COL774
Machine Learning
Minor Solutions
Semester I, 2020-21

Q.4

(a)

$$\theta^{ML}_{j=l|k} = \frac{\sum_{i=1}^{m} 1\{y^{(i)}=k\} 1\{x_j^{(i)} = l\}}{\sum_{i=1}^{m} 1\{y^{(i)}=k\}} \quad \forall j,k,l$$

(ML estimate)

(b)

$$\theta^{ML-S}_{j=l|k} = \frac{\left[\sum_{i=1}^{m} 1\{y^{(i)}=1\} 1\{x_j^{(i)}=l\}\right] + 1}{\left[\sum_{i=1}^{m} 1\{y^{(i)}=1\}\right] + |V|}$$

(Smoothed estimate)

Justification- When we do not see any documents in class k, with attributes value $x_j = l$, then estimate above helps avoid giving 0 probabilities to test documents (of class k label) which have an occurrence of word w_l i.e. $x_j = l$. This estimate is generally useful with low data. We are effectively saying that prior probability of seeing each word in a class is $\frac{1}{|V|}$ (uniform). Effectively we are adding a document with single occurrence of every word (for each class) to the training mix.

(c) Since we ~~only~~ have ~~two~~ $|V| = 2$, we only have two parameters $\theta_{j=e|k}$ for $e = 1$ & $e2$.

Further $\theta_{j=1|k} = 1 - \theta_{j=2|k}$ $\forall k$

Let $\theta_{j=1|k}$ ~~θ~~ ~~(simplifying the notation)~~ ~~(simplifying the notation)~~

$P(\theta_{j=1|k}) = (\theta_{j=1|k})(1-\theta_{j=1|k})$

Then $P(\cancel{\theta})$ ~~$= \theta(1-\theta)$~~

prior probability

Now, MAP estimate $= \theta MAP$

$= \arg\max_{\theta} P(\theta | D)$

$P(\theta|D) \propto P(D|\theta)\,P(\theta)$

$P(\cancel{D|\theta})$

~~Data corresponds data of documents with class~~

We are interested in-

$P(\theta_{j=1|k}) \propto P(D|\theta_{j=1|k})\,P(\theta_{j=1|k})$

Now, let us estimate

$P(D|\theta_{j=1|k})$

② First note that ~~we are~~ since we are interested in maximizing:

$$\arg\max_{\theta_j=1|k} P(D|\theta_j=1|k)\, P(\theta_j=1|k)$$

Only those components of data which depend on ~~occurrence~~ $\theta_j = 1|k$ will matter (rest will come out as a constant)

$$\equiv \cancel{\arg\max_{\theta_j=1|k} K \cdot \prod_{i=1}^{m} P(x_j^{(i)}=1 \mid \theta_j=1|k) \cdot 1\{x_j^{(i)}=1\} \cdot 1\{y^{(i)}=1\}}$$

$\cancel{P(\theta_j=1|k)}$

other terms

Taking log inside argmax:-

$$\arg\max_{\theta_j=1|k} \log P(D|\theta_j=1|k) + \log P(\theta_j=1|k)$$

$$\arg\max_{\theta_j=1|k} \sum_{i=1}^{m} \left(\log P(x_j^{(i)}=1 \mid \theta_j=1|k)\right) 1\{y^{(i)}=k\} 1\{x_j^{(i)}=1\}$$

$$+ \log\left[1 - P(x_j^{(i)}=0 \mid \theta_j=1|k)\right] \cdot 1\{y^{(i)}=k\} 1\{x_j^{(i)}=\ldots$$

$$+ (\log K) + \log P(\theta_j=1|k)$$

Terms which do not depend on $\theta_j = 1|k$

$$\Rightarrow \underset{\theta_{j=1|k}}{\arg\max} = \sum_{i=1}^{m} 1\{y^{(i)}=k\}\, 1\{x_j^{(i)}=1\} \log[\theta_{j=1|k}]$$

$$+ \sum_{i=1}^{m} 1\{y^{(i)}=k\}\, 1\{x_j^{(i)}=0\} \log(1-\theta_{j=1|k})$$

$$+ \log(\theta_{j=1|k})(1-\theta_{j=1|k})$$

$$+ \log k \equiv f(\theta_{j=1|k})$$

We want to maximize $f(\theta_{j=1|k})$ wrt $\theta_{j=1|k} \Rightarrow$

$$\cancel{\nabla_{\theta_{j=1|k}}} f(\theta_{j=1|k}) \cancel{= \sum_{i=1}^{m} \frac{1\{y^{(i)}=k\}1\{x_j^{(i)}=1\}}{\theta_{j=1|k}}}$$

$$\cancel{+ \sum_{i=1}^{m} (-1) \frac{1\{y^{(i)}=k\}1\{x_j^{(i)}=0\}}{1-\theta_{j=1|k}}}$$

$$= \left[\sum_{i=1}^{m} 1\{y^{(i)}=k\}\, 1\{x_j^{(i)}=1\}\right] \log[(\theta_{j=1|k})$$

$$+ \log \cancel{p(\theta_{j=1|k})} (\theta_{j=1|k})$$

$$+ \left[\sum_{i=1}^{m} 1\{y^{(i)}=k\}\, 1\{x_j^{(i)}=0\}\right] \log(1-\theta_{j=1|k})$$

$$+ \log(1-\theta_{j=1|k})$$

③ Combining terms we get:-

$$\cancel{\nabla_{\theta_{j=1|k}} f(\theta}$$

$$f(\theta_{j=1|k}) = \overbrace{\left[\sum_{i=1}^{m} 1\{y^{(i)}=k\} 1\{x_j^{(i)}=1\} + 1\right]}^{\alpha_1} \log \theta_{j=1|k}$$

$$+ \overbrace{\left[\sum_{i=1}^{m} 1\{y^{(i)}=k\} 1\{x_j^{(i)}=0\} + 1\right]}^{\alpha_0} \log (1-\theta_{j=1|k})$$

$$= \alpha_1 \log \theta_{j=1|k} + \alpha_0 \log (1-\theta_{j=1|k})$$

$$\Rightarrow \nabla_{\theta_{j=1|k}} f(\theta_{j=1|k})$$

$$= \frac{\alpha_1}{\theta_{j=1|k}} + \frac{\alpha_0}{1-\theta_{j=1|k}} (-1)$$

Equating to zero, we get

$$\cancel{\alpha_1 (1-\theta_{j=1|k}}$$

$$\frac{\alpha_1}{\theta_{j=1|k}} = \frac{\alpha_0}{1-\theta_{j=1|k}}$$

$$\Rightarrow \alpha_1 (1-\theta_{j=1|k}) = \alpha_0 (\theta_{j=1|k})$$

$$\Rightarrow \alpha_1 = (\alpha_1 + \alpha_0)(\theta_{j=1|k})$$

$$\Rightarrow \theta_{j=1|k} = \frac{\alpha_1}{\alpha_1 + \alpha_0}$$ → substituting for α_1 & α_2

$$= \frac{\left[\sum_{i=1}^{m} 1\{y^{(i)}=k\} 1\{x_j^{(i)}=1\}\right] + 1}{\left[\sum_{i=1}^{m} 1\{y^{(i)}=k\} 1\{x_j^{(i)}=1\} + \sum_{i=1}^{m} 1\{y^{(i)}=k\} 1\{x_j^{(i)}=0\}\right] + 1 + 1}$$

$\sum_{i=1}^{m} 1\{y^{(i)}=k\} \Leftarrow$

$$= \frac{\sum_{i=1}^{m} 1\{y^{(i)}=k\} 1\{x_j^{(i)}=1\} + 1}{\sum_{i=1}^{m} 1\{y^{(i)}=k\} + 2 \quad (=|V|)}$$

which is exactly the smoothed estimate.

Hence, proved

COL774

Machine Learning

Minor Solutions

Semester I, 2020-21

Q.5.

Let $X1$:- whether Person 1 is corona +ve or not

$X_1 = 1$ P1 corona +ve

0 o.w.

$X2$:- whether Person 2 is corona +ve or not

$X_2 = 1$ if P2 is corona +ve

0 o.w.

$T1$:- whether P1 tests +ve for corona

$T1 := 1$ P1 tests +ve

0 o.w.

$T2$:- whether P2 tests +ve for corona

1 if P2 tests +ve

0 o.w.

$$P(X_1 = 1, X_2 = 1 \mid T_1 = 1, T_2 = 1)$$

$$= \frac{P(T_1 = 1, T_2 = 1 \mid X_1 = 1, X_2 = 1)\; P(X_1 = 1, X_2 = 1)}{P(T_1 = 1, T_2 = 1)}$$

$$= \frac{\underbrace{P(T_1 = 1 \mid X_1 = 1) \cdot P(T_2 = 1 \mid X_2 = 1)}_{\text{from indepd}} \; \underbrace{P(X_1 = 1)\, P(X_2 = 1)}_{\text{from indepdc}}}{\sum_{X_1, X_2} P(T_1 = 1, T_2 = 1 \mid X_1, X_2)}$$

Denom:- $$\sum_{X_1} P(T_1 = 1 \mid X_1) \cdot P(X_1) \quad \sum_{X_2} P(T_2 = 1 \mid X_2) \cdot P(X_2)$$

$$= \left[\frac{P(T_1=1|X_1=1)\,P(X_1=1)}{\left\{ P(T_1=1|X_1=1)\,P(X_1=1) + P(T_1=1|X_1=0)\,P(X_1=0) \right\}} \right] \left[\frac{P(T_2=1|X_2=1)\,P(X_2=1)}{P(T_2=1|X_2=1)\,P(X_2=1) + P(T_2=1|X_2=0)\,P(X_2=0)} \right]$$

A B

A & B are identical.

$$A = \frac{0.98 \times 0.01}{0.98 \times .01 + 0.03 * 0.99}$$

$$= \frac{98 \times 10^{-4}}{98 \times 10^{-4} + 297 \times 10^{-4}}$$

$$= \frac{98}{98+297} = \frac{98}{395} = \overline{B}$$

$$\Rightarrow P(X_1=1, X_2=1 | T_1=1, T_2=1) = \left(\frac{98}{395}\right)^2$$

(b) i $P(E, B, A, T) = P(E)\,P(B)\,P(A|E,B)\,P(T|A)$

parameter (independent)

$= 1 + 1 + 4 + 2 = \overline{8}$

(Needs Justification)

ii $E \not\perp B | T$:- Explaining Away. if T is known (say true) then knowing that Earthquake happened explains the ~~phenomenon of~~ Tina calling (effect ~~flowing thing~~ Alarm) probability of B will go down in this case, for example.

[Also does not follow from network structure]