Notes on Bayesian Information Criterion Calculation for X-Means Clustering

October 2, 2012

This documents attempts to derive and explain the Bayesian Information Criterion (BIC) as written in [Pelleg and Moore(2000)].

First an explanation of the notation: Indices i, j, ... will be used for points. n, m, ... will be used to index over cluster. Some terms like μ and D when subscripted by n refer to the value for cluster n and when subscripted by a variable in parentheses, e.g. (i), refer to the value for the cluster to which data point (e.g. x_i) belongs.

The Pelleg and Moore's x-means algorithm is built around the "identical spherical assumption". The data is modeled by n gaussians, each with identical variance σ and differing positions μ_n.

Given that model, the probability of a data point i located at x_i is

$$P(x_i) = \sum_{n=1}^{K} \underbrace{P(x_i \in D_n)}_{\substack{\text{prob. data point i is} \\ \text{an element of cluster } D_n}} \cdot \underbrace{P(x_i | x_i \in D_n)}_{\substack{\text{prob. element i is positioned} \\ \text{at } x_i \text{ if in cluster } D_n}} \tag{1}$$

When computed under the maximum likelihood, the first term, $P(x_i \in D_n)$, is

$$P(x_i \in D_n) = \frac{R_n}{R} = \frac{R_{(i)}}{R} \tag{2}$$

where R_n is the number of elements in the cluster.[1]

The second term is a multivariate gaussian distribution centered at μ_n and with variance σ^2,

$$P(x_i | x_i \in D_n) = \frac{1}{(2\pi\sigma^2)^{M/2}} \exp\left(-\frac{1}{2\sigma^2}\|x_i - \mu_n\|^2\right) \tag{3}$$

Combining (1), (2), and (3)

$$P(x_i) = \frac{R_n}{R} \frac{1}{(2\pi\sigma^2)^{M/2}} \exp\left(-\frac{1}{2\sigma^2}\|x_i - \mu_{(i)}\|^2\right) \tag{4}$$

[1] A fuller derivation can be found in [Daume III(2009)]

Converting to log-likelihoods,[2]

$$
\begin{aligned}
l(D) &= \log \prod_i P(x_i) \\
&= \sum_i \log P(x_i) \\
&= \sum_i \left(\log \frac{R_n}{R} + \log \left(\frac{1}{(2\pi\sigma^2)^{M/2}} \right) - \frac{1}{2\sigma^2} \|x_i - \mu_{(i)}\|^2 \right) \qquad (5) \\
&= \sum_{n=1}^{K} \sum_{x_i \in D_n} \left(\log \frac{R_n}{R} + \log \left(\frac{1}{(2\pi\sigma^2)^{M/2}} \right) - \frac{1}{2\sigma^2} \|x_i - \mu_{(i)}\|^2 \right) \qquad (6) \\
&= \sum_{n=1}^{K} \left[R_n \left(\log \frac{R_n}{R} - \frac{M}{2} \log \left(2\pi\sigma^2 \right) \right) - \frac{1}{2\sigma^2} \sum_{x_i \in D_n} \|x_i - \mu_{(i)}\|^2 \right] \qquad (7)
\end{aligned}
$$

In [Pelleg and Moore(2000)] section 3.2, the authors incorrectly state the maximum likelihood estimate for the variance for identical spherical Gaussians is

$$\hat{\sigma}^2 = \frac{1}{R-K} \sum_i \|x_i - \mu_{(i)}\|^2 \qquad (8)$$

where the hat above the σ is used to denote the estimator. The actual maximum likelihood estimate for the variance is determined by setting $\frac{\partial l(D)}{\partial \sigma}$ to 0.

$$
\begin{aligned}
\frac{\partial l(D)}{\partial \sigma} &= \sum_{n=1}^{K} \left[-\frac{MR_n}{\sigma} + \frac{1}{\sigma^3} \sum_{x_i \in D_n} \|x_i - \mu_{(i)}\|^2 \right] \\
&= \sigma^{-3} \sum_{n=1}^{K} \left[-MR_n\sigma^2 + \sum_{x_i \in D_n} \|x_i - \mu_{(i)}\|^2 \right] \\
&= \sigma^{-3} \left[-MR\sigma^2 + \sum_i \|x_i - \mu_{(i)}\|^2 \right] \qquad (9)
\end{aligned}
$$

which is 0 when

$$\hat{\sigma}^2 = \frac{1}{MR} \sum_i \|x_i - \mu_{(i)}\|^2 \qquad (10)$$

Again using a hat to denote the variance estimate.[3]

Both estimates for the variance are proportional to the sum of the distances from each data point to the nearest centroid. This sum can be broken down into the sum by cluster,

[2] The logarithms in all equations in this and [Pelleg and Moore(2000)] are base-e. Different bases can be used, but they would lead to additional constant factors.

[3] A more detailed derivation of the maximum likelihood estimate is given in [Daume III(2009)].

$$\sum_i \|x_i - \mu_{(i)}\|^2 = \sum_n \sum_{i \in D_n} \|x_i - \mu(i)\|^2 \tag{11}$$

Each individual cluster is a spherical gaussian, which means the unbiased estimator of the variance is

$$\hat{\sigma}_n^2 = \frac{1}{M(R_n - 1)} \sum_{i \in D_n} \|x_i - \mu_n\|^2 \tag{12}$$

Substitution of (12) into (11) yields

$$\sum_i \|x_i - \mu_{(i)}\|^2 = M \sum_n (R_n - 1)\, \hat{\sigma}_j^2 \tag{13}$$

Assuming the "identical spherical assumption" means all the Gaussians have the same variance

$$\hat{\sigma}_j^2 = \hat{\sigma}^2 \tag{14}$$

then (13) becomes

$$\begin{aligned} \sum_i \|x_i - \mu_{(i)}\|^2 &= M \left(\sum_n (R_n) - \sum_n (1) \right) \hat{\sigma}^2 \\ &= M\,(R - K)\,\hat{\sigma}^2 \end{aligned} \tag{15}$$

or,

$$\hat{\sigma}^2 = \frac{1}{M(R - K)} \sum_i \|x_i - \mu_{(i)}\|^2 \tag{16}$$

which is the same as (8) up to a factor of M. So Pelleg and Moore are using the unbiased estimator for each cluster. Now using the maximum likelihood assumption from (13),

$$\hat{l}(D) = \sum_{n=1}^{K} \left[R_n \left(\log \frac{R_n}{R} - \frac{M}{2} \log\left(2\pi\hat{\sigma}^2\right) \right) - \frac{1}{2\hat{\sigma}^2} M\,(R_n - 1)\,\hat{\sigma}^2 \right] \tag{17}$$

$$= \sum_{n=1}^{K} \left[R_n \log R_n - R_n \log R - \frac{R_n M}{2} \log\left(2\pi\hat{\sigma}^2\right) - \frac{1}{2} M\,(R_n - 1) \right] \tag{18}$$

Using $\sum_{n=1}^{K} R_n = R$,

$$\hat{l}(D) = \sum_{n=1}^{K} R_n \log R_n - R \log R - \frac{RM}{2} \log\left(2\pi\hat{\sigma}^2\right) - \frac{M}{2}\,(R - K) \tag{19}$$

Now, consider two hypothesis, ϕ_1 and ϕ_2 (denoted M_j in the article, but I want to be clear this has nothing to do with the number of dimensions M).

K, R_n, and σ are all functions of the models, ϕ. In our case, ϕ_1 is the clustering result after minimizing with a fixed number of clusters, and ϕ_2 is the result after splitting one of the clusters into two and doing k-means only over that original cluster. Adding more clusters will decrease the variance (and increase the likelihood). (Or it could leave it unchanged, if the additional cluster is empty.) To avoid constantly adding centroids, information criteria are used. The main ones are the Akaike information criterion, and the one used in [Pelleg and Moore(2000)], the Bayesian information criterion.[4]

$$BIC(\phi) = \hat{l}_\phi(D) - \frac{p_\phi}{2} \cdot \log R \tag{20}$$

Both work by penalizing the likelihood more as the complexity of the model (e.g. the number of parameters) increases. So, instead of checking the likelihoods only, hypothesis ϕ_2 is better than hypothesis ϕ_1 if $BIC(\phi_2) > BIC(\phi_1)$.

For clarity, the maximum likelihood can be broken into the sum of two parts: a model-dependent part and a model-independent part.

$$\hat{l}(D,\phi) = \sum_{n=1}^{K(\phi)} R_n(\phi)\log R_n(\phi) - R\log R - \frac{RM}{2}\log\left(2\pi\hat{\sigma}(\phi)^2\right) - \frac{M}{2}\left(R - K(\phi)\right) \tag{21}$$

wrong sign

$$= \left[\sum_{n=1}^{K(\phi)} R_n(\phi)\log R_n(\phi) - \frac{MK(\phi)}{2} - \frac{RM}{2}\log\left(\hat{\sigma}(\phi)^2\right)\right] - \left[\frac{MR}{2} + R\log R + \frac{RM}{2}\log 2\pi\right] \tag{22}$$

$$= \hat{l}_{\text{model-dependent}}(D,\phi) + \hat{l}_{\text{model-independent}}(D) \tag{23}$$

Using the definition of the BIC and eliminating the model-independent terms,

$$\hat{l}(D,\phi_2) - \frac{p_{\phi_2}}{2}\log R > \hat{l}(D,\phi_1) - \frac{p_{\phi_1}}{2}\log R$$
$$\hat{l}_{\text{model-dependent}}(D,\phi_2) - \frac{p_{\phi_2}}{2}\log R > \hat{l}_{\text{model-dependent}}(D,\phi_1) - \frac{p_{\phi_1}}{2}\log R \tag{24}$$

[4] Actually this is the Schwarz information criterion used as an approximation for a Bayes factor, only valid when the number of points under consideration is large. See [Kass and Wasserman(1995)] for details.

This gives a final test of the form

$$\left[\sum_{n=1}^{K(\phi_2)} R_n(\phi_2)\log R_n(\phi_2) - \frac{MK(\phi_2)}{2} - \frac{RM}{2}\log\left(\hat{\sigma}(\phi_2)^2\right)\right] - \frac{p_{\phi_2}}{2}\log R$$
$$> \left[\sum_{n=1}^{K(\phi_1)} R_n(\phi_1)\log R_n(\phi_1) - \frac{MK(\phi_1)}{2} - \frac{RM}{2}\log\left(\hat{\sigma}(\phi_1)^2\right)\right] - \frac{p_{\phi_1}}{2}\log R \tag{25}$$

If this inequality holds, ϕ_2 is considered a better model the ϕ_1.

References

[Daume III(2009)] Hal Daume III. Gaussian Mixture Models, 2009. URL `http://www.cs.utah.edu/~hal/courses/2009F_ML/out/17-gmm.pdf`.

[Kass and Wasserman(1995)] Robert E Kass and Larry Wasserman. A Reference Bayesian Test for Nested and Its Relationship to the Schwarz Criterion. *Journal of the American Statistical Association*, 90(431), 1995.

[Pelleg and Moore(2000)] Dan Pelleg and Andrew Moore. X-means: Extending K-means with efficient estimation of the number of clusters. *Proceedings of the Seventeenth International*, 2000. URL `http://staff.utia.cas.cz/nagy/skola/Projekty/Classification/Xmeans.pdf`.