CSE 527 Lecture Note #13

Lecturer: Prof. Larry Ruzzo
Notes: Daehyun Baek
November 20, 2001

Bayesian Model Selection (BMS)

BMS is another way of doing notion of hypothesis A vs. hypothesis B to explain the data. The Bayesian information criterion(BIC) score is similar to chi-square value in the traditional hypothesis testing.

Bayesian Information Criterion (BIC)

: This is also an approximately statistical approach, not rigorously defined.

$$BIC = 2\log P(D \mid M, \hat{\theta}) - d\log n$$

where D: Observed data,
M: Model (The Model is actually a family of models with unit variance and unknown mean therefore $\hat{\theta}$ is needed),
$\hat{\theta}$: The maximum likelihood estimator (LME) of parameters in the model,
d : The number of free parameters, and
n : The number of data points.

Note: BIC score is good for comparing models. A model with a higher BIC is a better model, since if data fits well to the model, the log likelihood would be higher.

General model → Mixture model { BIC with multiple models
BIC with multiple parameter estimators

Multiple parameters get higher likelihood but it is penalized by the second term ($d\log n$). If we mix two Gaussian parameters, $\hat{\theta}$ becomes a pair. The likelihood will increase if there're more model parameters. The second term, $d\log n$ denotes a penalty term that also increases if more parameters are used. Intuitively, more data points need higher precision so it should be penalized.

Minimum Description Length (MDL) problem
Idea: Simpler is better. → Completely heuristic approach.
Let's define M as a model, θ as a parameter (vector), and D as observed data.

$$P(\theta \mid D) = \frac{P(D \mid \theta)P(\theta)}{P(D)}$$: Bayes' theorem

This is based on the 2-stage experiment. 1) Pick θ and 2) Draw data points according to θ. In this experiment, the above equation is a rigorous description of the model and $P(\theta \mid D)$ denotes the posterior probability. It gives us a way to update the prior $P(\theta)$ after seeing the data points D.

Idea: $P(\theta)$ is the prior probability, which is subjective. Based on subjective belief, we can estimate $P(\theta \mid D)$, the distribution of θ given the data.

By the law of total probability,

(1) $P(D) = \int_{\theta} P(D \mid \theta) P(\theta) d\theta$: Probability of data given the parameter

Suppose we have M_1, M_2, θ_1, and θ_2.

(2) $P(M_1 \mid D) = \dfrac{P(D \mid M_1) P(M_1)}{\sum_{i=1}^{2} P(D \mid M_i) P(M_i)}$: $P(M_1)$ is the prior probability here.

Notice that $P(\theta \mid D)$ is independent of θ. From (1),

(3) $P(D \mid M_i) = \int_{\theta} P(D \mid M_i \theta_i) P(\theta_i) d\theta_i$: Integrated likelihood over the parameter

Now, $\dfrac{P(M_2 \mid D)}{P(M_1 \mid D)}$: Posterior odds ratio

The odds ratio needed to establish data from which the model came becomes the following equation.

$$\underbrace{\frac{P(M_2 \mid D)}{P(M_1 \mid D)}}_{\text{Posterior Odds}} = \underbrace{\frac{P(D \mid M_2)}{P(D \mid M_1)}}_{\text{Bayes factor}} \underbrace{\frac{P(M_2)}{P(M_1)}}_{\text{Prior Odds}}$$

Example of prior odds: Situation in a bath with mixed fair and biased coins.
Note that Bayes factor explains the favors of probability of data to the given models.

Thus, the goal here is to determine the posterior odds that is updating the prior odds after seeing the data as the data explains the model. Eventually, we want to estimate (3) because $BIC \cong 2 \log P(D \mid M)$. However, the integral in (3) often is unsolvable in practice. So, let's define $g(\theta)$ as follows.

$$g(\theta) = \log P(D \mid \theta)P(\theta)$$

By Taylor series expansion, $g(\theta)$ can be expanded as the below equation.

(4) $$g(\theta) = g(\tilde{\theta}) + (\tilde{\theta} - \theta)g'(\tilde{\theta}) + \frac{1}{2}(\tilde{\theta} - \theta)^2 g''(\tilde{\theta}) + \Lambda$$

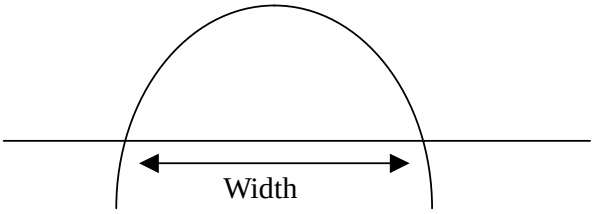
At the mode, $g'(\tilde{\theta}) = 0$. So the second term in (4) goes away. Therefore, (4) can be approximately simplified as follows.

$$g(\theta) \cong g(\tilde{\theta}) + \frac{1}{2}(\tilde{\theta} - \theta)^2 g''(\tilde{\theta})$$

$$P(D \mid M) = \int e^{g(\theta)} d\theta \cong \int e^{g(\tilde{\theta}) + \frac{1}{2}(\tilde{\theta} - \theta)^2 g''(\tilde{\theta})} d\theta = e^{g(\tilde{\theta})} \int e^{\frac{1}{2}(\tilde{\theta} - \theta)^2 g''(\tilde{\theta})} d\theta$$

$$\cong P(D \mid \hat{\theta})P(\hat{\theta})(normal)$$

Hint: The posterior mode would be close to the peak. The second derivative would be negative since there's a convex near the peak. The width is related to one of variances in this distribution.



Therefore, the second derivative term can be simplified. → $g''(\tilde{\theta}) \cong -\frac{1}{\sigma^2}$

$$\log P(D \mid M) \approx \log P(D \mid \hat{\theta}) + \log P(\hat{\theta}) + \frac{1}{2}\log\frac{2\pi}{\sigma^2}$$

M (By approximation not described here)

$$\approx 2\log P(D \mid \hat{\theta}) - d\log P(n) + O(1)$$

$$\therefore BIC \approx 2\log P(D \mid \hat{\theta}) - d\log P(n)$$

} For multiple parameters

This concluded equation doesn't have any prior probability term, which was omitted

during the approximations by the assumption of unit information prior. The unit information prior is a proper assumption, when we are not sure about the prior probability.

Interpretation of BIC values: BIC difference of 10 favors one model over the other by the factor of about 150.

Approximations in this approach

1) Taylor series third and higher order expansions are ignored.
2) In the posterior distributions, the mode is observed to be near local maxima.
3) O(1) doesn't go to zero if data set becomes bigger
4) $g''(\tilde{\theta}) \cong -\frac{1}{\sigma^2}$

Covariance Model

$\sum_k = \lambda_k D_k A_k D_k^T$: Covariance matrix for k_{th} cluster

λ_k , D_k , and A_k explain volume, orientation, and shape of the distribution, respectively.

More flexible, but more parameters

Equal volume spherical model (EI): similar to k-means model

$\sum_k \lambda I$

Unequal volume spherical model (VI)

$\sum_k \lambda_k I$

Diagonal model: cluster shapes parallel to axes

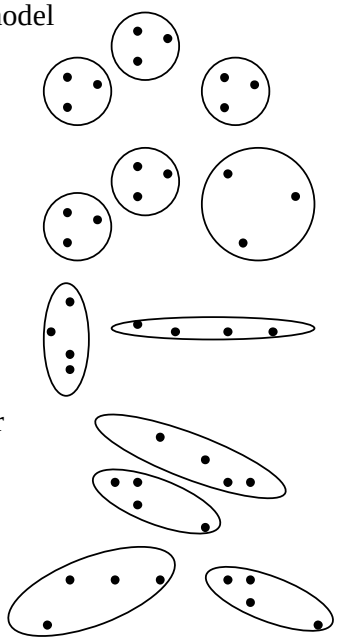
$\sum_k \lambda_k B_k$ where B_k is diagonal, $|B_k| = 1$

EEE elliptical model: cluster shapes parallel to each other

$\sum_k \lambda D A D^T$

Unconstrained model (VVV)

$\sum_k \lambda_k D_k A_k D_k^T$



Bottom line: BIC allows to choose the best possible model whereas MLE will always favor VVV model. In general, VVV model is the best model in terms of the highest likelihood, however it needs larger number of parameters.