

Дипломна работа

Диана Генева <dageneva@qtrp.org>

2019

Съдържание

A Приложение за Максимизиране на ентропията	2
---	---

Приложение А

Приложение за Максимизиране на ентропията

Нека имаме входни данни $\mathcal{D} = (x_1, y_1), \dots, (x_n, y_n)$, където $x_i \in X$, а $y_i \in Y$.

Търсим това разпределение p , което приближава разпределението, генерирано данните в \mathcal{D} и което се държи равномерно иначе.

Тоест търсеното разпределение p , трябва да изпълнява:

$$p(x, y) = \tilde{p}(x)p(y|x),$$

Тук с \tilde{p} означаваме емпиричното разпределение, дефинирано като:

$$\forall (x, y) \in X \times Y : \tilde{p}(x, y) = \frac{\#(x, y)}{n}, \text{ където}$$

$$\#(x, y) = \begin{cases} \text{брой срещания на } (x, y) \text{ в } \mathcal{D}, & \text{ако } (x, y) \in \mathcal{D} \\ 0, & \text{иначе} \end{cases}$$

С други думи:

$$\begin{aligned} p(x) &= \sum_{y \in Y} p(x, y) = \sum_{y \in Y} \tilde{p}(x)p(y|x) \\ &= \tilde{p}(x) \sum_{y \in Y} p(y|x) \\ &= \tilde{p}(x) \end{aligned}$$

и да максимизира ентропията:

$$H_p(X, Y) = - \sum_{(x, y) \in X \times Y} p(x, y) \log(p(x, y))$$

Нека имаме още множество от характеристични функции \mathcal{H} , $|\mathcal{H}| = K$, които са от вида $h_i : X \times Y \rightarrow [0, 1]$.

Ако с $E(q, h)$ означим очакването на h , спрямо разпределение q , тоест:

$$E(q, h) = \sum_{(x, y) \in X \times Y} q(x, y)h(x, y)$$

То искаме за търсеното p да е изпълнено:

$$E(p, h) = E(\tilde{p}, h), \forall h \in \mathcal{H}$$

Ако означим:

$$P = \{p | E(p, h) = E(\tilde{p}, h), \forall h \in \mathcal{H}\}$$

тогава искаме да намерим

$$\begin{aligned} \hat{p} &= \operatorname{argmax}_{p \in P} H_p(X, Y) \\ &= \operatorname{argmax}_{p \in P} \left(- \sum_{(x, y) \in X \times Y} p(x, y) \log(p(x, y)) \right) \\ &= \operatorname{argmax}_{p \in P} \left(- \sum_{(x, y) \in X \times Y} p(x, y) \log(\tilde{p}(x)p(y|x)) \right) \\ &= \operatorname{argmax}_{p \in P} \left(- \sum_{(x, y) \in X \times Y} p(x, y) \log(\tilde{p}(x)) - \sum_{(x, y) \in X \times Y} p(x, y) \log(p(y|x)) \right) \\ &= \operatorname{argmax}_{p \in P} \left(- \sum_{(x, y) \in X \times Y} \tilde{p}(x)p(y|x) \log(\tilde{p}(x)) + H_p(Y|X) \right) \\ &= \operatorname{argmax}_{p \in P} \left(- \sum_{x \in X} \tilde{p}(x) \log(\tilde{p}(x)) \sum_{y \in Y} p(y|x) + H_p(Y|X) \right) \\ &= \operatorname{argmax}_{p \in P} \left(- \sum_{x \in X} \tilde{p}(x) \log(\tilde{p}(x)) + H_p(Y|X) \right) \\ &\quad - \sum_{x \in X} \tilde{p}(x) \log(\tilde{p}(x)) \text{ е константа спрямо } p, \text{ следователно:} \\ &= \operatorname{argmax}_{p \in P} H_p(Y|X) \end{aligned}$$

За да решим тази оптимизационна задача, ще ползваме множители на Лагранж. Тъй като имаме K ограничения за всяка от характеристичните функции и трябва да отчетем, че търсим разпределение, задачата ще има вида:

$$\Lambda(p, \lambda) = H_p(Y|X) + \sum_{i=1}^K \lambda_i (E(p, h_i) - E(\tilde{p}, h_i)) + \sum_{x \in X} \mu_x \sum_{y \in Y} p(y|x) - 1$$

Нека фиксираме едно $x_0 \in X, y_0 \in Y$.

$$\begin{aligned}
& \frac{\partial (\Lambda(p, \lambda))}{\partial p(y_0|x_0)} = \\
& = \frac{\partial H_p(Y|X)}{\partial p(y_0|x_0)} + \frac{\partial (\lambda_i(E(p, h_i) - E(\tilde{p}, h_i)))}{\partial p(y_0|x_0)} + \frac{\left(\sum_{x \in X} \mu_x \sum_{y \in Y} p(y|x) - 1 \right)}{\partial p(y_0|x_0)} \\
& = \frac{\partial \left(- \sum_{(x,y) \in X \times Y} \tilde{p}(x)p(y|x) \log(p(y|x)) \right)}{\partial p(y_0|x_0)} + \frac{\partial \left(\sum_{i=1}^K \lambda_i \left[\sum_{(x,y) \in X \times Y} \tilde{p}(x)h_i(x, y) (p(y|x) - \tilde{p}(y|x)) \right] \right)}{\partial p(y_0|x_0)} + \mu_{x_0} \\
& = -\tilde{p}(x_0)\log(p(y_0|x_0)) - \tilde{p}(x_0) + \sum_{i=1}^K \lambda_i \tilde{p}(x_0)h_i(x_0, y_0) + \mu_{x_0}
\end{aligned}$$

Искаме да я нулираме:

$$\begin{aligned}
& -\tilde{p}(x_0)\log(p(y_0|x_0)) - \tilde{p}(x_0) + \sum_{i=1}^K \lambda_i \tilde{p}(x_0)h_i(x_0, y_0) + \mu_{x_0} = 0 \leftrightarrow \\
& \tilde{p}(x_0)\log(p(y_0|x_0)) = \sum_{i=1}^K \lambda_i \tilde{p}(x_0)h_i(x_0, y_0) - \tilde{p}(x_0) + \mu_{x_0} \leftrightarrow \\
& \log(p(y_0|x_0)) = \sum_{i=1}^K \lambda_i h_i(x_0, y_0) - 1 + \frac{\mu_{x_0}}{\tilde{p}(x_0)} \leftrightarrow \\
& p(y_0|x_0) = \exp \left(\sum_{i=1}^K \lambda_i h_i(x_0, y_0) \right) \exp \left(-1 + \frac{\mu_{x_0}}{\tilde{p}(x_0)} \right) \tag{A.0.1}
\end{aligned}$$

Производната по μ_{x_0} ни дава:

$$\begin{aligned} \sum_{y \in Y} p(y|x_0) &= 1 \leftrightarrow \\ \sum_{y \in Y} \exp \left(\sum_{i=1}^K \lambda_i h_i(x_0, y) - 1 + \frac{\mu_{x_0}}{\tilde{p}(x_0)} \right) &= 1 \leftrightarrow \\ \exp \left(-1 + \frac{\mu_{x_0}}{\tilde{p}(x_0)} \right) \sum_{y \in Y} \exp \left(\sum_{i=1}^K \lambda_i h_i(x_0, y) \right) &= 1 \\ \leftrightarrow \\ \exp \left(-1 + \frac{\mu_{x_0}}{\tilde{p}(x_0)} \right) &= \frac{1}{\sum_{y \in Y} \exp \left(\sum_{i=1}^K \lambda_i h_i(x_0, y) \right)} \end{aligned}$$

Заместваме в [Уравнение A.0.1](#):

$$p(y_0|x_0) = \frac{\exp \left(\sum_{i=1}^K \lambda_i h_i(x_0, y_0) \right)}{\sum_{y \in Y} \exp \left(\sum_{i=1}^K \lambda_i h_i(x_0, y) \right)}$$

Следователно вида на търсеното \hat{p} е $\hat{p}(x, y) = \pi \prod_{i=1}^K e^{\lambda_i h_i(x, y)}$, като π е нормализиращата константа. Ще покажем че \hat{p} , което минимизира ентропията също минимизира и условното правдоподобие.

Нека с Q означим всички разпределения с желание вид. Имаме, че $Q = \{p \mid p(x, y) = \pi \prod_{i=1}^K e^{\lambda_i h_i(x, y)}\}$. За да намерим оптималното разпределение, ще ни е нужно да дефинираме разстояние между разпределения - "Разстояние" на Кулбек-Лайблър:

$$D(p, q) = \sum_{(x, y) \in X \times Y} p(x, y) \log \left(\frac{p(x, y)}{q(x, y)} \right)$$

"Разстоянието" на Кулбек-Лайблър всъщност не е разстояние в математическия смисъл (в смисъла на метрика), тъй като не е симетрична функция, но често се използва за разстояние между разпределения, тъй като има този интуитивен смисъл. Затова ще продължим да го наричаме разстояние, пропускайки кавичките.

С това сме готови да покажем следните твърдения:

Твърдение 1. За всеки две разпределения p и q , $D(p, q) \geq 0$, като $D(p, q) = 0 \iff p = q$

Доказателство: Тъй като p е разпределение и е изпълнено, че $\sum_{(x, y) \in X \times Y} p(x, y) = 1$,

можем да приложим неравенството на Йенсен:

$$\sum_{i=1}^{\infty} p(x_i, y_i) f(z_i) \leq f \left(\sum_{i=1}^{\infty} p(x_i, y_i) z_i \right), \forall i : z_i \in \mathbb{R},$$

където f е вдлъбната. Ако за f е изпълнено, че $f'' < 0$, то равенство се достига, когато z_i е константа.

$$\begin{aligned}
 -D(p, q) &= - \sum_{(x,y) \in X \times Y} p(x, y) \log \left(\frac{p(x, y)}{q(x, y)} \right) \\
 &= \sum_{(x,y) \in X \times Y} p(x, y) \log \left(\frac{q(x, y)}{p(x, y)} \right) \\
 &\leq \log \left(\sum_{(x,y) \in X \times Y} p(x, y) \frac{q(x, y)}{p(x, y)} \right) \\
 &\leq \log \left(\sum_{(x,y) \in X \times Y} q(x, y) \right) = 0 \\
 &\Leftrightarrow D(p, q) \geq 0
 \end{aligned}$$

Тъй като втората производна на логаритъма е винаги отрицателна, равенство при неравенството на Йенсен се достига, когато $\frac{q(x, y)}{p(x, y)}$ е константа, тоест:

$$\begin{aligned}
 q(x, y) &= Cp(x, y) \\
 \sum_{(x,y) \in \mathcal{D}} q(x, y) &= \sum_{(x,y) \in \mathcal{D}} Cp(x, y) \\
 &\Leftrightarrow C = 1 \\
 &\Leftrightarrow p(x, y) = q(x, y) \forall (x, y) \in X \times Y
 \end{aligned}$$

Твърдение 2. За всеки $p_1, p_2 \in P, q \in Q$ е изпълнено:

$$\sum_{(x,y) \in X \times Y} p_1(x, y) \log(q(x, y)) = \sum_{(x,y) \in X \times Y} p_2(x, y) \log(q(x, y))$$

Доказателство:

$$\begin{aligned}
& \sum_{(x,y) \in X \times Y} p_1(x,y) \log(q(x,y)) \\
&= \sum_{(x,y) \in X \times Y} p_1(x,y) \log \left(\pi \prod_{h_i \in \mathcal{H}} e^{\lambda_i h_i(x,y)} \right) \\
&= \sum_{(x,y) \in X \times Y} p_1(x,y) \left(\log(\pi) + \log \left(\prod_{h_i \in \mathcal{H}} e^{\lambda_i h_i(x,y)} \right) \right) \\
&= \sum_{(x,y) \in X \times Y} p_1(x,y) \left(\log(\pi) + \sum_{h_i \in \mathcal{H}} \log(e^{\lambda_i h_i(x,y)}) \right) \\
&= \sum_{(x,y) \in X \times Y} p_1(x,y) \left(\log(\pi) + \sum_{h_i \in \mathcal{H}} \lambda_i h_i(x,y) \right) \\
&= \sum_{(x,y) \in X \times Y} p_1(x,y) \log(\pi) + \sum_{(x,y) \in X \times Y} p_1(x,y) \sum_{h_i \in \mathcal{H}} \lambda_i h_i(x,y) \\
&= \log(\pi) \sum_{(x,y) \in X \times Y} p_1(x,y) + \sum_{(x,y) \in X \times Y} \sum_{h_i \in \mathcal{H}} p_1(x,y) \lambda_i h_i(x,y) \\
&= \log(\pi) \cdot 1 + \sum_{(x,y) \in X \times Y} \sum_{h_i \in \mathcal{H}} p_1(x,y) \lambda_i h_i(x,y) \\
&= \log(\pi) \cdot 1 + \sum_{h_i \in \mathcal{H}} \lambda_i \sum_{(x,y) \in X \times Y} p_1(x,y) h_i(x,y)
\end{aligned}$$

Тъй като $p_2 \in P$ и $E(p_1, h) = E(\tilde{p}, h) = E(p_2, h) \forall h \in \mathcal{H}$:

$$= \log(\pi) \cdot 1 + \sum_{h_i \in \mathcal{H}} \lambda_i \sum_{(x,y) \in X \times Y} p_2(x,y) h_i(x,y)$$

Използваме и че $\sum_{(x,y) \in X \times Y} p_2(x,y) = 1$

$$\begin{aligned}
&= \log(\pi) \sum_{(x,y) \in X \times Y} p_2(x,y) + \sum_{h_i \in \mathcal{H}} \lambda_i \sum_{(x,y) \in X \times Y} p_2(x,y) h_i(x,y) \\
&= \sum_{(x,y) \in X \times Y} p_2(x,y) \log(q(x,y))
\end{aligned}$$

Твърдение 3. Ако $p \in P, q \in Q, r \in P \cap Q$, то $D(p, q) = D(p, r) + D(r, q)$

Доказателство:

$$D(p, r) + D(r, q) =$$

$$\begin{aligned} &= \sum_{(x,y) \in X \times Y} p(x, y) \log \left(\frac{p(x, y)}{r(x, y)} \right) + \sum_{(x,y) \in X \times Y} r(x, y) \log \left(\frac{r(x, y)}{q(x, y)} \right) \\ &= \sum_{(x,y) \in X \times Y} p(x, y) \log(p(x, y)) - \sum_{(x,y) \in X \times Y} p(x, y) \log(r(x, y)) + \\ &\quad \sum_{(x,y) \in X \times Y} r(x, y) \log(r(x, y)) - \sum_{(x,y) \in X \times Y} r(x, y) \log(q(x, y)) \end{aligned}$$

по **Твърдение 2** за $p, r \in P$ и $r \in Q$

$$\begin{aligned} &= \sum_{(x,y) \in X \times Y} p(x, y) \log(p(x, y)) - \sum_{(x,y) \in X \times Y} p(x, y) \log(r(x, y)) + \\ &\quad \sum_{(x,y) \in X \times Y} \cancel{r(x, y) \log(r(x, y))} - \sum_{(x,y) \in X \times Y} r(x, y) \log(q(x, y)) \end{aligned}$$

$$= \sum_{(x,y) \in X \times Y} p(x, y) \log(p(x, y)) - \sum_{(x,y) \in X \times Y} r(x, y) \log(q(x, y))$$

по **Твърдение 2** за $p, r \in P$ и $q \in Q$

$$\begin{aligned} &= \sum_{(x,y) \in X \times Y} p(x, y) \log(p(x, y)) - \sum_{(x,y) \in X \times Y} \cancel{p(x, y) \log(q(x, y))} \\ &= \sum_{(x,y) \in X \times Y} p(x, y) \log \left(\frac{p(x, y)}{q(x, y)} \right) = D(p, q) \end{aligned}$$

Твърдение 4. Ако $r \in P \cap Q$, то r е единствено и $r = \hat{p}$

Доказателство:

Нека $r \in P \cap Q$. Условието $r = \hat{p}$ значи, че $r = \operatorname{argmax}_{p \in P} H_p(X, Y)$, тоест ще покажем, че за всяко $p \in P : H_r(X, Y) \geq H_p(X, Y)$.

Нека $u \in Q$, такова че $u(x, y) \neq 0, \forall (x, y) \in X \times Y$. Всъщност всяко разпределение q от Q е такова, защото $\sum_{(x,y) \in X \times Y} e^\bullet > 0$, а $\pi \neq 0$, защото π е константа и ако

$\pi = 0$, тогава $\sum_{(x,y) \in X \times Y} q(x, y) = 0$ и не изпълнява условието за разпределение.

Нека фиксираме произволно $p \in P$. Тогава от **Твърдение 3** следва, че

$$D(p, u) = D(p, r) + D(r, u)$$

$$D(p, u) \stackrel{\text{Твърдение 1}}{\geq} D(r, u)$$

$$\begin{aligned} \sum_{(x,y) \in X \times Y} p(x, y) \log \left(\frac{p(x, y)}{u(x, y)} \right) &\geq \sum_{(x,y) \in X \times Y} r(x, y) \log \left(\frac{r(x, y)}{u(x, y)} \right) \\ - H_p(X, Y) - \sum_{(x,y) \in X \times Y} p(x, y) \log(u(x, y)) &\geq -H_r(X, Y) - \sum_{(x,y) \in X \times Y} r(x, y) \log(u(x, y)) \end{aligned}$$

по **Твърдение 2** за $p, r \in P$ и $u \in Q$ следва :

$$-H_p(X, Y) - \sum_{(x,y) \in X \times Y} p(x, y) \log(u(x, y)) \geq -H_r(X, Y) - \sum_{(x,y) \in X \times Y} \cancel{p(x, y) \log(u(x, y))}$$

$$H_r(X, Y) \geq H_p(X, Y)$$

Следователно $r = \operatorname{argmax}_{p \in P} H_p(X, Y)$

Сега нека видим защо r е единствено. Нека $r' = \operatorname{argmax}_{p \in P} H_p(X, Y)$. Тогава:

$$H_{r'}(X, Y) = H_r(X, Y) \leftrightarrow D(r, u) = D(r', u)$$

но $D(r, u) = D(r, r') + D(r', u)$ по **Твърдение 3**

$$\Rightarrow D(r, r') = 0$$

$$\stackrel{\text{Твърдение 1}}{\Rightarrow} r = r'$$

Дефинираме функцията $L(p)$:

$$L(p) = \sum_{(x,y) \in X \times Y} \tilde{p}(x, y) \log(p(x, y))$$

Твърдение 5. Ако $r \in P \cap Q$, то r е единствено и $r = \operatorname{argmax}_{q \in Q} L(q)$

Доказателство: Искаме да покажем, че за всяко $q \in Q : L(r) \geq L(q)$.

Нека фиксираме едно $q \in Q$, а \tilde{p} е емпиричното разпределение и следователно $\tilde{p} \in P$ по дефиниция.

Тогава от **Твърдение 3** следва, че:

$$D(\tilde{p}, q) = D(\tilde{p}, r) + D(r, q)$$

$$D(\tilde{p}, q) \stackrel{\text{Твърдение 1}}{\geq} D(\tilde{p}, r)$$

$$\begin{aligned} \sum_{(x,y) \in X \times Y} \tilde{p}(x, y) \log \left(\frac{\tilde{p}(x, y)}{q(x, y)} \right) &\geq \sum_{(x,y) \in X \times Y} \tilde{p}(x, y) \log \left(\frac{\tilde{p}(x, y)}{r(x, y)} \right) \\ - \cancel{H_{\tilde{p}}(X, Y)} - L(q) &\geq -\cancel{H_{\tilde{p}}(X, Y)} - L(r) \\ \Leftrightarrow L(r) &\geq L(q) \end{aligned}$$

Сега нека $r' = \operatorname{argmax}_{q \in Q} L(q)$, тоест $L(r) = L(r') \Rightarrow D(\tilde{p}, r) = D(\tilde{p}, r')$.

Но по **Твърдение 3**, $D(\tilde{p}, r') = D(\tilde{p}, r) + D(r, r') \Rightarrow D(r, r') = 0 \stackrel{\text{Твърдение 1}}{\Leftrightarrow} r' = r$, следователно r е единствено.

Сега, нека дефинираме условно правдоподобие на разпределение p при дадено множество \mathcal{D} като:

$$\widehat{L}_{\mathcal{D}}(Y|X) = \prod_{(x,y) \in X \times Y} p(y|x)^{\#(x,y)}$$

Тъй като логаритъмът е вдлъбната и монотонно растяща функция, често се разглежда за удобство:

$$\log(\widehat{L}_{\mathcal{D}}(Y|X)) = \sum_{(x,y) \in X \times Y} \#(x,y) \log(p(y|x))$$

Тъй като $\hat{p} \in P \cap Q$, по горното твърдение имаме:

$$\begin{aligned} \hat{p} &= \operatorname{argmax}_{p \in Q} L(p) \\ &= \operatorname{argmax}_{p \in Q} \left(\sum_{(x,y) \in X \times Y} \tilde{p}(x,y) \log(p(x,y)) \right) \\ &= \operatorname{argmax}_{p \in Q} \left(\sum_{(x,y) \in X \times Y} \tilde{p}(x,y) \log(\tilde{p}(x)p(y|x)) \right) \\ &= \operatorname{argmax}_{p \in Q} \left(\sum_{(x,y) \in X \times Y} \tilde{p}(x,y) \log(\tilde{p}(x)) + \sum_{(x,y) \in X \times Y} \tilde{p}(x,y) \log(p(y|x)) \right) \\ &= \operatorname{argmax}_{p \in Q} \left(\sum_{(x,y) \in X \times Y} \tilde{p}(x,y) \log(p(y|x)) \right) \\ &= \operatorname{argmax}_{p \in Q} \left(0 + \sum_{(x,y) \in \mathcal{D}} \frac{\#(x,y)}{n} \log(p(y|x)) \right) \\ &= \operatorname{argmax}_{p \in Q} \left(\sum_{(x,y) \in \mathcal{D}} \#(x,y) \log(p(y|x)) \right) \\ &= \operatorname{argmax}_{p \in Q} \left(\log(\widehat{L}_{\mathcal{D}}(Y|X)) \right) \end{aligned}$$

От [Твърдение 4](#) и [Твърдение 5](#), че ако вземем разпределение от сечението на P и Q , то е единствено и е равно на $\hat{p} = \operatorname{argmax}_{p \in P} H_p(X, Y) = \operatorname{argmax}_{p \in P} H_p(Y|X) = \operatorname{argmax}_{q \in Q} L(q) = \operatorname{argmax}_{q \in Q} \log(\widehat{L}_{\mathcal{D}}(Y|X))$