Дипломна работа

Диана Генева <dageneva@qtrp.org>

2019

Съдържание

e

Приложение А

Приложение за Максимизиране на ентропията

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

Това търсим разпределение p, което приближава разпределението, генерирало данните в $X \times Y$ и което се държи равномерно иначе.

Тоест търсеното разпределение p, трябва да изпълнява:

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

където с $\tilde{p}$ означаваме емпиричната разпределение $\left(\forall (x, y) \in X \times Y : \tilde{p}(x, y) = \frac{\#(x, y)}{n}\right)$.
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:

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

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

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:

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

тоест

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

Ако означим:

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

тогава искаме да намерим

$$
\begin{aligned}
\hat{p} &= argmax_{p\in P} H(X,Y) \\
&= argmax_{p\in P} \left(- \sum_{(x,y)\in X\times Y} p(x,y)\, log(p(x,y)) \right) \\
&= argmax_{p\in P} \left(- \sum_{(x,y)\in X\times Y} p(x,y)\, log(\tilde{p}(x)p(y|x)) \right) \\
&= argmax_{p\in P} \left(- \sum_{(x,y)\in X\times Y} p(x,y) log(\tilde{p}(x)) - \sum_{(x,y)\in X\times Y} p(x,y)\, log(p(y|x)) \right) \\
&= argmax_{p\in P} \left(- \sum_{(x,y)\in X\times Y} \tilde{p}(x)p(y|x) log(\tilde{p}(x)) + H(Y|X) \right) \\
&= argmax_{p\in P} \left(- \sum_{x\in X} \tilde{p}(x) log(\tilde{p}(x)) \sum_{y\in Y} p(y|x) + H(Y|X) \right) \\
&= argmax_{p\in P} \left(- \sum_{x\in X} \tilde{p}(x) log(\tilde{p}(x)) + H(Y|X) \right)
\end{aligned}
$$

$- \sum_{x\in X} \tilde{p}(x) log(\tilde{p}(x))$ е константа спрямо p, следователно:

$$= argmax_{p\in P} \left(H(Y|X) \right)$$

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:

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

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

$$\frac{\partial\left(\Lambda(p,\lambda)\right)}{\partial p(y_0|x_0)}$$

$$= \frac{\partial H(Y|X)}{\partial p(y_0|x_0)} + \frac{\partial\left(\lambda_i(E(p,h_i) - E(\tilde{p},h_i))\right)}{\partial p(y_0|x_0)} + \frac{\left(\sum_{x\in X}\mu_x \sum_{y\in Y} p(y|x) - 1\right)}{\partial p(y_0|x_0)}$$

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

$$= -\tilde{p}(x_0)log(p(y_0|x_0)) - \tilde{p}(x_0) + \sum_{i=1}^{K}\lambda_i\tilde{p}(x_0)h_i(x_0,y_0) + \mu_{x_0}$$

Искаме да я нулираме:

$$-\tilde{p}(x_0)log(p(y_0|x_0)) - \tilde{p}(x_0) + \sum_{i=1}^{K}\lambda_i\tilde{p}(x_0)h_i(x_0,y_0) + \mu_{x_0} = 0 \longleftrightarrow$$

$$\tilde{p}(x_0)log(p(y_0|x_0)) = \sum_{i=1}^{K}\lambda_i\tilde{p}(x_0)h_i(x_0,y_0) - \tilde{p}(x_0) + \mu_{x_0} \longleftrightarrow$$

$$ln(p(y_0|x_0)) = \sum_{i=1}^{K}\lambda_i h_i(x_0,y_0) - 1 + \frac{\mu_{x_0}}{\tilde{p}(x_0)} \longleftrightarrow$$

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

Производната по μ_{x_0} ни дава:

$$\sum_{y\in Y} p(y|x_0) = 1 \longleftrightarrow$$

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

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

$$\longleftrightarrow$$

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

Заместваме в Уравнение A.0.1:

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

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:

$$\sum_{i=1}^{n} p(x_i,y_i) f(z_i) \leq f\left(\sum_{i=1}^{n} p(x_i,y_i) z_i\right), \forall (z_1,\ldots,z_n) \in \mathbb{R}^n,$$

където f е вдлъбната. Ако f е строго вдлъбната (f' е строго намаляваща), равенство се достига, когато z_i е константа.

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

Тъй като логаритъмът е строго вдлъбната функция, равенство при неравенството на Йенсен се достига, когато $\frac{q(x,y)}{p(x,y)}$ е константа, тоест $\frac{q(x,y)}{p(x,y)} = 1 \iff p(x,y) = q(x,y)$ за произволно $(x,y) \in X \times Y$.

Твърдение 2. За всеки $p_1, p_2 \in P, q \in Q$ е изпълнено $\sum\limits_{(x,y)\in X\times Y} p_1(x,y)\, log(q(x,y)) = \sum\limits_{(x,y)\in X\times Y} p_2(x,y)\, log(q(x,y))$

Доказателство:

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

Тъй като $p_2 \in P$:

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

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

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

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:

$$
\begin{aligned}
&D(p,r) + D(r,q) \\
&= \sum_{(x,y)\in X\times Y} p(x,y)\, log\left(\frac{p(x,y)}{r(x,y)}\right) + \sum_{(x,y)\in X\times Y} r(x,y)\, log\left(\frac{r(x,y)}{q(x,y)}\right) \\
&= \sum_{(x,y)\in X\times Y} p(x,y)\, log(p(x,y)) - \sum_{(x,y)\in X\times Y} p(x,y)\, log(r(x,y)) + \\
&\quad \sum_{(x,y)\in X\times Y} r(x,y)\, log(r(x,y)) - \sum_{(x,y)\in X\times Y} r(x,y)\, log(q(x,y)) \\
&\overset{\text{Твърдение 1}}{\leq} \\
&\quad \sum_{(x,y)\in X\times Y} p(x,y)\, log(p(x,y)) - \cancel{\sum_{(x,y)\in X\times Y} p(x,y)\, log(r(x,y))} + \\
&\quad \cancel{\sum_{(x,y)\in X\times Y} p(x,y)\, log(r(x,y))} - \sum_{(x,y)\in X\times Y} r(x,y)\, log(q(x,y)) \\
&= \sum_{(x,y)\in X\times Y} p(x,y)\, log(p(x,y)) - \sum_{(x,y)\in X\times Y} r(x,y)\, log(q(x,y)) \\
&\overset{\text{Твърдение 1}}{\leq} \\
&\quad \sum_{(x,y)\in X\times Y} p(x,y)\, log(p(x,y)) - \sum_{(x,y)\in X\times Y} p(x,y)\, log(q(x,y)) \\
&= \sum_{(x,y)\in X\times Y} p(x,y)\, log\left(\frac{p(x,y)}{q(x,y)}\right) = D(p,q)
\end{aligned}
$$

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 $r = argmax_{p\in P} H(p)$, тоест ще покажем, че за всяко $p \in P : H(r) \geq H(p)$.

Нека u е равномерното разпределение върху $X \times Y$, тоест $u(x,y) = \dfrac{1}{n}, \forall (x,y) \in X \times Y$. Следователно $u \in Q$, защото можем да изберем $\pi = \dfrac{1}{n}$ и $\lambda_i = 0, \forall i \in \{1 \ldots K\}$

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

$$
\begin{aligned}
& D(p,u) = D(p,r) + D(r,u) \\
& D(p,u) \overset{\text{Твърдение 1}}{\geq} D(r,u) \\
& \sum_{(x,y)\in X\times Y} p(x,y)\, log\left(\frac{p(x,y)}{u(x,y)}\right) \geq \sum_{(x,y)\in X\times Y} r(x,y)\, log\left(\frac{r(x,y)}{u(x,y)}\right) \\
& -H(p) - \sum_{(x,y)\in X\times Y} p(x,y)\, log(u(x,y)) \geq -H(r) - \sum_{(x,y)\in X\times Y} r(x,y)\, log(u(x,y)) \\
& \overset{\text{Твърдение 2}}{\longleftrightarrow} \\
& -H(p) - \cancel{\sum_{(x,y)\in X\times Y} p(x,y)\, log(u(x,y))} \geq -H(r) - \cancel{\sum_{(x,y)\in X\times Y} p(x,y)\, log(u(x,y))} \\
& H(r) \geq H(p)
\end{aligned}
$$

Следователно $r = argmax_{p\in P} H(p)$

Сега нека видим защо r е единствено. Нека $r' = argmax_{p\in P} H(P)$. Тогава:

$$
\begin{aligned}
& H(r') = H(r) \longleftrightarrow D(r,u) = D(r',u) \\
& \text{но } D(r,u) = D(r,r') + D(r',u) \text{ по Твърдение 2} \\
& \implies \quad D(r,r') = 0 \\
& \overset{\text{Твърдение 1}}{\implies} \quad r = r'
\end{aligned}
$$

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

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

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

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

Дефинираме функцията $L(p)$:

$$
\begin{aligned}
L(p) = \sum_{(x,y)\in X\times Y} \tilde{p}(x,y)\, log(p(x,y)) &= \sum_{(x,y)\in D} \tilde{p}(x,y)\, log(p(x,y)) + \sum_{(x,y)\notin D} \tilde{p}(x,y)\, log(p(x,y)) \\
&= \sum_{(x,y)\in D} \tilde{p}(x,y)\, log(p(x,y)) + 0 \\
&= \sum_{(x,y)\in D} \frac{1}{n} log(p(x,y)) = C\, log\left(\widehat{L}_{X\times Y}(p)\right)
\end{aligned}
$$

Тоест L е пропорционална на логаритъм от правдоподобието. Тоест, ако p максимизира L, то максимизира и правдоодобието, и обратното.

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

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 $\tilde{p} \in P$.

Тогава от Твърдение 3 следва, че:

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

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

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

$$-\cancel{H(\tilde{p})} - L(q) \geq -\cancel{H(\tilde{p})} - L(r)$$

$$\longleftrightarrow L(r) \geq L(q)$$

Сега нека $r' = argmax_{q \in Q} L(q)$, тоест $L(r) = L(r') \implies D(\tilde{p}, r) = D(\tilde{p}, r')$.

Но по Твърдение 3, $D(\tilde{p}, r') = D(\tilde{p}, r) + D(r, r') \implies D(r, r') = 0 \overset{\text{Твърдение 1}}{\longleftrightarrow} r' = r$, следователно r е единствено.

От Твърдение 4 и Твърдение 5, че ако вземем разпределение от сечението на P и Q, то е единствено и е равно на $\hat{p} = argmax_{p \in P} H(p) = argmax_{q \in Q} L(q)$. Тъй като L е пропорционално на правдоподобието, за да намерим търсеното разпределение е достатъчно да максимизираме правдоподобието.