Homework review

CSE 5243 SP20

Homework #1

Problem 1. (20 points)

Language modeling is the problem of modeling the joint probability of any natural language (e.g., English) utterance. For example, for the English utterance

$$the\ dog\ is\ chasing\ a\ cat$$

it aims to estimate the joint probability

$$P(the, dog, is, chasing, a, cat)$$

A common simplification strategy is to convert the joint probability into a product of conditional probabilities using chain rule. Please show how to do that.

Solution:

Assume the->A, dog->B, is->C, chasing->D, a->E, cat->F

P(the, dog, is, chasing, a, cat) =P(A,B,C,D,E,F)

$$=\frac{P(A,B,C,D,E,F)}{P(A,B,C,D,E)}\frac{P(A,B,C,D,E)}{P(A,B,C,D)}\frac{P(A,B,C,D)}{P(A,B,C)}\frac{P(A,B,C)}{P(A,B)}\frac{P(A,B)}{P(A)}P(A)$$

$$=P(F|A,B,C,D,E)\ P(E|A,B,C,D)\ P(D|A,B,C)\ P(C|A,B)\ P(A|B)\ P(A)$$

Problem 2. (20 points)

Let X be a random variable denoting age. Consider a random sample of size n = 20. $X = (69, 74, 68, 70, 72, 67, 66, 70, 76, 68, 72, 79, 74, 67, 66, 71, 74, 75, 75, 76)$.

a) Find the mean, median, and mode of X.

b) Let us use the normal distribution to model the random variable X. Write down its probability density function (use the sample mean and standard deviation).

Solution:

(a) sort: [66, 66, 67, 67, 68, 68, 69, 70, 70, 71, 72, 72, 74, 74, 74, 75, 75, 76, 76, 79]
mean = 71.45 median =71.5 mode =74

(b) μ = 71.45

σ= $\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x-u)^2}$ =3.8179 (Here we use unbiased sample variance to estimate population variance)

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

Problem 3. (20 points)

Similarity/distance between data points plays an important role in data analysis. However, the results can vary depending on the similarity/distance measure used, and in practice one should choose a measure that works best for the specific type of data and analysis under investigation.

Suppose we have the following 2-D data set:

	A_1	A_2
$\mathbf{x}_1$	1.3	1.6
$\mathbf{x}_2$	2.1	1.7
$\mathbf{x}_3$	1.9	1.9
$\mathbf{x}_4$	1.8	1.6
$\mathbf{x}_5$	1.5	2.0

a) Consider the data as 2-D data points. Given a new data point, $x = (1.7, 1.8)$ as a query, rank the points based on similarity (most similar/closest ones first) with the query using Euclidean distance, Manhattan distance, Jaccard similarity, and cosine similarity.

b) Normalize the data set to make the norm of each data point equal to 1. Use Euclidean distance on the transformed data to rank the data points.

(a)

Euclidean Distance:

X1: $\sqrt{(1.7-1.3)^2+(1.8-1.6)^2} = 0.447$

X2: $\sqrt{(1.7-2.1)^2+(1.8-1.7)^2} = 0.412$

X3: $\sqrt{(1.7-1.9)^2+(1.8-1.9)^2} = 0.224$

X4: $\sqrt{(1.7-1.8)^2+(1.8-1.6)^2} = 0.224$

X5: $\sqrt{(1.7-1.5)^2+(1.8-2.0)^2} = 0.283$

Rank: x3 (tie), x4 (tie), x5, x2, x1

Manhattan Distance:

X1: $|1.7-1.3|+|1.8-1.6| = 0.6$

X2: $|1.7-2.1|+|1.8-1.7| = 0.5$

X3: $|1.7-1.9|+|1.8-1.9| = 0.3$

X4: $|1.7-1.8|+|1.8-1.6| = 0.3$

X5: $|1.7-1.5|+|1.8-2.0| = 0.4$

Rank: x3 (tie), x4 (tie), x5, x2, x1

Jaccard Similarity

X1: $(1.7*1.3+1.8*1.6)/(1.7^2+1.8^2+1.3^2+1.6^2-1.7*1.3-1.8*1.6) = 0.9622$

X2: $(1.7*2.1+1.8*1.7)/(1.7^2+1.8^2+2.1^2+1.7^2-1.7*2.1-1.8*1.7) = 0.9750$

X3: $(1.7*1.9+1.8*1.9)/(1.7^2+1.8^2+1.9^2+1.9^2-1.7*1.9-1.8*1.9) = 0.9925$

X4: $(1.7*1.8+1.8*1.6)/(1.7^2+1.8^2+1.8^2+1.6^2-1.7*1.8-1.8*1.6) = 0.9917$

X5: $(1.7*1.5+1.8*2.0)/(1.7^2+1.8^2+1.5^2+2.0^2-1.7*1.5-1.8*2.0) = 0.9872$

Rank: x3, x4, x5, x2, x1

Cosine Similarity

X1: $(1.7*1.3+1.8*1.6)/((1.7^2+1.8^2)*(1.3^2+1.6^2))^{0.5} = 0.9972$

X2: $(1.7*2.1+1.8*1.7)/((1.7^2+1.8^2)*(2.1^2+1.7^2))^{0.5} = 0.9911$

X3: $(1.7*1.9+1.8*1.9)/((1.7^2+1.8^2)*(1.9^2+1.9^2))^{0.5} = 0.9996$

X4: $(1.7*1.8+1.8*1.6)/((1.7^2+1.8^2)*(1.8^2+1.6^2))^{0.5} = 0.9962$

X5: $(1.7*1.5+1.8*2.0)/((1.7^2+1.8^2)*(1.5^2+2.0^2))^{0.5} = 0.9936$

Rank: x3, x1, x4, x5, x2

(b) Normalization : $X \rightarrow \frac{X}{||X||}$

Euclidean Distance:

X1: $\sqrt{(0.687 - 0.631)^2 + (0.727 - 0.776)^2} = 0.0745$

X2: $\sqrt{(0.687 - 0.777)^2 + (0.727 - 0.629)^2} = 0.1333$

X3: $\sqrt{(0.687 - 0.707)^2 + (0.727 - 0.707)^2} = 0.0286$

X4: $\sqrt{(0.687 - 0.747)^2 + (0.727 - 0.664)^2} = 0.0873$

X5: $\sqrt{(0.687 - 0.6)^2 + (0.727 - 0.8)^2} = 0.1133$

Rank: x3, x1, x4, x5, x2

Problem 4. (30 points)

A flu is going around and it is believed that 3 in 1,000 people now have it. John just had a flu test and the result was positive. The test can accurately identify 97% of patients who have flu. If a patient doesn't have flu, 99% of the time the test result will be negative.

a) What is the probability that John has flu? Show how you get to the answer.

b) John didn't believe the test result and just had the same test one more time. The result was still positive. Now what is the new probability that John has flu?

c) What if the result of the second test was negative?

(a) Define:

flu: someone has flu $\quad\quad$ $\widetilde{flu}$: someone does not have flu

+: result is positive $\quad\quad$ -:negative

$P(flu) = \frac{3}{1000} = 0.003$ $\quad\quad$ $P(\widetilde{flu}) = 0.997$

$P(+|flu) = 0.97$ $\quad\quad$ $P(+|\widetilde{flu}) = 0.03$

$P(-|flu) = 0.99$ $\quad\quad$ $P(-|\widetilde{flu}) = 0.01$

$$P(flu|+) = \frac{P(+|flu)P(flu)}{P(+)} = \frac{P(+|flu)P(flu)}{P(+|flu)P(flu)+P(+|\widetilde{flu})P(\widetilde{flu})}$$

$$= \frac{0.97*0.003}{0.97*0.003+0.01*0.997} = 0.2259$$

(b) $P(flu|++)=\frac{P(++|flu)P(flu)}{P(++)}=\frac{P(++|flu)P(flu)}{P(++|flu)P(flu)+P(++|\widetilde{flu})P(\widetilde{flu})}$

$$=\frac{P(+|flu)^2P(flu)}{P(+|flu)^2P(flu)+P(+|\widetilde{flu})^2P(\widetilde{flu})}$$

$$=\frac{0.97^2*0.003}{0.97^2*0.003+0.01^2*0.997}=0.9659$$

Why $P(++|flu)=P(+|flu)^2$?

Recall chain rule P(AB|C)=P(A|BC)P(B|C)

$$P(++|flu)=\frac{P(++flu)}{P(flu)}=\frac{P(++flu)}{P(+flu)}\frac{P(+flu)}{P(flu)}$$

$=P(+|+flu)$P(+|flu) = $P(+|flu)^2$

(Due to + and + are independent)

$$\text{(C) } P(flu|\pm)=\frac{P(+-|flu)P(flu)}{P(+\ -)}=\frac{P(+-|flu)P(flu)}{P(+-|flu)P(flu)+P\left(+-|\widetilde{flu}\right)P(\widetilde{flu})}$$

$$=\frac{P(+|flu)P(-|flu)P(flu)}{P(+|flu)P(-|flu)P(flu)+P\left(+|\widetilde{flu}\right)P\left(-|\widetilde{flu}\right)P(\widetilde{flu})}$$

$$=\frac{0.97*0.03*0.003}{0.97*0.03*0.003+0.001*0.99*0.997}=0.0088$$

Homework #2[1]

Problem 1 (10 points)

For the following group of data:

100, 200, 400, 500, 700, 1000, 3000

a) Calculate its mean and standard deviation.

b) Normalize the above group of data by min-max normalization with min = 0 and max = 1.

c) In z-score normalization, what value should the first number 100 be transformed to? What about the last number 3000?

Solution:

(a) mean = 842.86

std = $\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x-u)^2}$ = 924.05

[1]The solution of Homework 2 is adapted from Vishal Sunder and Xiao Liu's submissions.

(b) For min-max normalization with $max_A = 1$ and $min_A = 0$,

$$v' = \frac{v - min_A}{max_A - min_A}$$

Applying this to every element, new set

$$0, 0.03448276, 0.10344828, 0.13793103, 0.20689655, 0.31034483, 1$$

(c) For z-score normalization,

$$v' = \frac{v - \mu_A}{\sigma_A}$$

With this, first element,

$$= \frac{100 - 842.86}{924.05}$$
$$= -0.804$$

And last element,

$$= \frac{3000 - 842.86}{924.05}$$
$$= 2.334$$

Problem 2 (10 points)

Given the following table,

X_1	X_2
-3	a
3	b
-4.4	a
6.0	a
-4.0	a
-12.0	b
1.2	a
16.0	b
-16.0	b
13.2	a

X_1	X_2
c_2	a
c_2	b
c_2	a
c_3	a
c_2	a
c_1	b
c_2	a
c_3	b
c_1	b
c_3	a

assuming that X_1 is discretized into three bins as follows:

$$c_1 = (-20, -5]; \; c_2 = (-5, 5]; \; c_3 = (5, 20]$$

Answer the following questions:

a) Construct the contingency table between the discretized X_1 and X_2 attributes. Include the marginal counts.

b) Compute the χ^2 statistic between them.

(a) Contingency table,

	a	b	Sum (row)
c_1	0 (1.2)	2 (0.8)	2
c_2	4 (3)	1 (2)	5
c_3	2 (1.8)	1 (1.2)	3
Sum (column)	6	4	10

(b)
$$e_{ij} = \frac{count(A = a_i) \times count(B = b_j)}{n},$$

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

$$= \frac{(0 - 1.2)^2}{1.2} + \frac{(2 - 0.8)^2}{0.8} + \frac{(4 - 3)^2}{3} + \frac{(1 - 2)^2}{2} + \frac{(2 - 1.8)^2}{1.8} + \frac{(1 - 1.2)^2}{1.2}$$

$$= 3.89$$

Problem 3 (50 points)

Assume we get some data from a car insurance company in Table 1, where there are 6 data instances representing 6 people, with 2 attributes (Age and Car) and 1 class label (Risk). Here Age is a continuous attribute. Now we will build decision trees for this data set.

Data Point	Age	Car	Risk
x_2	40	Vintage	H
x_6	25	SUV	L
x_4	45	SUV	L
x_3	20	Sports	H
x_5	40	Sports	L
x_1	45	Sports	L

Table 1: Data for Problem 3. *Age* is numeric and *Car* is categorical. *Risk* gives the class label for each point: high (H) or low (L).

a) Let us consider a multi-way split for the Car attribute (using its unique values for partition). What is the information gain if we choose the Car attribute to split the root node? (5 points)

(a)

$$
\begin{aligned}
Info(D) &= I(2,4) \\
&= -\frac{2}{6}\log_2\frac{2}{6} - \frac{4}{6}\log_2\frac{4}{6} \\
&= 0.92
\end{aligned}
$$

Car	H_i	L_i	$I(H_i, L_i)$
Vintage	1	0	0
SUV	0	2	0
Sports	1	2	0.92

$$
\begin{aligned}
Info_{car}(D) &= \frac{1}{6} * 0 + \frac{2}{6} * 0 + \frac{3}{6} * 0.92 \\
&= 0.46 \\
Gain(car) &= Info(D) - Info_{car}(D) \\
&= 0.92 - 0.46 \\
&= 0.46
\end{aligned}
$$

b) Let us consider the binary splits for the Car attribute. Using information gain as the measure, which binary split of the Car attribute is the best at the root node? (5 points)

c) Between (**a**) and (**b**), which one do you prefer for splitting the root node using the Car attribute? Hint: Consider the GainRatio measure. (5 points)

d) Now, construct an entire decision tree for the given data set, using information gain as the split point evaluation measure. You can use your calculations or conclusions in (**a-c**). (30 points)

e) Classify the point (Age=27, Car=SUV) based on the constructed decision tree in (**d**). (5 points)

(b) Let's consider the three possible splits:

(a)

Car	H_i	L_i	$I(H_i, L_i)$
{Vintage,SUV}	1	2	0.92
Sports	1	2	0.92

$$Gain = 0.92 - (\frac{3}{6} * 0.92 + \frac{3}{6} * 0.92)$$
$$= 0$$

(b)

Car	H_i	L_i	$I(H_i, L_i)$
Vintage	1	0	0
{SUV,Sports}	1	4	0.72

$$Gain = 0.92 - (\frac{1}{6} * 0 + \frac{5}{6} * 0.72)$$
$$= 0.32$$

(c)

Car	H_i	L_i	$I(H_i, L_i)$
SUV	0	2	0
{Vintage,Sports}	2	2	1.00

$$Gain = 0.92 - (\frac{2}{6} * 0 + \frac{4}{6} * 1.00)$$
$$= 0.25$$

Therefore, we choose the split $\{\{SUV, Sports\}, Vintage\}$ which has the highest gain of 0.32

(c) For (a),

$$\begin{aligned} SplitInfo_a(D) &= -\frac{1}{6} * \log_2 \frac{1}{6} - \frac{2}{6} * \log_2 \frac{2}{6} - \frac{3}{6} * \log_2 \frac{3}{6} \\ &= 1.46 \\ GainRatio_a(D) &= \frac{Gain}{SplitInfo} \\ &= \frac{0.46}{1.46} \\ &= 0.32 \end{aligned}$$

For (b),

$$\begin{aligned} SplitInfo_b(D) &= -\frac{1}{6} * \log_2 \frac{1}{6} - \frac{5}{6} * \log_2 \frac{5}{6} \\ &= 0.65 \\ GainRatio_b(D) &= \frac{Gain}{SplitInfo} \\ &= \frac{0.32}{0.65} \\ &= 0.49 \end{aligned}$$

Hence, we choose (b), as it has a higher gain ratio.

(d) For the first split, lets compare the age and car attribute.
For age,

Step 1: Sorted Values and insert possible splitting values

Step 2: calculate the frequencies of each category

Step 3: calculate information gain

	20	25	40	45
	22	27	42	
	<=	>	<=	>	<=	>	
H	1	1	1	1	2	0	
L	0	4	1	3	2	2	
Info	0.4170		0.6059		0.4621		
Gain	0.2195		0.0306		0.1744		
SplitInfo	0.4506		0.6365		0.6365		
Gainratio	**0.4871**		0.0481		0.2740		

The gainratio is equal to (c). So both ways should work.
Assume we use {{SUV,sports},vintage}as the first split.

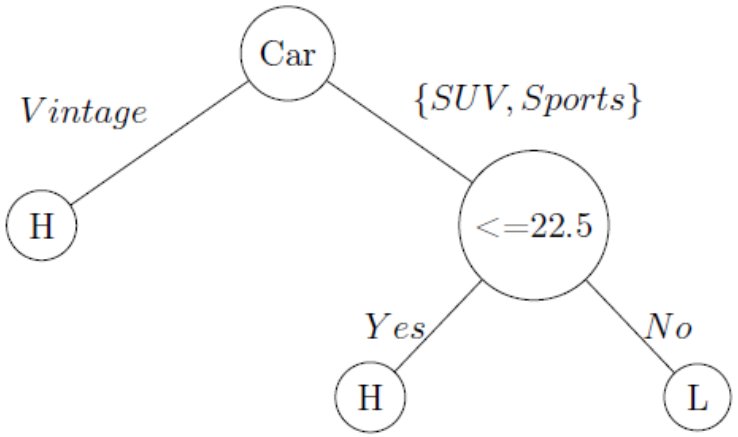
(d) For the second split, the node {vintage} only contain one class {H}. No need to split.
Lets split the node{SUV, sports}

Assuming that we split the first stage according to (c), $Info_{age}(D) = -\frac{4}{5} * log_2\frac{4}{5} - \frac{1}{5} * log_2\frac{1}{5} = 0.72$ we now split for the age which is a continuous attribute,

H	L	L	L,L
20	25	40	45

	22.5		32.5		42.5	
	<=	>	<=	>	<=	>
H	1	0	1	0	1	0
L	0	4	1	3	2	2
Gain	0.72-0=0.72		0.72-0.40=0.32		0.72-0.55=0.17	

Highest gain = 0.72 corresponding to split point 22.5.



(e) According to the decision-tree above, the point (Age=27, Car=SUV) will be classified as L.