Question Set

David Barbas Rebollo

Advanced Machine Learning
May, 2021

1 Theoretical Questions

1.1 Question 1 (1 point)

Given two distribution p and q, demonstrate if $D(p||q) + D(q||p)$ is or is not symmetric and satisfy the triangle inequality.

This formula represents the measure of the difference in probability between two distributions over the same variable x and it is known as Kullback-Leiber divergence.

$$D(p||q) + D(q||p) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

It is a non-symmetric measure, meaning:

$$D(p||q) + D(q||p) \neq D(q||p) + D(p||q)$$

For this demonstration the following distributions are proposed, p and q:

$$p(x) = \begin{cases} 1/4 & x = 0 \\ 1/2 & x = 1 \\ 1/4 & x = 2 \end{cases} \qquad q(x) = \begin{cases} 3/5 & x = 0 \\ 1/5 & x = 1 \\ 1/5 & x = 2 \end{cases}$$

$$D(p(x)||q(x)) = \frac{1}{4}\log\frac{5}{12} + \frac{1}{2}\log\frac{5}{2} + \frac{1}{4}\log\frac{5}{4} = 0.295$$

$$D(p(x)||q(x)) = \frac{3}{5}\log\frac{12}{5} + \frac{1}{5}\log\frac{2}{5} + \frac{1}{5}\log\frac{4}{5} = 0.297$$

Now that it has been established this is a non-symmetric we introduce a new distribution u to find a case which does satisfy the triangle inequality.

$$u(x) = \begin{cases} 0.27 & x = 0 \\ 0.38 & x = 1 \\ 0.35 & x = 2 \end{cases}$$

Given:

$$D(p||u) = 0.377$$

$$D(u||q) = 0.353$$

$$D(p||q) \leq D(p||u) + D(u||q)$$
$$0.295 = 0.377 + 0.353$$
$$0.295 < 0.730$$

Therefore, $D(p||q) + D(q||p)$ is non-symmetric and does not satisfy the triangle inequality.

1.2 Question 2 (1 point)

For a given value of $Y = y$, is it possible that $H(X|Y = y) \geq H(X)$? Provide a demonstration.

Let (X, Y) have the following joint distribution:

Y \ X	1	2	Σ_y
1	0	3/4	3/4
2	1/8	1/8	1/4
Σ_x	1/8	1/8	1

$$H(X) = H(\frac{1}{8}, \frac{7}{8}) = 0.544bits$$
$$H(X|Y=1) = 0bits$$
$$H(X|Y=2) = 1bits$$

Therefore, when $y = 2$, $H(X|Y=y) \geq H(X)$.

$$H(X|Y) = \frac{3}{4}H(X|Y=1) + \frac{1}{4}H(X|Y=2) = 0.25bits$$

$H(X|Y=y)$ may be greater than, less than or equal to $H(X)$. However, on average $H(X|Y=y) \leq H(X)$

1.3 Question 3 (2 points

Complete example in page 22 in the slides to compute $H_A(\theta\omega)$. Define a new string with length 6 that includes the prefix "aba" and compute the table.

Given the PFA from page 22 and the following algorithm, the table for the string "aaba" and "abaaaa" were computed.

Initialization: For $0 \leq j < |Q|$:

$$H_0(j) = 0$$
$$c_0(j) = I(j)$$

Recursion: For $0 \leq j < |Q| - 1; 1 \leq t \leq |\omega|$:

$$c_t(j) = \frac{\Sigma_{i=0}^{|Q|-1} c_{t-1}(i)p(i, \omega_t, j)}{\Sigma_{k=0}^{|Q|-1}\Sigma_{i=0}^{|Q|-1} c_{t-1}(i)p(i, \omega_t, k)}$$
$$p(\theta_{t-1} = i|\theta_t = j, \omega_{1,t}) = \frac{\Sigma_{i=0}^{|Q|-1} c_{t-1}(i)p(i, \omega_t, j)}{\Sigma_{k=0}^{|Q|-1} c_{t-1}(k)p(k, \omega_t, j)}$$
$$H_t(j) = \Sigma_{i=0}^{|Q|-1} H_{t-1}(i)p(\theta_{t-1} = i|\theta_t = j, \omega_{1,t})$$

$$-\Sigma_{i=0}^{|Q|-1} p(\theta_{t-1} = i|\theta_t = j, \omega_{1,t}) \log p(\theta_{t-1} = i|\theta_t = j, \omega_{1,t})$$

Termination: For $0 \leq j < |Q| - 1; T = |\omega| + 1$:

$$H_T(j) = \Sigma_{i=0}^{|Q|-1} H_{T-1}(i) p(\theta_{T-1} = i|\theta_T = j, \omega_{1,T})$$
$$-\Sigma_{i=0}^{|Q|-1} p(\theta_{T-1} = i|\theta_T = j, \omega_{1,T}) \log p(\theta_{T-1} = i|\theta_T = j, \omega_{1,T})$$
$$c_T(j) = T(j)$$

Table 1: Table computed for the string "aaba"

	-	a	a	b	a
H_0	0.0				
c_0	1.0				
H_1	0.0	0.0	0.0		
c_1	0.0	0.2	0.115		
H_2	0.0	0.0	0.0	0.0	0.784
c_2	0.0	0.8	0.462	0.133	0.735
H_3	0.0		0.0	0.439	
c_3	0.0		0.115	0.847	
H_4	0.0		0.0		0.953
c_4	0.0		0.308		0.265

Table 2: Table computed for the string "abaaaa"

	-	a	a	b	a	a	a
H_0	0.0						
c_0	1.0						
H_1	0.0	0.0	0.0				
c_1	0.0	0.2	0.115				
H_2	0.0	0.0	0.0	0.440	0.440	0.440	0.440
c_2	0.0	0.8	0.167	0.733	0.6	0.6	0.6
H_3	0.0		0.0				
c_3	0.0		0.833				
H_4	0.0			0.65	0.440	0.440	0.440
c_4	0.0			0.267	0.4	0.4	0.4

1.4 Question 4 (2 points

Given the example in page 39 of the slides, carry out exercise 1 to compute the increasing of λ_1, λ_2 or λ_3 (just two of them).

The ISS algorithm is used to solve this question. Specifically, the increasing of λ_1 and λ_2 are computed. The following equations are used.

$$\lambda_i = \lambda_i + \delta_i$$

$$\delta_i = \frac{1}{M} \log \frac{\tilde{p}(f_i)}{p_{\lambda_i}(f_i)}$$

$$p(y|x) = \frac{1}{Z(x)} \exp(\sum_{i=1}^{k} \delta_i f_i(x, y))$$

$$Z(x) = \sum_{y} \exp(\sum_{i=1}^{k} \delta_i f_i(x, y))$$

Due to having only one feature active at any moment, $M = 1$.

For δ_1, $f_1 = f(\omega_1, c_0)$.

$$\delta_1 = \log \frac{\tilde{p}(f_1)}{p_{\lambda_1}} = \log \frac{\tilde{p}(\omega_1, c_0)}{\tilde{p}(\omega_1) p_{\lambda_1}(c_0 \mid \omega_1)}$$

$$\tilde{p}(\omega_1) = \frac{2}{5}$$

$$\tilde{p}(\omega_1, c_0) = \frac{2}{5}$$

$$p_{\lambda_1}(c_0 \mid \omega_1) = \frac{1}{2}$$

$$\delta_1 = \log \frac{\frac{2}{5}}{\frac{2}{5}\frac{1}{2}} = 0.693$$

For δ_2, $f_2 = f(\omega_0, c_1)$.

$$\delta_2 = \log \frac{\tilde{p}(f_2)}{p_{\lambda_2}} = \log \frac{\tilde{p}(\omega_0, c_1)}{\tilde{p}(\omega_0) p_{\lambda_1}(c_1 \mid \omega_0)}$$

$$\tilde{p}(\omega_0) = \frac{3}{10}$$

$$\tilde{p}(\omega_0, c_1) = \frac{1}{10}$$

$$p_{\lambda_1}(c_1 \mid \omega_0) = \frac{1}{2}$$

$$\delta_1 = \log \frac{\frac{1}{10}}{\frac{3}{10}\frac{1}{2}} = -0.916$$

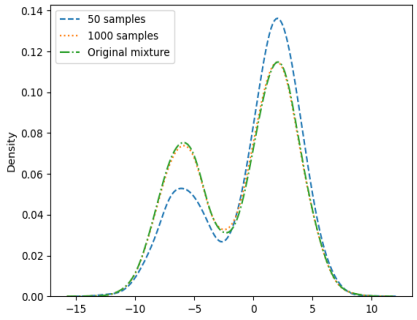
1.5 Question 5 (2 points

Repeat experiment 61 of the slides but with a training sample of size 1000. Explain your results and the conclusions. Be concise.

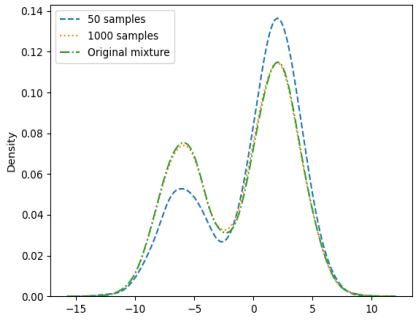
Let a gaussian mix of two unidimensional distributions with known mean (-6 and 2) and equal and known variance (4) where π_1, π_2 (0.4, 0.6) are unknown.

The experiment done is to estimate this mixture but with 3 component gaussian mixture and see how regulation affects the ability to perform this approximation.

In order to do so, 50 and 1000 samples were generated with the original distribution and approximated with the 3 component gaussian mixture one over 10,000 iterations.



(a) EM algorithm without regularization.



(b) EM algorithm with 0.1 regularization.

Table 3: Computed components after EM algorithm.

Samples	γ	π_1	π_2	π_3
50	0.0	0.289	0.711	0.0
1000	0.0	0.391	0.584	0.025
50	0.1	0.289	0.711	0.0
1000	0.1	0.392	0.586	0.022
Original	-	0.4	0.6	0.0

From the results obtained it is possible to observe that using more samples have a determining effect in order to make a better approximation of a mixture. Also, given the model was trained over many iterations it is visible that regularization has very little effect. In addition, using a high number of iterations the model overfits the samples instead of learning the real distribution.

2 Practical Questions

2.1 Question 6 (2 points)

Try to obtain the best possible results in the previous task (the mark in this exercise will depend a lot on the obtained results). The results could be checked by the professor if the experiments does not look fair. Explain your work and provide some conclusions. Hints: increase the number of training samples, both negative and/or positive samples.

This exercise studies the classification of samples after training with the MMI algorithm. The of this exercise is to train a PCFG which is able to caputure the features of right-angled, equilateral and isosceles triangles and adjust the parameters of training to obtain a error under 20%.

The following experiment shows the default example provided with the default training samples and default parameters for (0.1, 0.1, 0.1).

Table 4: Execution with default parameters and default data.

	EQ	IS	RA	Error	Error(%)
EQ	860	140	0	140	14.0
IS	410	548	42	452	45.2
RA	0	57	943	57	5.7
Total				649	21.63

Given the first example, different experimentation was done to yield better results. However, neither increasing the training samples or adjusting the parameters increased the performance of the classification. The parameters were all adjusted to {0.01, 0.05, 0.1, 0.2, 0.3, 0.5} for the default and

extended training set. Despite of this, the error value obtained was greater than 20% or even 25%.

After all this experimentation the parameters were individually adjusted to (0.6, 0.0, 0.6) for equilateral, isosceles and right-angled triangles respectively and finally yield the best results obtained.

Table 5: Execution with the best parameters and default data.

	EQ	IS	RA	Error	Error(%)
EQ	1000	0	0	0	0.0
IS	497	426	77	574	57.4
RA	0	24	976	24	2.4
Total				598	19.93

Given the results of the best execution it is possible to conclude the task was completed successfully. Having an error or 19.93%, just under 20%. However, overall it was clear the models had special difficulties detecting isosceles triangles compared to the other types.