4/19/2019 Discussion # 3

Entropy: log (base 2)

Definition:

If a random variable (RV) X has k different values, $a_1, a_2, \ldots, a_k$, it is entropy is given by

$$H[X] = -\sum_{k=1}^{k} p(X=a_k) \log p(X=a_k)$$

Example:

aside: $-\log a = \log \frac{1}{a}$

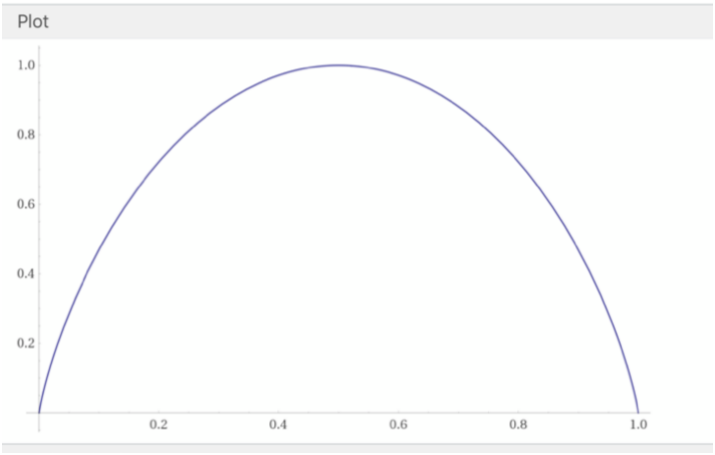
$k=2$. $p(X=0) = p$

$p(X=1) = 1-p$.

$$H(X) = \mathbb{E}\left[\log \frac{1}{p(x)}\right] = p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p} = H_2(p)$$

$$= H_b(p)$$

they are the same thing, just different notation.



Conditional entropy:

Definition:

Given two RVs X and Y

$$H[Y|X] = \sum_k p(X=a_k)\, H[Y|X=a_k]$$

Information Gain: (Mutual information)

$$\text{Gain} = H[Y] - H[Y|X] = I(X;Y)$$

properties:

$$0 \le H(x) \le \log |K|$$

K is a set

$|K|$: # of elements in set K. (Cardinality).

$$I(X;Y) = H(y) - H(y|x) = H(x) - H(x|y)$$

$$= I(Y;X)$$

$$0 \le H(Y|X) \le H(Y)$$

when $H(Y|x) = H(Y)$ $\quad X \perp\!\!\!\perp Y$

$$H(x) \ge H(x|Y)$$

Note: $H(x) \ge H(x|Y)$ didn't mean $H(x) \ge H(x|Y=y)$

Decision Tree algorithm:

Iterative Dichotomiser 3: (ID3):

Select the feature with largest information gain.

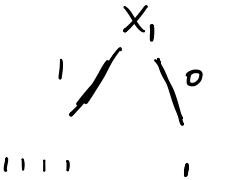
pb 1:

a) $p(Y=1) = \frac{5}{8}$ $\quad p(Y=0) = \frac{3}{8}$

$$H_b(\tfrac{3}{8}) = -\frac{5}{8}\log\frac{5}{8} - \frac{3}{8}\log\frac{3}{8} = 0.9544.$$

b) $H(Y|X_1)$

$p(X_1=0) = \frac{1}{2}$, $p(X_1=1) = \frac{1}{2}$



$p(Y=1|X=1) = 1$

$p(Y=0|X=1) = 0$

$p(Y=1|X=0) = \frac{1}{4}$

$p(Y=0|X=0) = \frac{3}{4}$

$$H(Y|X_1) = p(X_1=0)H(Y|X_1=0) + p(X_1=1)H(Y|X_1=1)$$

$$= \frac{1}{2}H_b(\tfrac{3}{4}) + \frac{1}{2}H_b(1)$$

$$\approx 0.4056.$$

(c) $H(Y|X_2)$, $H(Y|X_3)$, $H(Y|X_4)$

$H(Y|X_2) = \frac{2}{8}H_b(1) + \frac{6}{8}H_b(\frac{1}{2}) = 0.75$

$H(Y|X_3) = \frac{1}{2}H_b(\frac{3}{4}) + \frac{1}{2}H_b(\frac{1}{2}) = 0.9056$

$H(Y|X_4) = \frac{2}{8}H_b(1) + \frac{6}{8}H_b(\frac{1}{2}) = 0.75$

(d) $I(Y; X_1) = 0.9544 - 0.4056 = 0.5488$

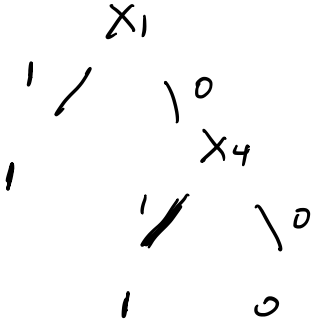
$I(Y; X_2) = 0.2044$

$I(Y; X_3) = 0.1407$

$I(Y; X_4) = 0.2044$

(e) after asking X_1:

sample:	X_2	X_3	X_4	Y
1	0	0	1	1
2	0	0	0	0
5	0	1	0	0
6	0	1	0	0



pb (2)

(a) $Var(V) = \sum_{x_i \in V} (x_i - u)^2$

$$\frac{\partial}{\partial u} Var(V) = -2 \sum_{x_i \in V} (x_i - u) = 0.$$

$$\sum_{x_i \in V} (x_i - u) = 0.$$

$$\sum_{x_i \in V} x_i - \sum_{x_i \in V} u = 0$$

$$|V| u = \sum_{x_i \in V} x_i$$

$$u = \frac{\sum_{x_i \in V} x_i}{|V|}$$

pb (3)

(a)

$\underline{\underline{C}} = \underline{\underline{A}}\,\underline{\underline{Z}}$

$$C_{11} = \sum_{j=1}^{n} a_{1j} z_{j1}$$

$$C_{22} = \sum_{j=1}^{n} a_{2j} z_{j2}$$

$$\vdots$$

$$C_{nn} = \sum_{j=1}^{n} a_{nj} z_{j2}$$

change of variable.

$$Tr(\underline{\underline{C}}) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} z_{ji} = \sum_{r=1}^{n}\sum_{s=1}^{n} a_{rs} z_{sr}$$

only exist when $s=i,\ r=j$

$$\frac{d\,Tr(\underline{\underline{A}}\,\underline{\underline{Z}})}{d z_{ij}} = \frac{d\left(\sum_{r=1}^{n}\sum_{s=1}^{n} a_{rs} z_{sr}\right)}{d z_{ij}} = a_{ji}$$

$$\therefore \frac{d\,Tr(\underline{\underline{A}}\,\underline{\underline{Z}})}{d\,\underline{\underline{Z}}} = A^T$$

(b) $$\frac{\partial \operatorname{Tr}(\underline{\underline{Z}}\,\underline{\underline{A}}\,\underline{\underline{Z}}^T)}{\partial \underline{\underline{Z}}} = \underline{\underline{Z}}\,\underline{\underline{A}}^T + \underline{\underline{Z}}\,\underline{\underline{A}}$$

Let $\underline{\underline{C}} = \underline{\underline{A}}\,\underline{\underline{Z}}^T \qquad D = ZAZ^T = \underline{\underline{Z}}\,\underline{\underline{C}}$

$$C_{11} = \sum_{t=1}^{n} a_{1t} z_{1t}$$

$$C_{12} = \sum_{t=1}^{n} a_{1t} z_{2t}$$

$$C_{21} = \sum_{t=1}^{n} a_{2t} z_{1t}$$

$$\vdots$$

$$C_{rs} = \sum_{t=1}^{n} a_{rt} z_{st}$$

$$d_{11} = \sum_{s=1}^{n} z_{1s} C_{s1} = \sum_{s=1}^{n} z_{1s} \sum_{t=1}^{n} a_{st} z_{1t}$$

$$= \sum_{s=1}^{n} \sum_{t=1}^{n} z_{1s} a_{st} z_{1t}$$

$$d_{22} = \sum_{s=1}^{n} \sum_{t=1}^{n} z_{2s} a_{st} z_{2t}$$

$$\operatorname{Tr}(D) = \sum_{r=1}^{n} \sum_{s=1}^{n} \sum_{t=1}^{n} z_{rs} a_{st} z_{rt}$$

($t=s$, $t \neq s$)

$$= \sum_{r=1}^{n} \sum_{s=1}^{n} z_{rs}^2 a_{ss} + \sum_{r}^{n} \sum_{s}^{n} \sum_{t,t\neq s}^{n} z_{rs} a_{st} z_{rt}$$

$$\therefore \operatorname{Tr}(D) = \sum_{r=1}^{n} \sum_{s=1}^{n} z_{rs}^2 a_{ss} + \sum_{r}^{n} \sum_{s}^{n} \sum_{t,t\neq s}^{n} z_{rs} a_{st} z_{rt}$$

$$Tr(D) = \sum_{r=1}^{n}\sum_{s=1}^{n} z_{rs}^2 a_{ss} + \sum_{r}^{n}\sum_{s}^{n}\sum_{t,t\neq s}^{n} z_{rs} a_{st} z_{rt}$$

Step 1:

$$\frac{d\,Tr(D)}{z_{ii}} = \frac{d}{dz_{ii}}\left(\underbrace{\sum_{r=1}^{n}\sum_{s=1}^{n} z_{rs}^2 a_{ss}}_{r=i,\ s=i} + \underbrace{\sum_{r}^{n}\sum_{s}^{n}\sum_{t,t\neq s}^{n} z_{rs} a_{st} z_{rt}}_{r=i,\ s=i,\ t\neq i;\ \ r=i,\ t=j,\ s\neq i}\right)$$

$$= 2 z_{ii} a_{ii} + \sum_{t=1,t\neq i}^{n} a_{it} z_{it} + \sum_{s=1,s\neq i}^{n} z_{is} a_{si}$$

↓ change of variable.

$$= 2 z_{ii} a_{ii} + \sum_{t=1,t\neq i}^{n} a_{it} z_{it} + \sum_{t=1,t\neq i}^{n} z_{it} a_{ti}$$

$$= z_{ii} a_{ii} + z_{ii} a_{ii} + \sum_{t=1,t\neq i}^{n} a_{it} z_{it} + \sum_{t=1,t\neq i}^{n} z_{it} a_{ti}$$

$$= \sum_{t=1}^{n} a_{it} z_{it} + \sum_{t=1}^{n} a_{ti} z_{it}$$

$$= \sum_{t=1}^{n} z_{it} (a_{it} + a_{ti})$$

$$\therefore \frac{d\,Tr(D)}{z_{ii}} = \sum_{t=1}^{n} z_{it} (a_{it} + a_{ti})$$

↑ diagnol term.

$$Tr(D) = \sum_{r=1}^{n}\sum_{s=1}^{n} z_{rs}^2 a_{ss} + \sum_{r}^{n}\sum_{s}^{n}\sum_{t,t\neq s}^{n} z_{rs}\, a_{st}\, z_{rt}$$

Step 2:

$$\frac{d\,Tr(D)}{z_{ij}} = \frac{d}{dz_{ij}}\left(\underbrace{\sum_{r=1}^{n}\sum_{s=1}^{n} z_{rs}^2 a_{ss}}_{r=i,\ s=j} + \underbrace{\sum_{r}^{n}\sum_{s}^{n}\sum_{t,t\neq s}^{n} z_{rs}\, a_{st}\, z_{rt}}_{r=i,\ s=j,\ t\neq j \quad\text{or}\quad r=i,\ t=j,\ s\neq j}\right)$$

$$= 2z_{ij}\,a_{jj} + \sum_{t,t\neq j}^{n} a_{jt} z_{it} + \sum_{s,\,s\neq j}^{n} z_{is}\, a_{sj}$$

$$= 2z_{ij}\,a_{jj} + \sum_{t,t\neq i}^{n} a_{jt} z_{it} + \sum_{t,t\neq j}^{n} z_{it}\, a_{tj}$$

$$= \sum_{t=1}^{n} a_{jt} z_{it} + \sum_{t=1}^{n} z_{it}\, a_{tj}$$

$$= \sum_{t=1}^{n} z_{it}\,(a_{jt} + a_{tj})$$

$$\therefore \frac{d\,Tr(D)}{z_{ij}} = \sum_{t=1}^{n} z_{it}\,(a_{jt} + a_{tj})$$

Combine step 1 and 2:

$$\frac{d\,Tr(D)}{z_{ii}} = \sum_{t=1}^{n} z_{it}\,(a_{it} + a_{ti})$$

Combine step 1 and 2:

$$\frac{\partial \mathrm{Tr}(D)}{z_{ii}} = \sum_{t=1}^{n} z_{it}(a_{it} + a_{ti})$$

$$\frac{\partial \mathrm{Tr}(D)}{z_{ij}} = \sum_{t=1}^{n} z_{it}(a_{jt} + a_{tj})$$

$$\frac{\partial \mathrm{Tr}(D)}{\underline{\underline{z}}} = \underline{\underline{z}}(A + A^T) = zA + zA^T$$

(c) Aside: $\mathrm{Tr}(A+B) = \mathrm{Tr}(A) + \mathrm{Tr}(B)$

$\mathrm{Tr}(AB) = \mathrm{Tr}[(AB)^T]$

$$E_r(w) = \mathrm{Tr}\left[(\underline{\underline{X}}\,\underline{\underline{w}} - \underline{\underline{T}})^T(\underline{\underline{X}}\,\underline{\underline{w}} - \underline{\underline{T}})\right]$$

$$= \mathrm{Tr}\left[\underline{\underline{w}}^T\underline{\underline{X}}^T\underline{\underline{X}}\,\underline{\underline{w}} - \underbrace{\underline{\underline{T}}^T\underline{\underline{X}}\,\underline{\underline{w}}}_{C} - \underbrace{\underline{\underline{w}}^T\underline{\underline{X}}^T\underline{\underline{T}}}_{C^T} + \underline{\underline{T}}^T\underline{\underline{T}}\right]$$

$$= \mathrm{Tr}\left[\underline{\underline{w}}^T\underline{\underline{X}}^T\underline{\underline{X}}\,\underline{\underline{w}}\right] - 2\mathrm{Tr}\left[\underline{\underline{T}}^T\underline{\underline{X}}\,\underline{\underline{w}}\right] + \mathrm{Tr}\left[\underline{\underline{T}}^T\underline{\underline{T}}\right]$$

(part A) (part B)

$$\nabla_w E_r(w) = (X^TX + X^TX)w - 2X^TT = 0.$$

$$\Rightarrow (\underline{\underline{X}}^T\underline{\underline{X}})\underline{\underline{W}} = \underline{\underline{X}}^T\underline{\underline{T}}$$

$$\underline{\underline{W}} = (\underline{\underline{X}}^T\underline{\underline{X}})^{-1}\underline{\underline{X}}^T\underline{\underline{T}}$$