h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES

fbmn
FACHBEREICH MATHEMATIK
UND NATURWISSENSCHAFTEN

Data Mining 1
Sommersemester 2020

Booklet

Name: Duong, Mai Chi (750093)
Teppe, Marius (??????)
Salome Weigmann (752324)

Studiengang: Angewandte Mathematik

Inhaltsverzeichnis

Arbeitsblatt 1 (Informationstheorie)

Aufgabe 1 - Asymmetrie der Kullback-Leibler-Divergenz

Wir zeigen durch ein Gegenbeispiel, dass die Kullback-Leibler-Divergenz nicht symmetrisch ist. Die Symmetrie-Eigenschaft gilt nur, falls

$$E\left(\frac{P(X=x)}{Q(X=x)}\right) \neq E\left(\frac{Q(X=x)}{P(X=x)}\right).$$

Wir definieren die Verteilungen P und Q mit

$$P(X=0)=\frac{1}{2}, P(X=1)=\frac{1}{2} \text{ und } Q(X=0)=\frac{9}{10}, Q(X=1)=\frac{1}{10}.$$

Dann ist die Kullback-Leibler-Divergenz gegeben durch

$$\begin{aligned} D_{KL}(p,q) &= -\sum_{i=1}^{m} P(X=x_i)\cdot\log\left(\frac{Q(X=x_i)}{P(X=x_i)}\right) \\ &= -\left(\frac{1}{2}\cdot\log\left(\frac{9/10}{1/2}\right)+\frac{1}{2}\cdot\log\left(\frac{1/10}{1/2}\right)\right) \\ &= +0,7369655 \end{aligned}$$

und

$$\begin{aligned} D_{KL}(q,p) &= -\sum_{i=1}^{m} Q(X=x_i)\cdot\log\left(\frac{P(X=x_i)}{Q(X=x_i)}\right) \\ &= -\left(\frac{9}{10}\cdot\log\left(\frac{1/2}{9/10}\right)+\frac{1}{10}\cdot\log\left(\frac{1/2}{1/10}\right)\right) \\ &= +0,53100. \end{aligned}$$

Es wird ersichtlich, dass

$$D_{KL}(p,q) \neq D_{KL}(q,p).$$

Aufgabe 2 - Indifferenzprinzip

Das Indifferenzprinzip (auch Prinzip vom unzureichenden Grund genannt) besagt, dass bei unterscheidbaren Ereignissen, die sich gegenseitig ausschließen, die Gleichverteilung angenommen werden sollte, wenn keine weiteren Informationen über die Ereignisse bekannt sind.
Dieses Prinzip lässt sich mit der Entropie und der Eigenschaft der Kullback-Leibler-Divergenz $D_{KL}(p,q) \geq 0$ begründen:

Unter der Annahme, dass dem Modell mit n Ereignissen eine diskrete Gleichverteilung zugrunde liegt, vereinfacht sich die Entropie zu:

$$H(X) = -\log(P(X = x_i)) = I_x(x_i),$$

welches den Informationsgehalt I_x des Ereignisses x_i beschreibt. Folglich besitzt jedes Ereignis den gleichen Informationsgehalt, sodass wir kein Ereignis über- bzw. unterschätzen.

Sei P nun die tatsächliche Verteilung der Ereignisse und Q die Approximation. Nimmt man für Q die Gleichverteilung an, also $Q(X = x_i) = \frac{1}{m}$, so ergibt sich folgende Kullback-Leibler-Divergenz:

$$\begin{aligned}
D_{KL}(p,q) &= -\sum_{i=1}^{m} P(X = x_i) log(\frac{Q(X = x_i)}{P(X = x_i)}) \\
&= \sum_{i=1}^{m} P(X = x_i) log(P(X = x_i)) - \sum_{i=1}^{m} P(X = x_i) log(Q(X = x_i)) \\
&= \sum_{i=1}^{m} P(X = x_i) log(P(X = x_i)) - \sum_{i=1}^{m} P(X = x_i) log(\frac{1}{m}) \\
&= \sum_{i=1}^{m} P(X = x_i) log(P(X = x_i)) - log(\frac{1}{m}) \sum_{i=1}^{m} P(X = x_i) \\
&= \sum_{i=1}^{m} P(X = x_i) log(P(X = x_i)) - log(\frac{1}{m}) * 1 \\
&\geq 0
\end{aligned}$$

Daraus kann geschlussfolgert werden:

$$H(p) = -\sum_{i=1}^{m} P(X = x_i) log(P(X = x_i)) \leq -log(\frac{1}{m}) = H_{max}(p)$$

Somit ist eine obere Schranke gefunden. Dies bedeutet, dass die mittlere Information pro Zeichen nicht größer werden kann als $H_{max}(p) = -log(\frac{1}{m})$.

Aufgabe 3 - Zusammenhang D_{KL} und $I(X,Y)$

Wir beweisen im Folgendem den Zusammenhang:

$$D_{KL}(P(X,Y), P(X)P(Y)) = I(X,Y).$$

Beweis:

$$\begin{aligned}
D_{KL}(P(X,Y), P(X)P(Y)) =& -\sum_{i=1}^{m}\sum_{j=1}^{n} P(X=x_i, Y=y_j) \cdot \log\left(\frac{P(X=x_i)\cdot P(Y=y_j)}{P(X=x_i, Y=y_j)}\right) \\
=& -\sum_{i=1}^{m}\sum_{j=1}^{n} P(X=x_i, Y=y_j) \cdot \log\left(P(X=x_i)\right) \\
& -\sum_{i=1}^{m}\sum_{j=1}^{n} P(X=x_i, Y=y_j) \cdot \log\left(\frac{P(Y=y_j)}{P(X=x_i, Y=y_j)}\right) \\
=& -\sum_{i=1}^{m} P(X=x_i) \cdot \log\left(P(X=x_i)\right) \\
& -\left[-\sum_{i=1}^{m}\sum_{j=1}^{n} P(X=x_i, Y=y_j) \cdot \log\left(\frac{P(X=x_i, Y=y_j)}{P(Y=y_j)}\right)\right] \\
=& H(X) - H(X|Y) \\
=& I(X,Y)
\end{aligned}$$

Aufgabe 4 - binärer symmetrischer Kanal

Bei einem binärem symmetrischem Kanal ist die Wahrscheinlichkeit einer Falschübermittlung gleich hoch.

Aus der zur Aufgabe gehörigen Graphik können die folgenden Wahrscheinlichkeiten entnommen werden:

$$\begin{aligned} P(Y=0|X=0) &= 1-\epsilon \\ P(Y=0|X=1) &= \epsilon \\ P(Y=1|X=0) &= \epsilon \\ P(Y=1|X=1) &= 1-\epsilon \end{aligned}$$

Zudem ist bekannt, dass eine gleichverteilte Inputverteilung X vorliegt:

$$\begin{aligned} P(X=0) &= \frac{1}{2} \\ P(X=1) &= \frac{1}{2}. \end{aligned}$$

Zu Berechnen sind $P(X,Y), P(X|Y), H(X), H(X|Y), I(X,Y)$:

Die gemeinsame Verteilung $P(X,Y)$ lässt sich wie folgt ermitteln:

$$P(X=x_i, Y=y_i) = P(Y=y_i|X=x_i) \cdot P(X=x_i). \tag{0.0.1}$$

Damit ergibt sich die folgende Verteilung:

$$\begin{aligned} P(X=0,Y=0) &= \frac{1-\epsilon}{2} \\ P(X=0,Y=1) &= \frac{\epsilon}{2} \\ P(X=1,Y=0) &= \frac{\epsilon}{2} \\ P(X=1,Y=1) &= \frac{1-\epsilon}{2}. \end{aligned}$$

Die bedingte Verteilung $P(X|Y)$ lässt sich auch mit Gleichung 0.0.1 berechnen:

$$\begin{aligned} P(X=0|Y=0) &= 1-\epsilon \\ P(X=0|Y=1) &= \epsilon \\ P(X=1|Y=0) &= \epsilon \\ P(X=1|Y=1) &= 1-\epsilon. \end{aligned}$$

Die Entropie der Inputverteilung X beträgt

$$\begin{aligned} H(X) &= -\sum_{i=1}^{m} P(X=x_i) \cdot \log(P(X=x_i)) \\ &= -2 \cdot \frac{1}{2} \cdot \log_2\left(\frac{1}{2}\right) = 1. \end{aligned}$$

Die bedingte Entropie von X unter Y lässt sie wie folgt berechnen.

$$\begin{aligned} H(X|Y) &= \sum_{j=1}^{n} P(Y = y_j) \cdot H(X|Y = y_j) \\ &= -\sum_{i=1}^{2}\sum_{j=1}^{2} P(Y = x_i, Y = y_j) \cdot \log_2 P(X = x_i|Y = y_j) \\ &= -\left[\frac{\epsilon}{2} \cdot \log_2(\epsilon) + \frac{1-\epsilon}{2} \cdot \log_2(1-\epsilon) + \frac{1-\epsilon}{2} \cdot \log_2(1-\epsilon) + \frac{\epsilon}{2} \cdot \log_2(\epsilon)\right] \\ &= -\epsilon \cdot \log_2(\epsilon) - (1-\epsilon)\log_2(1-\epsilon). \end{aligned}$$

Somit erhalten wir eine Transformation von X und Y von

$$\begin{aligned} I(X,Y) &= H(X) - H(X|Y) \\ &= 1 + \epsilon \cdot \log_2(\epsilon) + (1-\epsilon)\log_2(1-\epsilon). \end{aligned}$$

Arbeitsblatt 2

Anhang

Abbildungsverzeichnis

Tabellenverzeichnis

Quellcodeverzeichnis

Literaturverzeichnis

[AB00] Adam, Anna ; Brot, Bernd: *Titel*. Springer, 0000