The Kindest Cut: Minimum Message Length Segmentation

Rohan A. Baxter and Jonathan J. Oliver
Dept. of Computer Science
Monash University, Clayton, 3168
Australia
rohan,jono@cs.monash.edu.au

April 15, 1996

Abstract

We consider some particular instances of the segmentation problem. We derive minimum message length (MML) expressions for stating the region boundaries for some one and two dimensional examples. It is the found the message length cost of stating region boundaries is dependent on the noise of the data in the separated regions and also the 'degree of separation' of the two regions.

The framework given here can be extended to different shaped cuts and also non-constant fits for the regions. Possible applications for the work presented here include its use in tree (i.e. CART) regression and in image segmentation.

1 Introduction

We consider some instances of the segmentation problem. The segmentation problem arises whereever it is desired to partition data into distinct homogenous regions. The regions have distinct boundaries. This distinguishes the segmentation problem from mixture modelling, where regions are represented by overlapping probability distributions, with no distinct boundaries. The segmentation problem is to decide whether to divide a region into one or more sub-regions and to choose where to make the division.

In the one-dimensional case, the region to be partitioned is a line segment and a region boundary is a cut-point. In the two-dimensional case, the region is a plane and region boundaries consist of lines partitioning the plane into sub-regions.

The segmentation problem arises in applications that partition data in areas such as data mining, A.I. and statistics. Decision tree algorithms, piecewise curve-fitting, and image processing all need to solve instances of the segmentation problem in some way. Different applications will represent the data in the regions in different ways. For instance, the line dividing the plane can be a straight line, a polynomial, a spline or have some other mathematical form.

The fit of a segmentation model to data can be expressed precisely using maximum likelihood estimation. However, choosing a segmentation model to maximize the likelihood results in a model with homogenous regions containing only one datum each. Therefore, heuristics for solving the segmentation problem usually involve 'penalizing' a segmentation for its model complexity.

Minimum encoding estimation penalizes model complexity by including the coding cost of the model parameters. In this paper, we derive minimum message length (MML) expressions for stating the region boundaries for an one dimensional instances of the image segmentation problem. It is the found the message length cost of stating region boundaries is dependent on the noise of the data in the separated regions and also the 'degree of separation' of the two regions.

The framework given here can be extended to different shaped cuts and also non-constant fits for the regions. Possible applications for the work presented here include its use in tree (i.e. CART) regression and in image segmentation.

Other approaches are discussed in Section (4.1). The MML method proposed here differs from previous minimum encoding approaches to segmentation by including coding penalties for stating the parameters of each region *and* also for stating the region boundary.

2 Fitting D constants to Univariate Data

Consider some data given as follows. We have n data points, each of which consists of a pair (x_i, y_i). The x_i are distributed uniformly between $[0, R]$, and

each y_i is to be classified as some constant, c_j. We assume that the errors in stating y_i by a constant, c_j are i.i.d Gaussian with an unknown σ_j.

The range $[0, R]$ can be cut into d pieces by $d-1$ segment boundaries, $\{v_0, v_1, ..., v_d\}$, where $v_0 = 0$ and $v_d = R$. For each of the d segments we estimate a constant, $\{c_1, ..., c_d\}$.

2.1 The Form of the Message

We consider sending a message for this data of the form:

```
<d>
<c0>...<c(d-1)>
<v1>...<v(d-1)>
<y1>...<yn>
```

The distance between successive x_i is assumed known. Since x_i is uniform, one can work out the number of x_i in any region from knowing the size of the region. The range of x_i is assumed to be known by the receiver *a priori*.

We can extend the following case to arbitrarily distributed, x_i, but consider uniform x_i to ease the explanation of the derivation.

3 Minimum Message Length Formulas

Wallace and Freeman [8] showed that the expected message length under some fairly general conditions (a locally flat prior and quadratic log-likelihood function) for sending x and parameters θ is:

$$E(MessLen(x, \theta)) = -\log h(\theta) - \log f(x|\theta) + 0.5 \log |F(\theta)| + \frac{d}{2} \log \kappa_d \qquad (1)$$

where

- $h(\theta)$ is the assumed known prior density on θ
- $f(x|\theta)$ is the assumed known conditional probability, or likelihood, of x give θ
- $|F(\theta)|$ is the determinant of the Fisher Information matrix.
- κ_d is lattice constant dependent on the dimension of θ, d.

3.1 The One Segment, $d = 1$, case

For fitting a constant with no cut points $d = 1$, our θ consists of two parameters, c_0 and σ_0.

We choose a noninformative prior on these, based on the population variance of y_i [6].

Since the likelihood is Gaussian $N(c_0, \sigma)$, the Fisher Information matrix in this case has two diagonal entries and is , $\frac{n^2}{\sigma^4}$.

There are two parameters so $d = 2$ and $\kappa_2 = \frac{5}{36\sqrt{3}}$ [2].

$$\begin{aligned} E(MessLen) &= -\log h(c_0, \hat{\sigma}_0) + 0.5 \log |F(c_0, \hat{\sigma}_0)| + 0.5 + 0.5 \log \kappa_2 \\ &\quad + n \log(\sqrt{2\pi}\hat{\sigma}_0) + \sum_{i=1}^{n} \frac{(y_i - \hat{c_0})^2}{2\hat{\sigma_0}^2} + \frac{d}{2} \end{aligned} \tag{2}$$

where $|F(c_0, \sigma_0)|$ is the determinant of the Fisher Information matrix, $h(c_0, \sigma_0)$ is the prior probability density on the constant, c_0, and n is the number of x_i. The last two terms of Equation (2) are the negative log-likelihood. When we note that $\sum_{i=1}^{n} \frac{(y_i - \hat{c_0})}{2\hat{\sigma_0}^2} = n\hat{\sigma}_0^2$, the last term simplifies to $\frac{n}{2}$.

3.2 The $d = 2$ case

We now consider the effect of stating the cut point, $\hat{v}$, imprecisely. Let the cut point have precision $AOPV$ (an acronym for Accuracy Of Parameter Value).

Let ϵ be the difference in the $\hat{v}$ stated in the message, and the maximum likelihood v estimated from the data. Assume ϵ is uniformly distributed in the range $[\frac{-AOPV}{2}, \frac{AOPV}{2}]$. We now need to state c_0 and c_1, the constants fitted to the data in the regions on each side of the cut point and also the cut point itself.

In the following we denote the set of x_i in region 0 fitted by constant c_0 as S_0. We do the same for the set of x_i in region 1 fitted by constant c_1, denoting it S_1. The residual errors are distributed as $N(0, \sigma_0)$ for region S_0 and as $N(0, \sigma_1)$ for region S_1.

The message length expression for the parameters is then written as follows:

$$\begin{aligned} MessLen(\theta) &= -\log Prob(c_0, \hat{\sigma}_0, c_1, \hat{\sigma}_1) + 0.5 \log |F(c_0, \hat{\sigma}_0, c_1, \hat{\sigma}_1)| + 0.5 + 0.5 \log \frac{1}{12} \\ &\quad - \log h(v) - \log AOPV(v) \end{aligned} \tag{3}$$

We note that, given our assumptions about uniform x, that $n(1 - \frac{\epsilon}{R})$ data items will lie clearly in their correct regions, but that $\frac{n\epsilon}{R}$ data items will be put in the 'wrong' region.

The message length expression for the data is:

$$MessLen(x|\theta) = -(1 - \frac{|\epsilon|}{R})(n_0 \log(\sqrt{2\pi}\hat{\sigma}_0) + n_1 \log(\sqrt{2\pi}\hat{\sigma}_1)$$

$$
\begin{aligned}
&+ \sum_{i \in S_0} \frac{(y_i - c_0)^2}{2\hat{\sigma}_0^2} + \sum_{i \in S_1} \frac{(y_i - c_1)^2}{2\hat{\sigma}_1^2}) \\
&-(\frac{|\epsilon|}{R}(n_0 \log(\sqrt{2\pi}\hat{\sigma}_0) + n_1 \log(\sqrt{2\pi}\hat{\sigma}_1)) \\
&+ \sum_{i \in S_1 \ but \ put \ in \ S_0} \frac{(y_i - c_0)^2}{2\hat{\sigma}_0^2} \\
&+ \sum_{i \in S_0 \ but \ put \ in \ S_1} \frac{(y_i - c_1)^2}{2\hat{\sigma}_1^2}) \qquad (4)
\end{aligned}
$$

We wish to determine the expected message length. The expected value of $|\epsilon|$ is $\frac{AOPV}{4}$, since (considering only positive values of ϵ)

$$
\begin{aligned}
\epsilon &= \int_{\hat{v}}^{\hat{v}+\frac{AOPV}{2}} \frac{x}{AOPV} dx \qquad (5) \\
&= \frac{AOPV}{4} \qquad (6)
\end{aligned}
$$

Now since

$$E(\sum_{i \in S_0} (y_i - c_0)^2) = n_0 \hat{\sigma}_0^2 \qquad (7)$$

and the expected cost of stating a datum in S_0 when it should have been in S_1 is:

$$E(\sum_{i \in S_0} (y_i - c_1 - d)^2) = n_0(\frac{\hat{\sigma}_1^2 + d^2}{2\hat{\sigma}_0^2}) \qquad (8)$$

where $d = |c_1 - c_0|$. To see this, expand out the quadratic on the left side of Equation (8):

$$y_i^2 + c_1^2 + d^2 - 2c_1 y_i - 2dy_i + 2c_1 d \qquad (9)$$

We note that, in expectation,

$$E(y_i) = c_1 \qquad (10)$$

Substituting this into the second last term of Equation (9) , we are left with

$$y_i^2 + c_1^2 + d^2 - 2y_i^2 c_1 = (y_i - c_1)^2 + d^2 \qquad (11)$$

We can now combine Equations (4) and (3) and take the partial derivative w.r.t. $AOPV$, set the result to 0 and solve for the optimal $AOPV$ to minimize the message length expression:

$$
\begin{aligned}
\frac{\partial E(MessLen(x,\theta))}{\partial AOPV} &= \frac{-1}{AOPV} - (\frac{n}{4R}(n_0 \log(\sqrt{2\pi}\hat{\sigma}_0) + n_1 \log(\sqrt{2\pi}\hat{\sigma}_1)) \\
&\quad \frac{1}{4R}(n_0 \frac{\hat{\sigma}_1^2 + d^2}{2\hat{\sigma}_0^2} + n_1 \frac{\hat{\sigma}_0^2 + d^2}{2\hat{\sigma}_0^1}) \\
&\quad = 0 \qquad (12)
\end{aligned}
$$

The expected message length has a minimum at:

$$AOPV = \frac{8R}{\frac{n_0\hat{\sigma}_1^2+d^2}{\hat{\sigma}_0^2} + \frac{n_1\hat{\sigma}_0^2+d^2}{\hat{\sigma}_1^2}} \quad (13)$$

If the σ_i^2 and n_i are the same for both regions, an assumption made in some segmentation applications, the following simpler $AOPV$ results:

$$AOPV = \frac{8R}{n(1+\frac{d^2}{\sigma^2})} \quad (14)$$

3.3 Discussion

We now check that the results in Equations (13) and (14) accord with our intuitions. The $AOPV$ can be interpreted as a volume in the parameter space. As n_0 and n_1 grow, we see that the volume decreases because the estimate of v can be stated more accurately.

The $AOPV$ is also naturally dependent on the distance, d, separating the constants, c_0 and c_1. For large d, we expect that the cutpoint will be clearer and so can be stated more accurately.

The $AOPV$ is also dependent on the ratio of σ_0^2 and σ_1^2. Note that this ratio only matters if $n_0 \neq n_1$.

3.4 Message Length Expression

Let $X = \frac{n_0(\hat{\sigma}_1^2+d^2)}{\hat{\sigma}_0^2} + \frac{n_1(\hat{\sigma}_0^2+d^2)}{\hat{\sigma}_1^2}$, so that the optimal $AOPV$ is $\frac{8R}{X}$.

We substitute the optimal $AOPV$ into the message length expression obtained by combining Equations (3) and (4)

$$\begin{aligned} E(MessLen(x,\theta)) = & -\log h(c_0,c_1) + 0.5\log|F(c_0,\hat{\sigma}_0)| + 0.5\log|F(c_1,\hat{\sigma}_1)| \\ & -\log h(v_1) - \log\frac{8R}{X} \\ & +(1-\frac{2}{X})(n_0\log(\sqrt{2\pi}\hat{\sigma}_0) + n_1\log(\sqrt{2\pi}\hat{\sigma}_1) + \frac{n}{2}) \\ & +(\frac{2}{X})(n_0\log(\sqrt{2\pi}\hat{\sigma}_0) + n_1\log(\sqrt{2\pi}\hat{\sigma}_1) \\ & +n_0\frac{\hat{\sigma}_1+d^2}{2\hat{\sigma}_0^2} + n_1\frac{\hat{\sigma}_0+d^2}{2\hat{\sigma}_0^1}) \end{aligned} \quad (15)$$

This expression generalises easily to more than 1 cut point.

4 The Simulation

4.1 The Criteria Tested

We tested the following criteria for simplest hypothesis testing case. Is there a cut-point or not?

- maximum likelihood (MaxLik). This is the case where no penalty term is applied, such as in CART [1].
- MML, using Equations (2) and (15) of this paper.
- AIC, using $-\log f(x|\theta) + k$ [5].
- BIC, using $-\log f(x|\theta) + \frac{k}{2}\log n$ [4].
- QUI, using $-\log f(x|\theta) + \log n$. We note that its use by Quinlan is in a slightly different context to here, but believe the rationale provided for its use still holds [7].
- DOM, using $\log f(x|\theta) + \log \binom{n}{d}$. We note that Dom used this penalty measure in the different, but related, context of segmenting binary strings and that our use of it here is not meant to imply that Dom would advocate its use here [3]. Dom requires that $d < 0.5n$, otherwise the complexity of the term decreases for increasing d, which is counter to prior beliefs about segmentation models in most applications.

4.2 Hypothesis Testing

We wrote a program in C to generate data between $[0, 100]$ with a cutpoint at $x = 50$. The program accepts N, d and σ as input and then calculates the message length estimates for two models:

- no cut-point: Equation (2)
- a cut-point: Equation (15)

We prefer the model with the minimum message length estimate.

In the first simulation, we generated data with $d = 0$. We searched for a cut point in two ways:

1. Consider every possible cut-point, see Table (1)
2. Consider only cut-points leading to regions with at least 3 data items, see Table (2)

The second way corresponds to a heuristic used by many segmentation algorithms and we see that it indeed does improve the methods we tested. The poor behaviour of the methods other than MML in Table (1) explains the popularity of the heuristic. The log-likelihood savings of small segments are high compared to the penalties and so the methods other than MML badly overfit the segmentation model.

We note that MML slight overfits in the case of larger n, e.g. $n = 80$. On closer examination, we found that MML in these cases was preferring segments of around size 5. We would have preferred MML to never overfit, but must acknowledge that the data will sometimes suggest two segments, even if there is only one.

	One Segment: $n = 10$, $\sigma_x = \sigma_y = 1.00$					
	MaxLik	MML	AIC	BIC/MDL	Quinlan	Dom
1 True	0	100	0	0	0	0
2	100	0	100	100	100	100
	One Segment: $n = 20$, $\sigma_x = \sigma_y = 1.00$					
	MaxLik	MML	AIC	BIC/MDL	Quinlan	Dom
1 True	0	100	0	0	0	0
2	100	0	100	100	100	100
	One Segment: $n = 40$, $\sigma_x = \sigma_y = 1.00$					
	MaxLik	MML	AIC	BIC/MDL	Quinlan	Dom
1 True	0	100	0	0	0	0
2	100	0	100	100	100	100
	One Segment: $n = 80$, $\sigma_x = \sigma_y = 1.00$					
	MaxLik	MML	AIC	BIC/MDL	Quinlan	Dom
1 True	0	97	31	0	0	0
2	100	3	100	100	100	100

Table 1: The predicted number of segments when there is one segment.

In the second simulation, we generated data with $d = 1$. We again searched for a cut point in two ways:

1. Consider every possible cut-point, see Table (3)

2. Consider only cut-points leading to regions with at least 3 data items, see Table (4)

As one would have expected from the first simulation, MML underfits for small n, but then improves for larger n. It is misleading just to consider whether the right number of segments was selected, we should check how far the estimated cutpoint was away from the true cutpoint at $x = 50$. We collected statistics on this and found that the MML estimate of the cut-point was more accurate than the other criteria.

	One Segment: $n = 10, \sigma_x = \sigma_y = 1.00$					
	MaxLik	MML	AIC	BIC/MDL	Quinlan	Dom
1 True	0	100	72	74	96	87
2	100	0	28	26	4	13
	One Segment: $n = 20, \sigma_x = \sigma_y = 1.00$					
	MaxLik	MML	AIC	BIC/MDL	Quinlan	Dom
1 True	0	100	59	89	100	96
2	100	0	41	11	0	4
	One Segment: $n = 40, \sigma_x = \sigma_y = 1.00$					
	MaxLik	MML	AIC	BIC/MDL	Quinlan	Dom
1 True	0	100	43	90	100	99
2	100	0	57	10	0	1
	One Segment: $n = 80, \sigma_x = \sigma_y = 1.00$					
	MaxLik	MML	AIC	BIC/MDL	Quinlan	Dom
1 True	0	92	47	96	100	99
2	100	8	53	4	0	1

Table 2: The predicted number of segments when there is one segment and the size of segments restricted to greater than 3.

Further simulations, varying d, σ_i and n_i have been conducted producing similar trends in segmentation criteria behaviour.

5 Conclusion

The MML framework can provide both parameter estimation and model selection without overfitting in segmentation problems. We have derived message length formulae for a simple segmentation problem. We tested the formulae and found that they performed well in determining the number of regions in the simple simulations conducted here. The framework here is being extended to the two-dimensional case for use in image segmentation.

References

[1] L. Breiman et al. *Classification and Regression Trees.* Wadsworth, 1984.

[2] J.H. Conway and N.J.A. Sloane. *Sphere Packings, Lattices and Groups.* Springer-Verlag,New York, 1988.

[3] B. Dom. MDL estimation with Small Sample Sizes including an application to the problem of segmenting binary strings using bernoulli models. Technical Report RJ 9997 (89085) 12/15/95, IBM Research Division, Almaden Research Center, 650 Harry Rd, San Jose, CA, 95120-6099, 1995.

	Two Segments: $n = 10$, $\sigma_x = \sigma_y = 1.00$					
	MaxLik	MML	AIC	BIC/MDL	Quinlan	Dom
1 True	0	100	43	50	85	70
2	100	0	57	50	15	30
	Two Segments: $n = 20$, $\sigma_x = \sigma_y = 1.00$					
	MaxLik	MML	AIC	BIC/MDL	Quinlan	Dom
1 True	0	96	29	55	96	72
2	100	4	71	45	4	28
	Two Segments: $n = 40$, $\sigma_x = \sigma_y = 1.00$					
	MaxLik	MML	AIC	BIC/MDL	Quinlan	Dom
1 True	0	77	5	36	90	63
2	100	23	95	64	10	37
	Two Segments: $n = 80$, $\sigma_x = \sigma_y = 1.00$					
	MaxLik	MML	AIC	BIC/MDL	Quinlan	Dom
1 True	0	29	0	13	79	30
2	100	71	100	87	21	70

Table 3: The predicted number of segments when there is two segments.

[4] Mengxiang Li. Minimum description length based 2-d shape description. In *IEEE 4th Int. Conf. on Computer Vision*, pages 512–517, May 1992.

[5] Z. Liang et al. Parameter estimation of finite mixtures using the EM algorithm and information criteria with applications to medical image processing. *IEEE Trans. on Nuclear Science*, 39(4):1126–1133, 1992.

[6] J.J. Oliver, Baxter R.A., and Wallace C.S. Unsupervised Learning using MML. In *Machine Learning: Proceedings of the Thirteenth International Conference*. Morgan Kaufmann Publishers, San Francisco, CA, 1996.

[7] J.R. Quinlan. Improved use of continuous attributes in c4.5. *Journal of Artificial Intelligence*, 4:77–90, 1996.

[8] C.S. Wallace and P.R. Freeman. Estimation and inference by compact coding. *J. R. Statist. Soc B*, 49(3):240–265, 1987.

	Two Segments: $n = 10, \sigma_x = \sigma_y = 1.00$					
	MaxLik	MML	AIC	BIC/MDL	Quinlan	Dom
1 True	0	100	72	74	96	87
2	100	0	28	26	4	13
	Two Segments: $n = 20, \sigma_x = \sigma_y = 1.00$					
	MaxLik	MML	AIC	BIC/MDL	Quinlan	Dom
1 True	0	96	0	0	0	0
2	100	4	100	100	100	100
	Two Segments: $n = 40, \sigma_x = \sigma_y = 1.00$					
	MaxLik	MML	AIC	BIC/MDL	Quinlan	Dom
1 True	0	75	0	0	0	0
2	100	25	100	100	100	100
	Two Segments: $n = 80, \sigma_x = \sigma_y = 1.00$					
	MaxLik	MML	AIC	BIC/MDL	Quinlan	Dom
1 True	0		39	0	0	0
2	100	61	100	100	100	100

Table 4: The predicted number of segments when there is two segments and the size of segments restricted to greater than 3.