Advanced Machine Learning: Exercises

Ramon Ruiz Dolz

April 2019

1 Exercise(*): *Compute H(C), H(S), H(C|S), H(S|C), H(C, S), and I(C; S).*

En este ejercicio se nos pide realizar los cálculos de la entropía de las variables Climatología (C) y Luminosidad (L) así como las entropías condicionales, la entropía conjunta y la intersección para estas mismas variables. Como punto de partida se toman los datos presentados en la diapositiva 13. Para la variable C, sus estados pueden ser, *despejado* (DES), *nublado* (NUB) y *lluvia* (LLU). Por otra parte, para la variable S, los estados posibles son desplazamiento *seguro* (SEG) o con *accidente* (ACC). A partir de las probabilidades conjuntas proporcionadas, es posible obtener las probabilidades de cada variable marginalizando. Por lo tanto obtenemos las siguientes probabilidades a priori para la variable C:

- P(DES) = 0.46
- P(NUB) = 0.33
- P(LLU) = 0.21

Y para la variable S:

- P(SEG) = 0.86
- P(ACC) = 0.14

Una vez obtenidas las distintas probabilidades a priori, ya es posible realizar los cálculos requeridos para el ejercicio. Mediante la formula de la entropía:

$$H(X) = -\sum_{i=1}^{|X|} p(i)log(p(i)) \tag{1}$$

Por lo tanto, se obtiene que:

$$H(C) = -(0.46log(0.46) + 0.33log(0.33) + 0.21log(0.21)) \simeq 0.468 \tag{2}$$

$$H(S) = -(0.86log(0.86) + 0.14log(0.14)) \simeq 0.176 \tag{3}$$

A demás de las entropías de cada variable, también se nos pide calcular las entropías condicionales de estas dos variables. Para ello hacemos uso de la siguiente fórmula:

$$H(Y|X) = -\sum_x \sum_y p(x,y)log(p(y|x)) \tag{4}$$

Que aplicada a nuestras variables nos conduce a los siguientes cálculos:

$$H(C|S) = \tag{5}$$
$$-((0.43log(0.5)+0.3log(0.35)+0.13log(0.15)+(0.03log(0.21)+0.03log(0.21)+0.08log(0.57))) =$$
$$-((-0.373) + (-0.042)) \simeq 0.415$$

$$\begin{aligned} H(S|C) = -((0.43log(0.93) + 0.03log(0.07)) + (0.3log(0.91) + 0.03log(0.09)) \\ + (0.13log(0.62) + 0.08log(0.38))) = -((-0.048) + (-0.044) + (-0.061)) \simeq 0.153 \end{aligned} \tag{6}$$

Una vez obtenidas tanto las entropías como las entropías condicionales asociadas a cada variable, es posible finalizar los cálculos de la entropía conjunta y la interección. Para ello se haran uso de las siguientes dos ecuaciones:

$$H(X,Y) = H(X) + H(Y|X) \tag{7}$$

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \end{aligned} \tag{8}$$

Con lo que se obtiene:

$$\begin{aligned} H(C,S) &\simeq 0.468 + 0.153 \\ &= 0.621 \end{aligned} \tag{9}$$

$$\begin{aligned} I(C;S) &\simeq 0.468 - 0.415 \\ &= 0.053 \end{aligned} \tag{10}$$

2 Exercise(***): *Reproduce an example similar to the previous example with two uni-dimensional distributions, with equal and known variance where the mean of each distribtuion and π_1, π_2 are unknown. Obtain the corresponding plots.*

En este ejercicio se nos pide reproducir el experimento mostrado en clase relacionado con el algoritmo EM. Para ello se han generado 100 muestras unidimensionales de forma aleatoria siguiendo dos distribuciones normales distintas con la varianza predefinida (2). Las muestras generadas se pueden observar en la Figure 1.

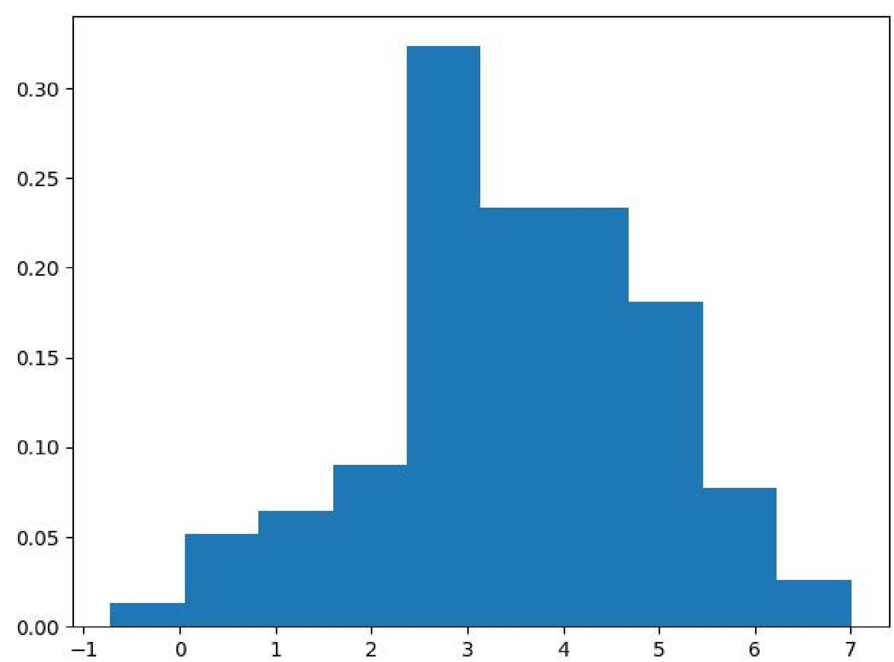

Figure 1: Distribución de las 100 muestras generadas aleatoriamente

Puesto que tanto los pesos como las medias son desconocidas, se inicializan aleatoriamente al empezar el algoritmo EM. Mediante la iteración de este algoritmo, a partir de los datos conocidos (distribución de las muestras y las varianzas) se trata de aprender la distribución de la mixtura de ambas Gaussianas. Al finalizar este algoritmo se estiman las medias (2.9 y 3.7) así como los parámetros π_1 (0.67) y π_2 (0.33). En la Figure 2 se puede observar gráficamente este resultado.

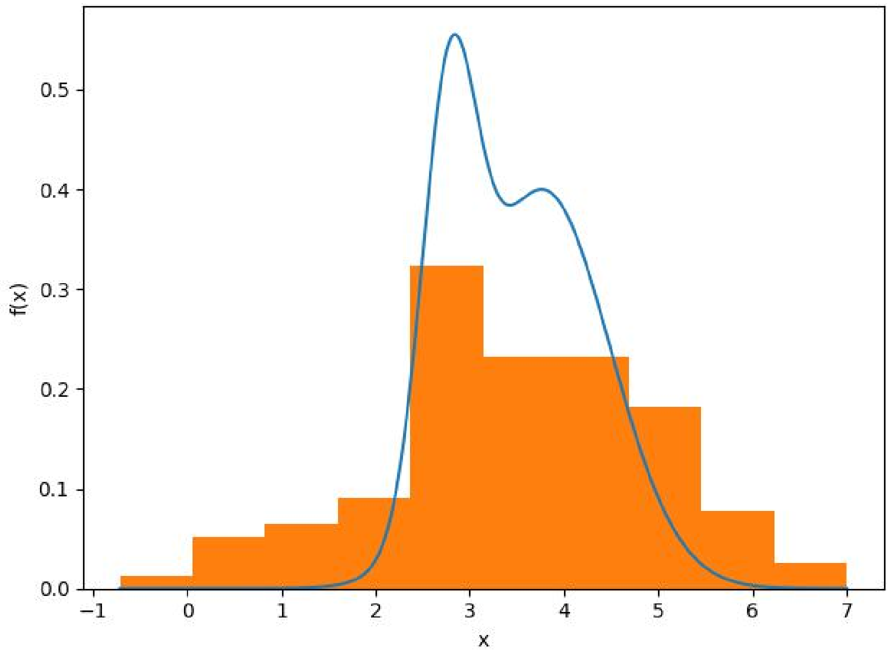

Figure 2: Mixtura de Gausianas aprendida a partir de las muestras

Como bien se puede apreciar, los parámetros aprendidos que definen a cada Gaussiana no son iguales y, por lo tanto, la distribución de probabilidades tras aplicar el EM sigue una distribución similar al de la naturaleza de las muestras.

Debido a la aleatoriedad del ejercicio se ha realizado un segundo experimento con el objetivo de aportar mayor riqueza. En este segundo experimento se han vuelto a generar 100 muestras unidimensionales de forma aleatoria siguiendo dos distribuciones Gausianas distintas con varianza conocida a priori (2). En la Figure 3 se puede observar la distribución del conjunto total de muestras generadas.

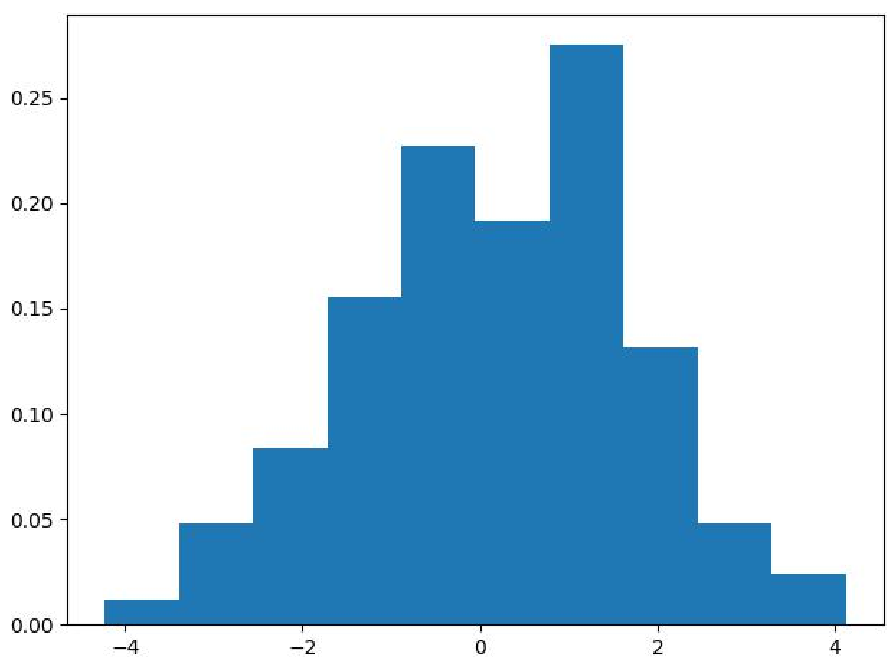

Figure 3: Distribución de las 100 muestras generadas aleatoriamente

Del mismo modo que en el ejercicio anterior, las medias y los pesos son desconocidos. Al finalizar el algoritmo, se estiman las medias (-0.5 y 0.8) así como los parámetros π_1 (0.498) y π_2 (0.502) obteniendo como resultado la mixtura observable en la Figure 4

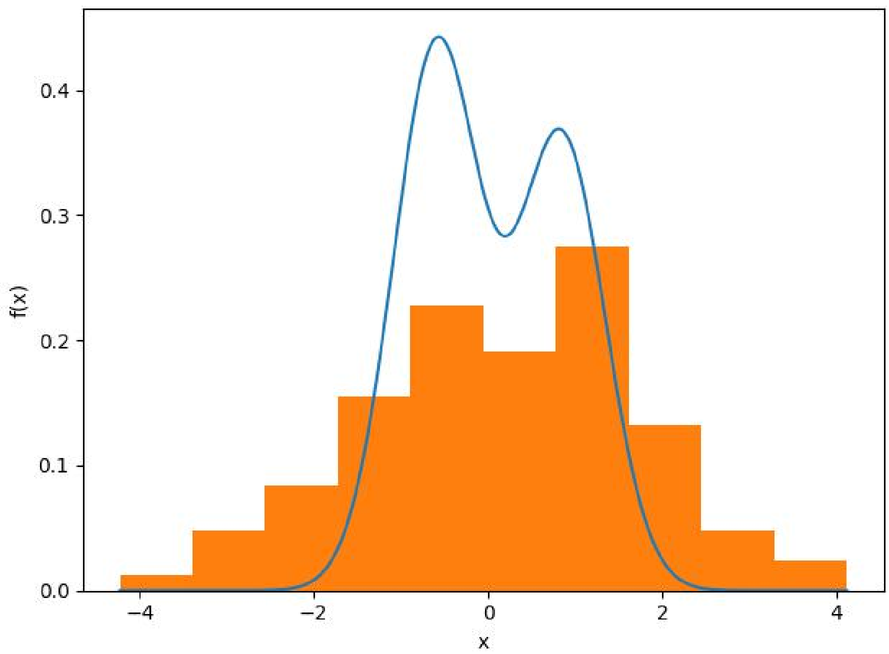

Figure 4: Mixtura de Gausianas aprendida a partir de las muestras

3 Exercise(***): *Reproduce an example similar to the previous example with two unidimensional distributions, with equal and known mean where the variances and π_1, π_2 are unknown. Obtain the corresponding plots.*

En este ejercicio se nos pide reproducir el mismo experimento que en el anterior, pero en este caso las medias son conocidas y las varianzas desconocida. Para ello, de forma similar al ejercicio anterior se han realizado dos experimentos con datos generados aleatoriamente distintos.

En el primer experimento se han generado 100 muestras unidimensionales siguiendo dos distribuciones normales distintas con medias conocidas (0 y 1). Podemos observar la distribución de las muestras generadas en este experimento en Figure 5.

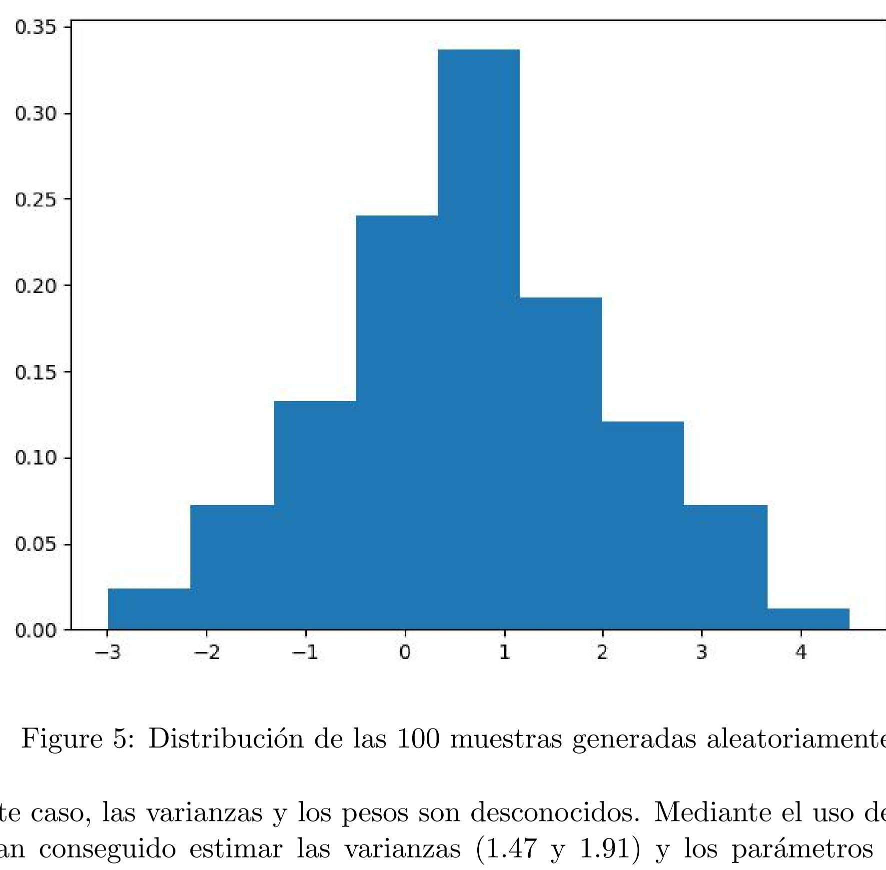

Figure 5: Distribución de las 100 muestras generadas aleatoriamente

En este caso, las varianzas y los pesos son desconocidos. Mediante el uso del algoritmo EM se han conseguido estimar las varianzas (1.47 y 1.91) y los parámetros π_1 (0.61) y

π_2 (0.39). Las mixturas de Gausianas resultantes se pueden observar gráficamente en la Figure 6

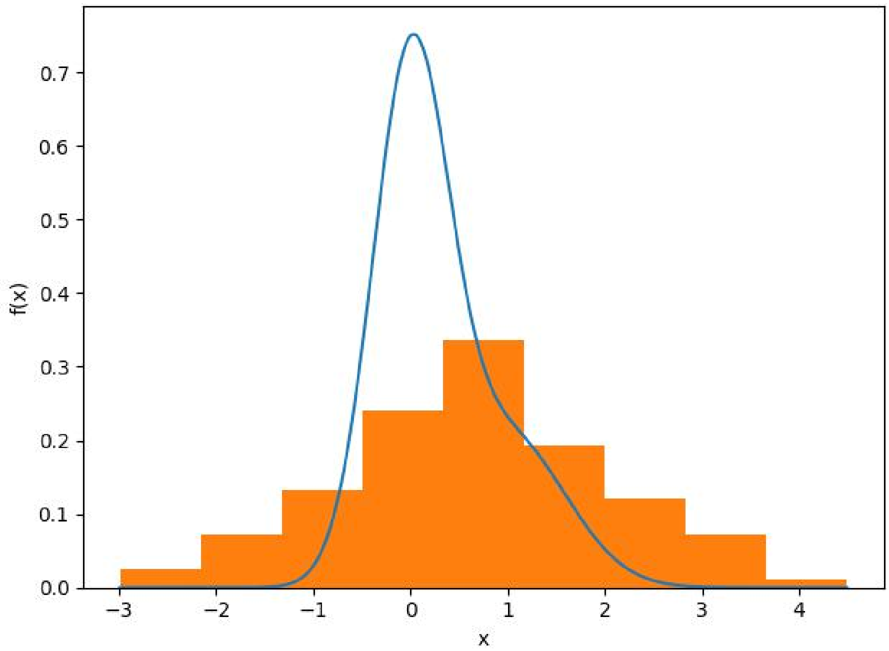

Figure 6: Mixtura de Gausianas aprendida a partir de las muestras

Finalmente, con tal de aportar riqueza al ejercicio se ha añadido un segundo experimento. De manera muy similar al experimento explicado anteriormente, se han generado 100 muestras unidimensionales de forma aleatoria con medias conocidas (0 y 1), siguiendo dos distribuciones Gausianas diferentes. En la Figure 7 se puede apreciar la distribución de las 100 muestras generadas.

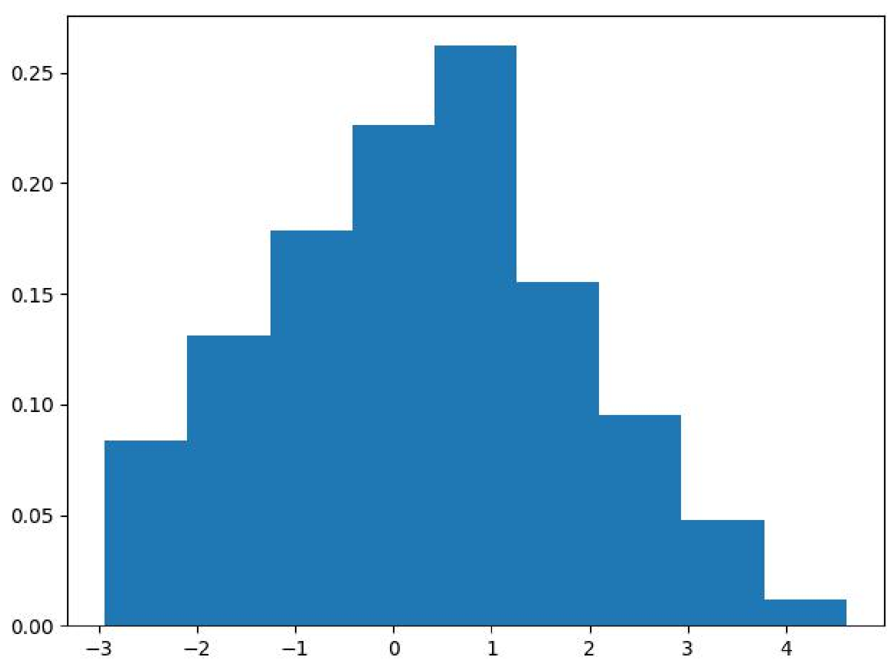

Figure 7: Distribución de las 100 muestras generadas aleatoriamente

A partir de estos datos generados se ha aprendido una mixtura de Gausianas haciendo uso del algoritmo EM. Mediante este algoritmo se han estimado las varianzas (2.19 y 2.12) y los parámetros π_1 (0.325) y π_2 (0.675). El modelo aprendido se puede observar en la Figure 8.

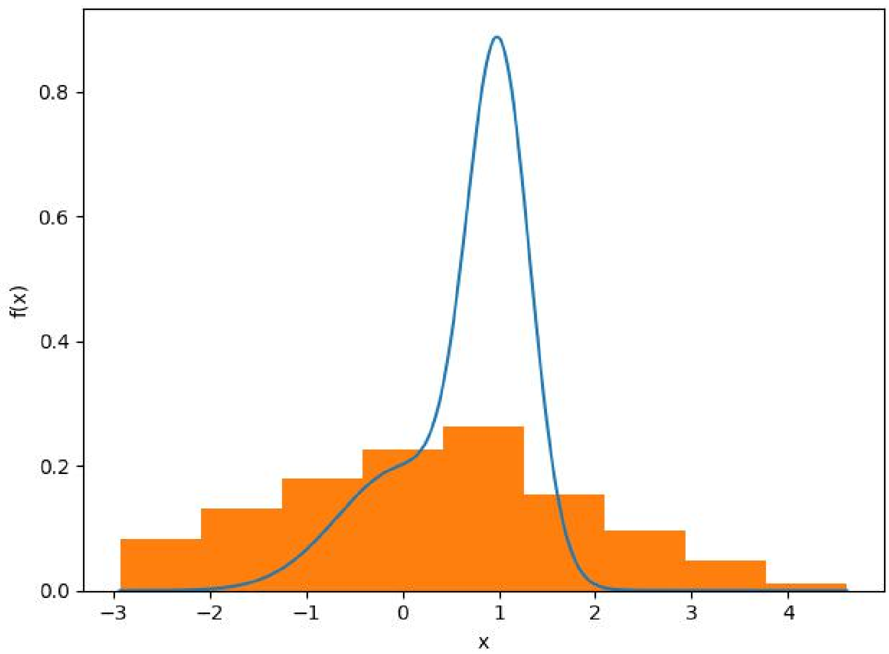

Figure 8: Mixtura de Gausianas aprendida a partir de las muestras