Дипломна работа

Диана Генева <dageneva@qtrp.org>

2019

Съдържание

1 Двойната звезда 2
1.1 Съчетаване чрез конкатенация на характеристичните вектори 2
1.1.1 Описание 2
1.1.2 Резултати 2
1.2 Съчетаване чрез максимизиране на ентропията 3
1.2.1 Описание 3

A Приложение за Максимизиране на ентропията 5

Глава 1

Двойната звезда

В предните раздели описахме получаването на класификатори за сигнали от реч и ЕЕГ поотделно. Сега въпросът е дали можем да съчетаем по някакъв начин данните или класификаторите с цел да получим по-добра класификационна точност. Разгледаните подходи са два. Първият е наивен и съчетава самите характеристични вектори, а вторият - вече получените класификатори.

1.1 Съчетаване чрез конкатенация на характеристичните вектори

1.1.1 Описание

Тук идеята е възможно най-проста - конкатенираме характеристичните вектори от речта и тези от ЕЕГ сигнала до получаване на нов вектор и трениране класификатора, описан в **??**. Единствената трудност е, че векторите за реч се взимат много по-често. За да има еднакъв брой вектори за двата сигнала, тези за речта се осредняват на всеки 200 ms.

1.1.2 Резултати

На таблица Таблица 1.1 е показана средната класификационна точност за всяка една от емоциите за двата класификатора поотделно и накрая за този, получен при конкатенацията на векторите. Всяка от Гаусовите смески на класификатора на реч има 8 гаусиани, докато този за ЕЕГ и полученият при конкатенация имат по 3 гаусиани. От таблицата се вижда, че този метод не довежда до подобрение. Това вероятно се дължи на малкото количество данни, тъй като векторите, с които работим, са $39 + 76 = 115$ мерни. Това означава, че за да видим повече проявления на характеристиките, ще са нужни много повече данни. Освен набавянето на допълнително данни, може да се намали пространството чрез факторен анализ, за да се подобри

обучението на класификатора.

Емоция	Само реч	Само ЕЕГ	Конкатенация
Гняв	85.00%	80.00%	85.00%
Щастие	75.00%	75.00%	50.00%
Неутрално	7.50%	97.50%	97.50%
Тъга	85.42%	87.50%	87.50%
Общо	**63.23%**	**85.00%**	**80.00%**

Таблица 1.1: Класификационна точност на класификатор за реч, класификатор за ЕЕГ и класификатор, получен при конкатенация на характеристичните вектори

1.2 Съчетаване чрез максимизиране на ентропията

1.2.1 Описание

Другият похват, който е приложен, използва модел, максимизиращ ентропията. Идеята е да намерим такова разпределение p, което се държи като емпиричното разпределение върху тренировъчните данни, но в същото време не прави допълнителни предположения извън тях. Тоест имайки входни данни от вида $\mathcal{D} = (x_1, y_1), \ldots, (x_n, y_n)$, където $x_i \in X = \mathbb{R}^n$ са характеристични вектори с етикети $y_i \in Y$, където $Y = \{1, \ldots, K\}$ представя множеството от търсените емоции, искаме да максимизираме ентропията:

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

Нека с h_1 бележим класификатора на реч, а с h_2 този на ЕЕГ. Тогава h_1, h_2 играят роля на характеристични функции, тъй като имат вида $h_i : X \times Y \mapsto [0, 1]$.

Очакването на всеки от класификаторите спрямо търсеното разпределение, трябва да съвпада с това на емпиричното. В A е показано, че търсеното $\hat{p}$ има вида $\hat{p}(x, y) = \pi exp\left(\lambda_1 h_1(x, y) + \lambda_2 h_2(x, y)\right)$, където π е нормализираща константа и има вида $\pi = \sum\limits_{y' \in Y} exp\left(p(x, y')\right)$. Тоест:

$$\hat{p}(y|x) = \frac{exp\left(\lambda_1 h_1(x, y) + \lambda_2 h_2(x, y)\right)}{\sum\limits_{y' \in Y} exp\left(\lambda_1 h_1(x, y') + \lambda_2 h_2(x, y')\right)}$$

В същото приложение е показано, че е достатъчно да максимизираме логаритъм от условното правдоподобие над $\mathcal{D}$, зададено с:

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

С $\#(x,y)$ бележим броя на срещанията на (x,y) в корпуса.

Тоест оптимизационната задача е:

$$\hat{\lambda}_1, \hat{\lambda}_2 = argmax_{\lambda_1,\lambda_2} \sum_{(x,y)\in X\times Y} \#(x,y) log(p(y|x))$$
$$= argmax_{\lambda_1,\lambda_2} \sum_{(x,y)\in\mathcal{D}} log\left(\frac{exp\left(\lambda_1 h_1(x,y) + \lambda_2 h_2(x,y)\right)}{\sum\limits_{y'\in Y} exp\left(\lambda_1 h_1(x,y') + \lambda_2 h_2(x,y')\right)}\right)$$

Намираме производните:

$$= \frac{\partial\left[\sum\limits_{(x,y)\in\mathcal{D}} log\left(\frac{exp\left(\lambda_1 h_1(x,y) + \lambda_2 h_2(x,y)\right)}{\sum\limits_{y'\in Y} exp\left(\lambda_1 h_1(x,y') + \lambda_2 h_2(x,y')\right)}\right)\right]}{\partial\lambda_1}$$
$$= \frac{\partial\left[\sum\limits_{(x,y)\in\mathcal{D}} \lambda_1 h_1(x,y) + \lambda_2 h_2(x,y) - \sum\limits_{(x,y)\in\mathcal{D}} log\left(\sum\limits_{y'\in Y} exp\left(\lambda_1 h_1(x,y') + \lambda_2 h_2(x,y')\right)\right)\right]}{\partial\lambda_1}$$
$$= \sum_{(x,y)\in\mathcal{D}} h_1(x,y) - \sum_{(x,y)\in\mathcal{D}} \frac{\sum\limits_{y'\in Y} exp\left(\lambda_1 h_1(x,y') + \lambda_2 h_2(x,y')\right) h_1(x,y')}{\sum\limits_{y'\in Y} exp\left(\lambda_1 h_1(x,y') + \lambda_2 h_2(x,y')\right)}$$

Съответно:

$$\frac{\partial\Lambda(\lambda_1,\lambda_2)}{\partial\lambda_2} = \sum_{(x,y)\in\mathcal{D}} h_2(x,y) - \sum_{(x,y)\in\mathcal{D}} \frac{\sum\limits_{y'\in Y} exp\left(\lambda_1 h_1(x,y') + \lambda_2 h_2(x,y')\right) h_2(x,y')}{\sum\limits_{y'\in Y} exp\left(\lambda_1 h_1(x,y') + \lambda_2 h_2(x,y')\right)}$$

1. Избираме $\lambda_1^0 = \lambda_2^0 = 0.5$ и пресмятаме $\hat{L}^0_{\mathcal{D}}(Y|X)$
2. За всяка стъпка t се прави следното:
   (а) Намираме $\hat{\lambda}_1$ и $\hat{\lambda}_2$
   (б) Вървим по градиента:

   $$\lambda_1^t = \lambda_1^{t-1} + C\hat{\lambda}_1$$
   $$\lambda_2^t = \lambda_2^{t-1} + C\hat{\lambda}_2$$

   (в) Пресмята се $\hat{L}^t_{\mathcal{D}}(Y|X)$. Ако $\hat{L}^t_{\mathcal{D}}(Y|X) \leq \hat{L}^{t-1}_{\mathcal{D}}(Y|X)$ (или $t > 200$), алгоритъмът приключва с отговор $\lambda_1^{t-1}, \lambda_2^{t-1}$

При получените по горния начин λ_1 и λ_2 и подаден вектор $x \in X$, новополученият класификатор работи по следния начин:

$$H(x) = argmax_{y\in Y}\left(\lambda_1 h_1(x,y) + \lambda_2 h_2(x,y)\right)$$

Приложение А

Приложение за Максимизиране на ентропията

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

Търсим това разпределение p, което приближава разпределението, генерирало 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:

$\forall (x, y) \in X \times Y : \tilde{p}(x, y) = \frac{\#(x, y)}{n}$, 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 $h_i : X \times Y \to [0, 1]$.

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:

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

Ако означим:

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

тогава искаме да намерим

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

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

$$= argmax_{p\in P} H_p(Y|X)$$

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 с определени свойства, задачата ще има вида:

$$\begin{aligned}
\Lambda(p,\tau,\lambda,\mu) = H_p(Y|X) &+ \sum_{x\in X} \tau_x \left[\sum_{y\in Y} (p(x,y) - \tilde{p}(x)p(y|x)) \right] \\
&+ \sum_{i=1}^{K} \lambda_i (E(p,h_i) - E(\tilde{p},h_i)) + \sum_{x\in X} \mu_x \sum_{y\in Y} p(y|x) - 1
\end{aligned}$$

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

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

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

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

Искаме да я нулираме:

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

Производната по μ_{x_0} ни дава:

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

Заместваме в Уравнение A.0.1:

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

Следователно вида на търсеното $\hat{p}$ е $\hat{p}(x,y) = \pi\prod\limits_{i=1}^{K} e^{\lambda_i h_i(x,y)}$, 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.

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

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

та, тоест:

$$q(x,y) = Cp(x,y)$$
$$\sum_{(x,y)\in\mathcal{D}} q(x,y) = \sum_{(x,y)\in\mathcal{D}} Cp(x,y)$$
$$\longleftrightarrow C = 1$$
$$\longleftrightarrow p(x,y) = q(x,y) \forall (x,y) \in X \times Y$$

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:

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

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

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

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

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

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:

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

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

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

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

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

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

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

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

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

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 $r = argmax_{p\in P} H_p(X,Y)$, 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

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

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

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

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

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

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

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

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

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

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

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

$$\implies \quad D(r,r') = 0$$

$$\overset{\text{Твърдение 1}}{\implies} \quad r = r'$$

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 $r = argmax_{q\in Q} L(q)$

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:

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

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

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

$$\cancel{-H_{\tilde{p}}(X,Y)} - L(q) \geq \cancel{-H_{\tilde{p}}(X,Y)} - L(r)$$

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

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.

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:

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

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

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

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