I(Y, X_i) = \underset{i}{\operatorname{argmin}}\, H(Y|X_i)$$

حال هر کدام از $H(Y|X_i)$ ها را محاسبه می‌کنیم.

$$H(Y) = -\sum_y P(Y=y)\log P(Y=y)$$

$$H(Y|X_1) = \sum_x P(X_1=x)\, H(Y|X_1=x)$$

$$= \frac{-\left(\frac{3}{7}\log\frac{3}{7} + \frac{4}{7}\log\frac{4}{7}\right) - \left(\frac{1}{3}\log\frac{1}{3} + \frac{6}{7}\log\frac{6}{7}\right)}{2} = \frac{+0.29 + 0.17}{2} = 0.23$$

$$H(Y|X_2) = \frac{-3\left(\frac{0}{4}\log 0 + \frac{4}{4}\log\frac{4}{4}\right) - 5\left(\frac{3}{5}\log\frac{3}{5} + \frac{2}{5}\log\frac{2}{5}\right) - 5\left(\frac{2}{5}\log\frac{2}{5} + \frac{3}{5}\log\frac{3}{5}\right)}{14}$$

$$\Rightarrow H(Y|X_2) = \frac{+0.29\times 5 + 4\times 0.29}{14} \simeq 0.18$$

$$H(Y|X_3) = \frac{-8\left(\frac{2}{8}\log\frac{2}{8} + \frac{6}{8}\log\frac{6}{8}\right) - 6\left(\frac{3}{6}\log\frac{3}{6} + \frac{3}{6}\log\frac{3}{6}\right)}{14}$$

$$H(Y|X_3) = \frac{+8\times 0.24 + 6\times 0.3}{14} = 0.26$$

$$H(Y|X_4) = \frac{-4(\frac{1}{4}\log\frac{1}{4} + \frac{3}{4}\log\frac{3}{4}) - 6(\frac{2}{6}\log\frac{2}{6} + \frac{4}{6}\log\frac{4}{6}) - 4(\frac{2}{4}\log\frac{2}{4} - \frac{2}{4}\log\frac{2}{4})}{14}$$

$$H(Y|X_4) = \frac{4\times 0.24 + 6\times 0.27 + 0.3\times 4}{14} = 0.27$$

پس X_2 چون از همه $H(Y|X_i)$ کوچک‌تری دارد را انتخاب می‌کنیم.

طول بدنه ایمیل (X_2)

کوتاه — متوسط — طولانی

H:2 S:3 | H:3 S:2 | H:0 S:3

حال X_1 و X_3 و X_4 را داریم. باید information gain هر سه را با ۳ متغیر محاسبه کنیم.

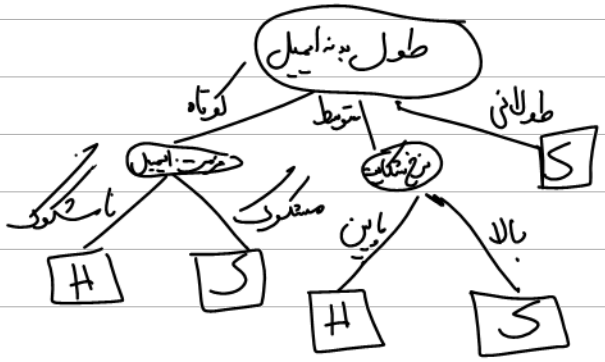
اگر کمی دقت کنیم مسیر سمت راست کامل آموزش بالا دارد. پس نمی‌توانیم با هیچ کدام از feature ها به $I(X;Y)=0$ برسیم. اما در دو مسیر دیگر اگر دقت کنیم برای مسیر طول متوسط X_1 و برای طول کوتاه X_3 به صورت دو به دو حدس می‌کنیم.

طول متوسط: پایین H:3 S:0 | بالا H:0 S:2

طول کوتاه: X_3 شلوغ H:0 S:3 | ناشلوغ H:2 S:0

می‌بینیم که این دو feature در این دو مسیر به گونه‌ای داده‌ها را حدس می‌کنند که $I(X;Y)$ ماکسیم می‌شود زیرا به ازای $H:n, S:0$ یا $H:0, S:n$ تبدیل می‌شود که می‌دانیم $H(Y|X_i)$ در این حالات ۰ است؟

درخت تصمیم نهایی که سعی بر max شدن information gain می‌شود:



نیازی به عمق بیشتر نداریم زیرا همه داده‌ها دسته‌بندی شدند.

مسئله دوم

حال داده های مجموعه را با توجه به درخت تصمیم پیش‌بینی می‌کنیم.

شماره	نرخ شکایت از فرستنده	طول بدنه ایمیل	فرمت ایمیل	دامنه ایمیل	تشخیص
۱	پایین	کوتاه	مشکوک	academic	نامشکوک
۲	بالا	کوتاه	نامشکوک	gmail	مشکوک
۳	بالا	کوتاه	مشکوک	گمنام	مشکوک
٤	بالا	متوسط	مشکوک	گمنام	مشکوک
٥	پایین	متوسط	مشکوک	academic	نامشکوک
٦	بالا	طولانی	نامشکوک	gmail	مشکوک

جدول ۱-۲

داده اول: طول بدنه ایمیل = کوتاه، فرمت ایمیل = مشکوک ← Spam

داده دوم: طول بدنه ایمیل = کوتاه، فرمت ایمیل = نامشکوک ← Ham

داده سوم: طول بدنه ایمیل = کوتاه، فرمت ایمیل = مشکوک ← Spam

داده چهارم: طول بدنه ایمیل = متوسط، نرخ شکایت = بالا ← Spam

داده پنجم: طول بدنه ایمیل = متوسط، نرخ شکایت = پایین ← Ham

داده ششم: طول بدنه ایمیل = طولانی ← Spam