Natural Language Processing

Homework-1

gpande@gatech.edu

2. The entropy of a discrete random variable X is defined as (use base e for all log operations unless specified otherwise):

$$H(X) = -\sum_{x \in X} P(x) \log P(x)$$

(a) Compute the entropy of the distribution $P(x) = \text{Multinomial}([0.2, 0.3, 0.5])$. [3 pts]

(b) Compute the entropy of the uniform distribution $P(x) = \frac{1}{m} \forall x \in [1, m]$. [3 pts]

(c) Consider the entropy of the joint distribution P(X, Y):

$$H(X,Y) = -\sum_{x \in X} \sum_{y \in Y} P(x,y) \log P(x,y)$$

How does this entropy relate to H(X) and H(Y), (i.e. the entropies of the marginal distributions) when X and Y are independent? [4 pts]

Solution: Next Page

2. $H(X) = -\sum_{x \in X} P(x) \log P(x)$

(a): $P(x) =$ Multinoulli $([0.2, 0.3, 0.5])$

Entropy $\Rightarrow$ $H(x) = -[0.2 \log 0.2,\ 0.3 \log 0.3,\ 0.5 \log 0.5]$

$= [0.32,\ 0.36,\ 0.34]$

(b): Entropy of uniform distribution

$P(x) = \frac{1}{m} \ \forall x \in [1, m]$

$H(x) = \int_1^m - P(x) \log P(x)$

$= -\int_1^m \frac{1}{m} \log m \, dm$

$\log m = u$

$dm = du$

$\frac{1}{m} dm = du$

$= -\int_1^m u \, du$

$= -\int_0^{e^u} u \, du = -\left(\frac{u^2}{2}\right)_0^{e^u}$

$= -\left[\frac{(e^u)^2}{2} - 0\right]$

$= \frac{(e^u)^2}{2} = \frac{m^2}{2}$

(c): Entropy of Joint Prob. distribution. $P(X,Y)$

$H(X,Y) = -\sum\sum P(x,y) \log P(x,y)$ — (i)

when X, and Y are independent than:

$H(Y|X) = H(Y)$ — (1)

$H(X|Y) = H(X)$

we know that:

$H(X,Y) = H(X) + H(Y|X)$

from equation (1)

$H(X,Y) = H(X) + H(Y)$

3. You are investigating articles from the New York Times and from Buzzfeed. Some of the articles contain *fake* news, while others contain *real* news (assume that there are only two types of news).
Note: for the following questions, write your answer using up to 3 significant figures.

(a) Fake news only accounts for 5% of all articles in all newspapers. However, it is known that 30% of all fake news comes from Buzzfeed. In addition, Buzzfeed generates 25% of all news articles. What is the probability that a randomly chosen Buzzfeed article is fake news? [3 pts]

(b) Suppose that 15% of all fake news comes from the New York Times (NYT). Furthermore, suppose that 60% of all real news comes from the NYT. Under all assumptions so far, what is the probability that a randomly chosen NYT article is fake news? [3 pts]

(c) Mike is an active reader of the New York Times: Mike reads 80% of all NYT articles. However, he also has a suspicion that the NYT is a bad publisher, and he believes that 25% of all NYT articles are fake news. Furthermore, the NYT generates 30% of all news articles. Under all assumptions so far, what is the probability that a randomly chosen article (from all newspapers) will be from the NYT, will be read by Mike and will be *believed* to be fake news? [4 pts]

………..

Solution: Next page

(3): (a) fake news = 5%

30% comes from Buzzfeed = P(fake, Buzzfeed)

25% news comes from Buzzfeed. = P(Buzzfeed)

$P(\text{fake news} \mid \text{Buzzfeed article}) = \dfrac{P(\text{fake news, Buzzfeed})}{P(\text{Buzzfeed article})}$

$= \dfrac{0.05 \times 0.3}{0.25} = 0.3 \times \dfrac{50}{250}$

$= 0.06$

(b) 15% of fake News comes from NYT.

60% of real " " " " .

$P(\text{fake News} \mid \text{NYT}) = \dfrac{P(\text{fake News, NYT})}{P(\text{NYT})}$

$= \dfrac{5 \times 15/100}{95 \times \frac{60}{100}} = \dfrac{0.75}{57.75}$

$= 0.013$

(c) $P(\text{Mike read} \mid \text{NYT}) = 0.8$

$P(\text{NYT}) = 0.3$

$P(\text{Believed fake} \mid \text{NYT}) = 0.25$

P(NYT and Read and Believed false)

$= P(\text{read news} \mid \text{fake, NYT}) * P(\text{fake news} \mid \text{NYT}) * P($

$= 0.8 * 0.25 * 0.3$

$= 0.06$

4. Suppose we have a probability density function (pdf) defined as:

$$f(x,y) = \begin{cases} C(x^2+2y), & 0 < x < 1 \text{ and } 0 < y < 1, \\ 0, & \text{otherwise.} \end{cases}$$

(a) Find the value of C. [2pts]

(b) Find the marginal distribution of X and Y. [4pts]

(c) Find the joint cumulative density function (cdf) of X and Y. [4pts]

Solution:

$$f(x,y) = \begin{cases} c(x^2+2y), & 0<x<1,\ 0<y<1 \\ 0, & \text{otherwise} \end{cases}$$

(a) value of c.

Summation of all the Prob. accsoss PDF = 1

$$\int_0^1\int_0^1 c(x^2+2y)\,dx\,dy = 1$$

$$\int_0^1\int_0^1 c(x^2\,dx\,dy + 2y\,dx\,dy) = 1$$

$$\int_0^1 c\left[\frac{x^3}{3}\right]_0^1 dy + 2c\,[x]_0^1\, y\,dy = 1$$

$$\int_0^1 \frac{c}{3}\,dy + 2c\, y\,dy = 1$$

$$\frac{c}{3}[y]_0^1 + 2c\left[\frac{y^2}{2}\right]_0^1 = 1$$

$$\frac{c}{3} + \frac{2c}{2} = 1$$

$$\frac{4c}{3} = 1$$

$$c = \frac{3}{4}$$

(b): Marginal distribution of X and Y.

$$f_X(x) = \int_0^1 C(x^2+2y)\,dy$$

$$= \frac{3}{4}x^2[y]_0^1 + \frac{3}{4}\times 2\times\left[\frac{y^2}{2}\right]_0^1$$

$$= \frac{3x^2}{4} + \frac{6}{4}\times\frac{1}{2}\times 2$$

$$= \frac{3}{4}(x^2+1)$$

Thus,

$$f_x(x) = \begin{cases} \frac{3}{4}(x^2+1), & 0<x<1 \\ 0, & \text{otherwise} \end{cases}$$

$$f_Y(Y) = \int_0^1 \frac{3}{4}(x^2+2y)\,dx$$

$$= \frac{3}{4}\left[\frac{x^3}{3}\right]_0^1 + \frac{3}{4}\times 2\,[x]_0^1\, y$$

$$= \frac{1}{4} + \frac{6}{4}y$$

$$= \frac{1}{4}(6y+1)$$

Thus,

$$f_y(Y) = \begin{cases} \frac{1}{4}(6y+1), & 0 \leq y \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

(c) Joint cumulative density function (cdf) of X and Y.

Joint cumulative function of 2 random variable is defined as:

$$F_{XY}(x,y) = P(X \le x, Y \le y)$$

$$F_{XY}(x,y) = \int_0^y \int_0^x \frac{3}{4}(x^2+2y)\,dx\,dy$$

$$= \int_0^y \frac{3}{4}\left[\frac{x^3}{3}\right]_0^x dy + \frac{3}{4} \times 2\,[x]_0^x\, y\,dy$$

$$= \int_0^y \frac{1}{4} \times x^3\,dy + \frac{6}{4} \times x\,y\,dy$$

$$= \frac{x^3}{4}[y]_0^y + \frac{6}{4}x\left[\frac{y^2}{2}\right]_0^y$$

$$= \frac{x^3 y}{4} + \frac{3}{4}xy^2$$

Thus,

$$f_{XY}(x,y) = \begin{cases} \frac{1}{4}(3xy^2 + x^3y), & 0 \le x \le 1, \; 0 \le y \le 1 \\ 0, & \text{otherwise} \end{cases}$$

5. **[Graduate Students Only]** A 2-D Gaussian distribution is defined as:

$$G(x,y) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{x^2+y^2}{2\sigma^2}\right)$$

Compute the following integral:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} G(x,y)\,(5x^2y^2 + 3xy + 1)\,dx\,dy$$

Hint: Think in terms of the properties of probability distribution functions. [5 pts]

Solution:

(5): 2-D Gaussian Distribution

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-\left(\frac{x^2+y^2}{2\sigma^2}\right)}$$

Compute:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} G(x,y)\,(5x^2y^2 + 3xy + 1)\,dx\,dy$$

for a standard 2-D gaussian distribution:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} G(x,y)\,dx\,dy = 1 \quad -(1)$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} G(x,y)\,x\,y\,dx\,dy = 0 = E(X,Y) \quad -(2)$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} G(x,y)\,x^2y^2\,dx\,dy = E(x^2, y^2) = Var(x^2 y^2) + [E(x)]$$

$$= \sigma^4$$

Hence

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} 5G(x,y)\,x^2y^2\,dx\,dy + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} 3\,G(x,y)\,xy\,dx\,dy + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} G(x,y)\,dx\,dy$$

$$= 5\sigma^4 + 0 + 1$$

$$= 5\sigma^4 + 1$$