The Bivariate 2-Poisson model for IR

Giambattista Amati[1] and Giorgio Gambosi[2]

[1] Fondazione Ugo Bordoni, Rome, Italy gba@fub.it
[2] Enterprise Engineering Department of University of Tor Vergata, Rome, Italy
giorgio.gambosi@uniroma2.it

1 Introduction

Harter's 2-Poisson model of Information Retrieval is a univariate model of the raw term frequencies, that does not condition the probabilities on document length [2]. A bivariate stochastic model is thus introduced to extend Harter's 2-Poisson model, by conditioning the term frequencies of the document to the document length. We assume Harter's hypothesis: the higher the probability $f(X = x|L = l)$ of the term frequency $X = x$ is in a document of length l, the more relevant that document is. The new generalization of the 2-Poisson model has 5 parameters that are learned term by term through the EM algorithm over term frequencies data.

We explore the following frameworks:

- We assume that the observation $\langle x, l\rangle$ is generated by a mixture of k Bivariate Poisson (k-BP) distributions (with $k \geq 2$) with or without some conditions on the form for the marginal of the document length, that can reduce the complexity of the model. We here reduce for the sake of simplicity to $k = 2$. In the case of the 2-BP we also assume the hypothesis that the marginal distribution of l is a Poisson. The elite set is generated by the BP of the mixture with higher value for the mean of term frequencies, λ_1.
- The covariate variable Z_3 of length and term frequency λ_3 could be learned from covariance [3, page 103]. Instead, we here consider Z_3 a latent random variable which is learned by extending the EM algorithm in a standard way.
- Our plan is to compare the effectiveness of the bivariate 2-Poisson model with respect to standard models of IR, and in particular with some additional baselines that are obtained in our framework as follows:
  - applying the Double Poisson Model, which is the 2-BP with the marginal distributions that are independent.
  - Reducing to the univariate case (standard 2-Poisson model) by normalizing the term frequency x to a smoothed value $\mathbf{tfn}$. For example, we can use the Dirichlet smoothing:

    $$\mathbf{tfn} = \frac{x + \mu \cdot \hat{p}}{l + \mu} \cdot \mu'$$

    where μ and μ' are parameters and $\hat{p}$ is the term prior.

2 The Bivariate 2-Poisson distribution

In order to define the bivariate 2-Poisson model we need first to remind the definition of a bivariate Poisson model, that can be introduced in several ways, for example as limit of a bivariate binomial, as a convolution of three univariate Poisson distributions, as a compounding of a Poisson with a bivariate binomial. We find that the trivariate reduction method of the convolution more convenient to easily extend Harter's 2-Poisson model to the bivariate case.

Let us consider the random variables Z_1, Z_2, Z_3 distributed according to Poisson distributions $P(\lambda_i)$, that is:

$$p(Z_i = x|\lambda_i) = e^{-\lambda_i}\frac{\lambda_i^x}{x!}$$

and the random variables $X = Z_1 + Z_3$ e $Y = Z_2 + Z_3$ distributed according to a bivariate Poisson distribution, $BP(\Lambda)$, where $\Lambda = (\lambda_1, \lambda_2, \lambda_3)$:

$$p(X = x, Y = y|\Lambda) = e^{-(\lambda_1+\lambda_2+\lambda_3)}\frac{\lambda_1^x}{x!}\frac{\lambda_2^y}{y!}\sum_{i=0}^{min(x,y)}\binom{x}{i}\binom{y}{i}i!\left(\frac{\lambda_3}{\lambda_1\lambda_2}\right)^i$$

The corresponding marginal distributions turn out to be Poisson

$$p(X = x|\Lambda) = \sum_{y=0}^{\infty} p(X = x, Y = y|\Lambda) = P(\lambda_1 + \lambda_3)$$
$$p(Y = y|\Lambda) = \sum_{x=0}^{\infty} p(X = x, Y = y|\Lambda) = P(\lambda_2 + \lambda_3)$$

with covariance $Cov(X, Y) = \lambda_3$.

Let us now consider the mixture $2BP(\Lambda_1, \Lambda_2, \alpha)$, where $\Lambda_1 = (\lambda_1^1, \lambda_2^1, \lambda_3^1)$ and $\Lambda_2 = (\lambda_1^2, \lambda_2^2, \lambda_3^2)$, of two bivariate Poisson distributions

$$p(x, y|\Lambda_1, \Lambda_2, \alpha) = \alpha \cdot BP(\Lambda_1) + (1 - \alpha) \cdot BP(\Lambda_2)$$

The corresponding marginal distributions are 2-Poisson

$$p(x|\Lambda_1, \Lambda_2, \alpha) = \alpha \cdot P(\lambda_1^1 + \lambda_3^1) + (1 - \alpha) \cdot P(\lambda_1^2 + \lambda_3^2) = 2P(\lambda_1^1 + \lambda_3^1, \lambda_1^2 + \lambda_3^2, \alpha)$$
$$p(y|\Lambda_1, \Lambda_2, \alpha) = \alpha \cdot P(\lambda_2^1 + \lambda_3^1) + (1 - \alpha) \cdot P(\lambda_2^2 + \lambda_3^2) = 2P(\lambda_2^1 + \lambda_3^1, \lambda_2^2 + \lambda_3^2, \alpha)$$

In our case, we consider the random variables *x*, number of occurrences of the term in the document, and $L^- = l - x$, document length out of the term occurrences, and set $X = x$ and $Y = L^- = l - x$ (hence, Y could possibly be 0): as a consequence, we have $x = X = Z_1 + Z_3$, $L^- = Y = Z_2 + Z_3$, and $l = X + Y = Z_1 + Z_2 + 2Z_3$.

Moreover, we want *x* to be distributed as a *2-Poisson* and L^- to be distributed as a *Poisson*. By assuming $\lambda_2^1 = \lambda_2^2 = \lambda_2$ and $\lambda_3^1 = \lambda_3^2 = \lambda_3$ we obtain

$$p(x|\Lambda_1, \Lambda_2, \alpha) = \alpha \cdot P(\lambda_1^1 + \lambda_3) + (1 - \alpha) \cdot P(\lambda_1^2 + \lambda_3) = 2P(\lambda_1^1 + \lambda_3, \lambda_1^2 + \lambda_3)$$
$$p(L^-|\Lambda_1, \Lambda_2, \alpha) = \alpha \cdot P(\lambda_2 + \lambda_3) + (1 - \alpha) \cdot P(\lambda_2 + \lambda_3) = P(\lambda_2 + \lambda_3)$$

This implies that, apart from α, we assume five latent variables in the model, $Z_1^1, Z_1^2, Z_2, Z_3, W$ each Z is Poisson distributed with parameters $\lambda_1^1, \lambda_1^2, \lambda_2, \lambda_3$ respectively and W is a binary random variable Bernoulli distributed with parameter α. The resulting bivariate distribution is

$$\begin{aligned} p(x, L^-|\Lambda_1, \Lambda_2, \alpha) &= \alpha \cdot p_1(x, L^-|\lambda_1, \lambda_2, \lambda_3) + (1-\alpha) \cdot p_2(x, L^-|\lambda_1^2, \lambda_2, \lambda_3) \\ &= \alpha \cdot BP(\lambda_1^1, \lambda_2, \lambda_3) + (1-\alpha) \cdot BP(\lambda_1^2, \lambda_2, \lambda_3) \end{aligned}$$

3 EM algorithm for the Bivariate Poisson

Given a set of observations $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, with $\boldsymbol{x}_i = (x_i, l_i)$, we wish to apply maximum likelihood to estimate the set of parameters Λ of a bivariate Poisson distribution $p(\boldsymbol{x}|\Theta)$ fitting such data. We wish to derive the value of Θ by maximizing the log-likelihood, that is computing

$$\Theta^* = \arg\max_{\Theta} \log \mathcal{L}(\Theta|\mathcal{X}) = \arg\max_{\Theta} \log \prod_{i=1}^{n} p(\boldsymbol{x}_i|\Theta)$$

In our case (see also [1]), we are interested to a mixture of 2 Bivariate Poisson with latent variables Z_1^1, Z_1^2, Z_2, Z_3, since with respect to the general case we have now $Z_2^1 = Z_2^2 = Z_2$ and $Z_3^1 = Z_3^2 = Z_3$. Then, for each observed pair of values $\boldsymbol{x}_i = (x_i, l_i)$, $w_i = 1$ if $\boldsymbol{x}_i$ is generated by the first component, and $w_i = 2$ if generated by the second one. Accordingly:

$$\boldsymbol{z}_i = (z_{i1}^1, z_{i1}^2, z_{i2}, z_{i3}) \text{ are such that } x_i = \begin{cases} z_{i1}^1 + z_{i3} \text{ if } w_i = 1 \\ z_{i1}^2 + z_{i3} \text{ if } w_i = 2 \end{cases}$$
$$\text{and } l_i = z_{i2} + z_{i3}$$

EM algorithm requires, in our case, to consider the complete dataset

$$(\mathcal{X}, \mathcal{Z}) = \{(\boldsymbol{x}_1, \boldsymbol{z}_1, w_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{z}_n, w_n)\}$$

and the set of parameters is $\Theta = \Lambda_1 \cup \Lambda_2 \cup \{\alpha\}$, with $\Lambda_k = \left\{\lambda_1^k, \lambda_2, \lambda_3\right\}$. Let also $\Lambda = \Lambda_1 \cup \Lambda_2$.

3.1 Maximization

Let us consider the k-th M-step for Θ. We can show the following estimates:

$$\alpha^{(k)} = \frac{1}{n} \sum_{i=1}^{n} p_i^{(k-1)} \text{ where } p_i^{(k-1)} = \frac{\alpha^{(k-1)} p(\boldsymbol{x}_i|\Lambda_1^{k-1})}{\alpha^{(k-1)} p(\boldsymbol{x}_i|\Lambda_1^{(k-1)}) + (1-\alpha^{(k-1)}) p(\boldsymbol{x}_i|\Lambda_2^{(k-1)})}$$

and p is the Bivariate Poisson with parameters Λ_i, and

$$\lambda_1^{1(k)} = \frac{\sum_{i=1}^{n} b_{1i}^{1\ (k-1)} p_i^{(k-1)}}{\sum_{i=1}^{n} p_i^{(k-1)}} \qquad \lambda_1^{2(k)} = \frac{\sum_{i=1}^{n} b_{1i}^{2\ (k-1)} (1 - p_i^{(k-1)})}{\sum_{i=1}^{n} (1 - p_i^{(k-1)})}$$

$$\lambda_2^{(k)} = \frac{1}{n} \sum_{i=1}^{n} b_{2i}{}^{(k-1)} \qquad \lambda_3^{(k)} = \frac{1}{n} \sum_{i=1}^{n} b_{3i}{}^{(k-1)}$$

where ${b_{hi}^{j}}^{(k)} = E[Z_h^j | W = j, \boldsymbol{x}_i, \Lambda^{(k)}]$ and ${b_{hi}}^{(k)} = E[Z_h | \boldsymbol{x}_i, \Lambda^{(k)}]$ with $h = 1, 2, 3$.

3.2 Expectation

We can show that the expectations ${b_{1i}^{j}}^{(k)}$ and ${b_{hi}}^{(k)}$ are:

$$
\begin{aligned}
{b_{3i}}^{(k)} &= \sum_{r=0}^{\min(x_i, l_i)} r \cdot p(Z_3 = r | \boldsymbol{x}_i, \Lambda) \text{ where } \Lambda^{(k)} = \Lambda_1^{(k)} \cup \Lambda_2^{(k)} \\
&= \sum_{r=0}^{\min(x_i, l_i)} r \frac{(1-\alpha) p(Z_3 = r, \boldsymbol{x}_i | W = 2, \Lambda^{(k)}) + \alpha p(Z_3 = r, \boldsymbol{x}_i | W = 1, \Lambda^{(k)})}{p(\boldsymbol{x}_i | \Lambda^{(k)})} \\
{b_{1i}^{1}}^{(k)} &= E[X_1 | W = 1, \boldsymbol{x}_i] - E[Z_3 | W = 1, \boldsymbol{x}_i] = x_i - {b_{3i}^{1}}^{(k)} \\
{b_{1i}^{2}}^{(k)} &= E[X | W = 2, \boldsymbol{x}_i, \Lambda^{(k)}] - E[Z_3 | W = 2, \boldsymbol{x}_i, \Lambda^{(k)}] = x_i - {b_{3i}^{2}}^{(k)} \\
{b_{2i}}^{(k)} &= E[Y | \boldsymbol{x}_i, \Lambda^{(k)}] - E[Z_3 | \boldsymbol{x}_i, \Lambda^{(k)}] = l_i - {b_{3i}}^{(k)}
\end{aligned}
$$

where

$$
p(Z_3 = r, \boldsymbol{x}_i | W = j, \Lambda^{(k)}) = P_0(r | {\lambda_3}^{(k)}) \cdot P_0(x - r | {\lambda_1^{j}}^{(k)}) \cdot P_0(l - r | {\lambda_2}^{(k)})
$$

and P_0 is the univariate Poisson, $p(\boldsymbol{x}_i | \Lambda^{(k)})$ is the mixture of the bivariate Poisson. Efficient implementation of the bivariate Poisson through recursion can be found in [4].

4 Conclusions

We have implemented the EM algorithm for the univariate 2-Poisson and we are currently extending the implementation to the bivariate case.

The implementation will be soon available together with the results of the experimentation at the web site `http://tinyurl.com/cfcm8ma`.

References

1. BRIJS, T., KARLIS, D., SWINNEN, G., VANHOOF, K., WETS, G., AND MANCHANDA, P. A multivariate poisson mixture model for marketing applications. *Statistica Neerlandica 58*, 3 (2004), 322–348.
2. HARTER, S. P. A probabilistic approach to automatic keyword indexing. part I: On the distribution of specialty words words in a technical literature. *Journal of the ASIS 26* (1975), 197–216.
3. KOCHERLAKOTA, S., AND KOCHERLAKOTA, K. *Bivariate discrete distributions*. Marcel Dekker Inc., New York, 1992.
4. TSIAMYRTZIS, P., AND KARLIS, D. Strategies for efficient computation of multivariate poisson probabilities. *Communications in Statistics-Simulation and Computation 33*, 2 (2004), 271–292.