Genetic Algorithm for Feature Selection with Bimodal Distribution Removal Application in Reader Type Classification

Xueting Sun
u5900182@anu.edu.au
Research School of Computer Science, Australian National University

Abstract.
In real world, data for Neural Network training can be insufficient or with noises, which will bring difficulties for the training. Bimodal Distribution Removal (BDR) is a technique to distinguish outliers from small data set and make the most of the training data. In this report, data set from the visual distractions experiment was used to predict first and second English language readers. Genetic Algorithm (GA) for feature selection is applied to find the optimal feature set for training in Artificial Neural Network (ANN). BDR is also applied in ANN training for generalization improvement. Pre-processing, comparisons of techniques and parameters, and performance evaluations will be demonstrated in the report. From experiments, Simply using BDR cannot significantly improve the performance, however, when applying ANN+BDR as evaluation function to conduct GA feature selection, the accuracy increases.

Keywords: Artificial Neural Network, Genetic Algorithm, Feature Selection, Bimodal Distribution Removal, Classification, English Reader, Visual Distractions

1 Introduction

With the development of electronic devices, reading in digital environments are becoming popular. Digital reading can help people attain information in a convenient way and becomes popular in people's life. Internet has become a significant resource where people gain and share information. For example, reading news online has become more popular than reading printed newspaper for a great many people. However, there are visual distractions when people read in digital environment, such as messages from online chat, popup windows in web pages, which would reduce the quality of information gained. This could be harmful for entities that share or sell information online. Compared with reading printed books, it would be more likely that readers are distracted while reading using electronic devices, e.g. laptops, mobile phones, etc. Copeland and Gedeon (2015) investigated the factors that affect the quality of reading in many aspects. It is shown that the text readability would significantly influence the distraction rate, eye movements, and comprehension based on the research on both first language English (L1) readers and second language English (L2) readers.

In this report, Reading Distractions data set is chosen for demonstrations. From the experiment, data about participants' reading conditions, fixations, distractions, reading scores, reading habits and text readability were collected. In Reading Distractions data set, there are 23 attributes while only 66 records. ANN turns out to perform worse with complicated dimensions and small-sized data set. As LeCun, Denker and Solla (1990) demonstrated, removal of unimportant parameters from the network can contribute to better generalization, fewer training patterns required and classification speed. That is to say, feature selection is vital for constructing a better model. However, a particular technique should be taken due to complicated computations when experimenting different sets of features. If there are p features, then there are 2^p feature sets. Whereas feature selection with Genetic Algorithm can significant reduce the computation required and bring better measurements for determining pattern's class (Brill et al., 1992). GA for feature selection can be used to check whether the current set of features is the optimal choice and whether subset can perform better. The choice of features can improve classification performance.

From the previous report, Bimodal Distribution Removal is applied in ANN, since outliers exist and influence the generations of the model. When observing error distributions, it can be found that bimodal error distributions exist in the training.

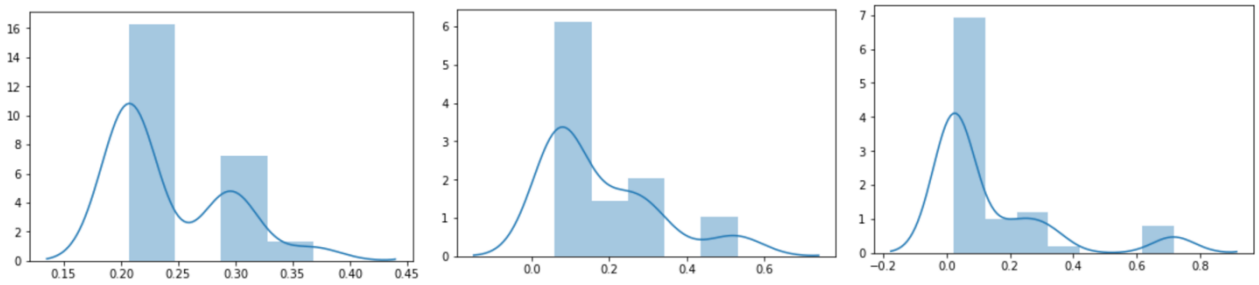

Figure 1. Bimodal Error Distribution (epoch num = 1, 50, 100)

When epoch = 1, error points focus on 0.2 and 0.3. When epoch = 50, 100, the curve flattens out but outliers still exist in the range of 0.5 to 0.7. That is to say, the technique BDR can be useful in this data set to detect and remove noisy data points for better training performance (Slade & Gedeon, 1993).

Since the number of records was small, more attentions should be paid to make the most of the data set and improve generalizations with less impact of noisy data. Thus, the particular technique Bimodal Distribution Removal (BDR) will be applied in the Reading Distractions dataset to distinguish outliers and remove some of them.

In this report, GA for feature selection will be conducted at the first. A set of features will be selected after evaluation, which is based on testing accuracy of the model. Then, ANN with BDR will be trained with the feature set mentioned above. Feature selection, BDR technique application, changes of parameters and their performance will be discussed in the following sections. Appropriate validation methods will be chosen as well. Holdout and Leave-one-out validations will be implemented for performance evaluations.

2 Method

2.1 Data Preprocessing

From the original spread-sheet, the third sheet was chosen. There are 66 records about 22 participants' experiment data. Some columns are dropped due to meaninglessness or redundancies.

- #0 'ParticipantID'
- #1 'Condition'
- #9 'Num fixations in text area/out of text area'
- #12 'ratio of fixation duration in text are to out of text area'

As #0 column is a string about when the participant participated in the experiment, which is meaningless for training. #1 Column contains like 'AE', is a combination of #3 column 'Condition.1' and #2 column 'Text Type', which is redundant. For columns like #9 and #12, they are both about ratios and the original statistics like #7 number of fixations in text area and #8 out of text area, #10 fixation duration in text area and #11 out of text area are all kept in the data set as well.

Columns like #2 'Text Type', #3 'Condition.1' and #18 'Time Taken' are not in numeric format and should be transformed. There are two text type E (Easy-to-read) and H (Hard-to-read). E can be encoded as 0 and H can be encoded as 1. Then condition A, B, C are encoded as 0, 1 and 2. For #18 column 'Time Taken', the original format is datetime like 'h:min:sec' and it can be transferred to the second unit. For the output in the neural network, the values in column 'L1/L2' can be encoded as 0 or 1.

What's more, features with different ranges exist in the data set. For example, the number of fixations can be 779 while the fixation durations can be 6.78E-06. There are a few features in very different ranges in the data set and values in 16 columns can all be normalized to the range of 0 to 1.

2.2 Neural Network Construction

ANN with 18 inputs, 1 hidden layer, 10 hidden neurons, 2 output neurons, 0.01 learning rate is set. Sigmoid is used as activation function, cross entropy as loss function and SGD as optimizer. In the previous report, Adam was applied, however the result is not good. Reddi, Kale & Kumar (2018) stated that Adam may not converge in some cases, especially in learning with large output spaces.

Holdout validation is applied. There are 51 patterns in the training set, 15 patterns in the testing set. From the experimentation, after 100 epochs, it tends to overfit and present 100% accuracy in training set after 351 epochs. However, Accuracy = 60% in testing set.

2.3 Genetic Algorithm for Feature Selection

In the previous section, the whole set of features are fed into the ANN. The number of features is 18. However, not all features are important to separate the classes and smaller set of features may be expected for better modeling. In order to select the relevant and meaningful features, Genetic Algorithm can be applied. Genetic Algorithm is a biology analogy. Feature selection can be considered as a group of individuals (solutions) to evolve until expected offspring (results) are found. The evaluation of offspring is based on fitness (testing accuracy in the context).

Initialization and Fitness Evaluation

In the Representation phase, chromosomes are defined. A chromosome contains 18 binary numbers. Indexes with 0 indicate the corresponding features are not selected, while 1 indicates the feature is chosen as part of input. The population is initialized as 20 at the first. Smaller population can reduce computations and time consuming, however, it will lead to bad solutions and low testing accuracies. Different populations and generations will be used and discussed in the results section.

In the context, testing accuracy after training in ANN with BDR will be used as evaluation function. Each chromosome in the population will act as an individual to begin a new training and will be assigned with fitness values (testing accuracy) after training. Then two parents will be selected among the current generation based on fitness. That is to say, individuals with higher accuracies have more chance to be selected. Their natures can be retained and have more probability to pass on to the next generation. In the context, sets of features with higher accuracies have more chance be parents and evolved to good generations with higher accuracies. Individuals with lower accuracies tends to be replaced in the following generations.

Crossover and Mutation

Chromosomes of the selected individuals have particular probabilities to crossover and mutate. Binary numbers in the parents' chromosomes have chance to exchange segments between crossover points in chromosomes. For mutation, 1 can flip to 0 and 0 can flip to 1. Thus, variation can be assured. Different sets of features with the overall trend of increasing accuracies will have chance to be evaluated and evolved. It helps to avoid getting into local minimum in the early stage. Parameters of crossover probability, mate and mutating probability will be discussed in the next section.

Max number of generations is set. The process of crossover, mutation, evaluation and selection is repeated for the pre-determined number of generations. Testing accuracies in different generations are recorded for comparisons. The optimal solution (particular set of features) will be found.

2.4 Bimodal Distribution Removal

From the observation of simple neural network, the performance is not good. Many factors might contribute

to this situation. For example, the small number of patterns in the data set might lead to a higher possibility of overfitting. Also, the outliers inside the data set would also reduce the performance of the neural network.

Bimodal Distribution Removal is a technique to detect outliers from valid but rare data points. It can discover noisy data in the real-world dataset based on variance of error and mean error of predicted data and real data. During the training, normalized variance of errors v_ts is recorded. When $v_{ts} < 0.1$, some of patterns in the training set may be removed based on some rules.

Patterns whose error greater than the mean error of the training set δ_{ts}will be taken and then the mean value δ_{ss} and standard deviation σ_{ss} will be recalculated. Patterns with $error \geq \delta_{ss} + \alpha * \sigma_{ss}$ will be permanently removed from the training set. The process of removal will be conducted every 50 epochs until $v_{ts} < 0.01$.

When implementing, error is calculated based on the squared difference between predicted probability that any of the classes are true and the target. That is, if there is 30% probability that the reader is L1 while actually the reader is L1, the error will be $0.7 * 0.7 = 0.49$. Then, errors of all patterns in the training set will be normalized for the further use.

For the setting of parameter α, it should be in the range of 0-1. As α increases, there will be less strict filtering of records that can be kept. Less patterns will be removed. At first, α is set to 0.5, however, training is never halted even epoch = 70,000, for the reason that variance is never below 0.04. There are 54 patterns in the training set at the beginning, and after epoch = 5900, the number of patterns in the training set stabilize at 9.

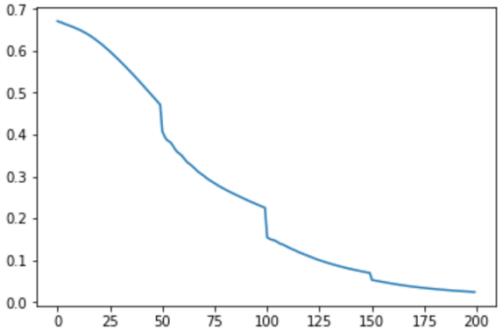

Figure 2. Loss trend in the first 200 epochs

From the trial, it can be seen parameter should be changed. Firstly, too many patterns are removed and it will impact generalization of the model. Secondly, appropriate training termination lacks. It can be seen around epoch = 200, overfitting may already happen.

2.5 Performance Evaluations

Holdout validation is used to evaluate performance at first. 66 records from Reading Distractions dataset are randomly splitted into 2 groups: 80% in training set and 20% in testing set. Due to the small number of records, holdout validation seems to 'waste' some data and usually only about 50 patterns can be trained. Also, the evaluation may heavily depends on how the division is made.

What's more, Leave-One-Out validation is used. Since the data set only contains 66 records, the computation will not be expensive. For 66 separate time, 65 patterns are for training and only one pattern will be tested. Average error is computed and used to evaluate the performance.

3 Results and Discussion

3.1 GA and ANN parameters

ANN + BDR in evaluation function

According Xu and Chen (2008), the number of hidden neurons is recommended to around $\frac{2}{3}$ size of input and output layer. Thus, the number of hidden neurons in ANN in evaluation function is set according to the individual's chromosome, usually in the range of [8, 15] . BDR is also applied in ANN as part of evaluation function. BDR parameters setting is discussed in the previous report.

Figure 3. Tournament Selection with size = 3, Roulette Wheel Selection

For initial setting, population = 20, generation = 15, crossover probability = 0.5, mutation probability = 0.5. For selection method, comparing with Roulette Wheel Selection, Tournament Selection turns out to converge faster and is more commonly used (Blickle & Thiele, 1995). From the Figure. 3, it can be seen fitness is more converged to higher accuracies in Tournament Selection, while in Roulette Wheel Selection, there are more oscillations. Testing accuracies are lower as well.

Figure 4. mutation probability = 0.2 & mutation probability = 0.5

Generation number is changed to 20. Mutation probability is also changed to 0.2 and 0.5. It can be seen there is an overall trend of increasing accuracies in the left figure. Good natures seem to be passed on to the next generation. After generation = 8, the max fitness of the generation is stabilized to 100%.

While in the right figure (mutation probability= 0.5), fitness does not converge and still significant oscillates. When mutation probability = 0.3, 0.4, there are still oscillations. Although the fitness changed greatly in mutation probability = 0.4, 0.5, the population with the higher mutation probability can have high accuracy in a particular generation. Max fitness usually appear in generation = 8, 9, 10.

When lowering crossover probability to 0.2, max and average fitness over generations are both decrease.

Only apply ANN in evaluation function

In the previous section, BDR is applied in ANN to distinguish and remove outliers with parameters $\alpha = 0.5$, (var_ts < 0.03) or (epoch > 500) (discussed in the next section). When only applying ANN and using Holdout Validation to return testing accuracy in GA evaluation function, testing accuracy decreases.

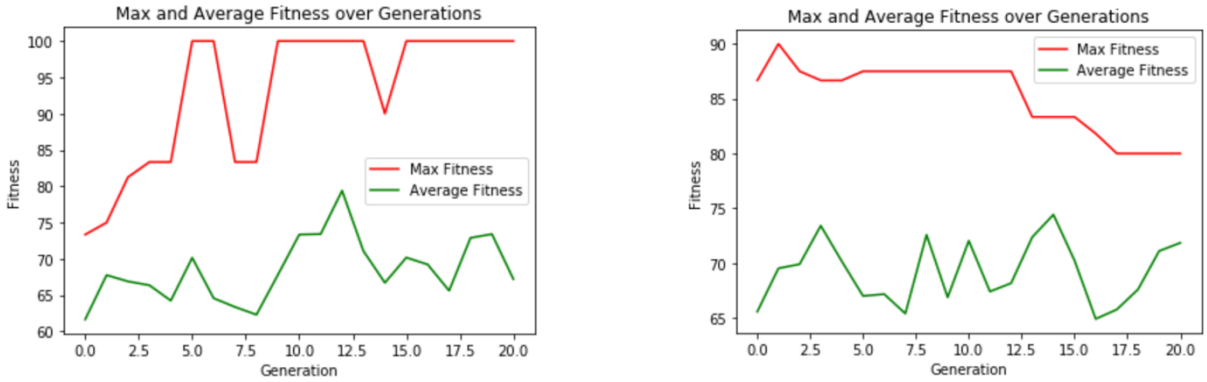

Figure 5. apply BDR in ANN & only apply ANN

After comparing all models, these features are selected to train the model.
[1, 2, 7, 10, 11, 12, 14, 16] with fitness 100%

3.2 Holdout Validation with all features selected

NN settings	Alpha	Termination rules	Deleted patterns	Accuracy	Terminate at
Hidden neurons = 15	0.5	(var_ts < 0.03) or (epoch > 500)	Num = 15	64.71%	Epoch = 209
Hidden neurons = 15	0.5	(var_ts < 0.03) or (epoch > 500)	Num = 9	70.59%	Epoch = 167
Hidden neurons = 15	0.5	(var_ts < 0.03) or (epoch > 500)	[14, 15, 23, 36, 45, 46, 53] [4, 23, 48] [12, 40, 42] Num = 13	77.8%	Epoch = 162
Hidden neurons = 10	0.5	(var_ts < 0.03) or (epoch > 500)	[5, 13, 22, 34, 36, 40] [10, 13, 37] Num = 9	66.67%	Epoch = 128
Hidden neurons = 10	0.5	(var_ts < 0.03) or (epoch > 500)	Num = 27	58.33%	Epoch = 501
Hidden neurons = 10	0.5	(var_ts < 0.03) or (epoch > 500)	Num = 21	53.33%	Epoch = 501
Hidden neurons = 15	0.6	(var_ts < 0.03) or (epoch > 500)	Num = 11	44.44%	Epoch = 205
Hidden neurons = 15	0.6	(var_ts < 0.03) or (epoch > 500)	Num = 8	71.43%	Epoch = 173
Hidden neurons = 15	0.6	(var_ts < 0.025) or (epoch > 500)	Num = 9	58.82%	Epoch = 167
Hidden neurons = 15	0.7	(var_ts < 0.025) or (epoch > 500)	Num = 23	40.82%	Epoch = 501

Table 1. Testing Accuracy in Holdout Validation with parameters tuning (without GA Feature Selection)

In this section, all 18 features are selected. According to the tests, it can be found that the model tend out to be better if the number of deleted pattern in the range of [8, 15]. That is to say, 75%-80% records in the training set should be kept. Alpha in the range of [0.5, 0.6] turns out to be good.

Secondly, termination rules need to be modified in this data set. According to BDR, the training cannot be halted until $variance < 0.01$. However, as mentioned above, even the epoch comes to 70,000, the variance is still above 0.02 and the training cannot terminate. For trial, the termination rules are changed to a higher variance (0.03 or 0.025) with an epoch upper bound (epoch = 500). From the form, the model with good performance mostly terminates before epoch = 200, since too many epochs for training lead to overfitting. Thus, it seems to be better with the termination rules $(var_{ts} < 0.03)\ or\ (epoch > 500)$.

However, due to drawbacks of Holdout validation, the evaluations can have a high variance. This may result from different divisions of training and testing set. Although BDR can help to remove outliers from training set, data in the testing set may be very different from data in the training set.

3.3 Leave-One-Out Validation with all features selected

In order to mitigate the impact of divisions of training and testing set, LOOCV is applied. For 66 separate times, all the data except for one record will be trained and the only one remaining pattern will be used to test. Then the accuracies of 66 predicted values will be evaluated.

NN settings	Alpha	Termination rules	Num of incorrect predictions
Hidden neurons = 15	0.5	(var_ts < 0.03) or (epoch > 500)	3 (index = 5, 8, 63)
Hidden neurons = 10	0.5	(var_ts < 0.03) or (epoch > 500)	3 (index = 5, 35, 51)
Hidden neurons = 10	0.6	(var_ts < 0.03) or (epoch > 500)	3 (index = 5, 20, 35)
Hidden neurons = 10	0.7	(var_ts < 0.03) or (epoch > 500)	3 (index = 5, 20, 35)
Hidden neurons = 10	1	(var_ts < 0.03) or (epoch > 500)	3 (index = 5, 19, 20)
Hidden neurons = 10	0.5	(var_ts < 0.01) or (epoch > 500)	2 (index = 5, 8)
Hidden neurons = 15	0.5	(var_ts < 0.02) or (epoch > 600)	2 (index = 5, 35)

Table 2. Testing Accuracy in LOO Validation with parameters tuning (without GA Feature Selection)

It can be found that, more data points trained, the performance is better. Unbalanced data points lead to poor performance, which may result from inappropriate divisions or original nature of the data set (e.g. 42 first English language reader and 24 second English language reader). More training patterns and application of technique BDR help improve the generalization. From the table, #5, #8, #20 and #35 patterns may be different from most of the patterns. When training, BDR can distinguish them and permanently remove them from the training set. After outliers removal and training, the error variance becomes smaller and the model will perform better when testing.

When hidden neuron = 15, $\alpha = 0.5, tenmination\ rules\ \{var_{ts} < 0.02\ or\ epoch > 600\}$, among 66 patterns in testing, the number of removed patterns is in the range of [19, 27], around 30% of the training set. The removal proportion 20% - 30% is relatively appropriate in BDR. However, the accuracy does not obviously increase and turns out to be instable. BDR may play a more important part in other scenarios, such as datasets with more patterns. BDR can detect, remove patterns and increase the accuracy more effectively.

3.4 Validation when using GA Feature Selection

From the previous section, feature set is determined as [1, 2, 7, 10, 11, 12, 14, 16]. Then these 8 features are fed into ANN with BDR to train the model. The number of hidden neurons is set to $\frac{2}{3}(input\ neurons + output\ neurons) = 7$. For BDR parameters, $(\text{var\_ts} < 0.03)$ or $(\text{epoch} > 500)$, $\alpha = 0.5$. It can be found that the overall testing accuracies increase in Holdout Validation (80% for training, 20% for testing).

In some cases, no pattern is deleted. The reason can be that outliers exist in the features unselected and now they are not recognized as outliers. In most of cases, accuracy tends to higher if the number of patterns removed is smaller. Some records which are not noisy data may be removed.

However, BDR performs well in evaluation function in GA feature selection. It contributes to detect and remove outliers, in order to assign a relatively accurate fitness value for each individual in generations.

Patterns Removed	Accuracy
none	83.33%
none	81.82%
none	77.78%
[3, 4, 5, 13, 14, 20, 22]	70.59%
[32, 33, 36, 41, 42, 43, 44, 45] [5, 6, 7, 13, 14, 15]	61.11%
[7, 15, 16, 33, 35]	90%

Table 3. Testing Accuracy in Holdout Validation after using GA feature selection

4 Conclusion and Future Work

In this report, Reading Distractions data set is used to demonstrate classifications of first English language readers and second English language readers. Due to the particular characteristics of the data set: small number of records and presence of noisy data, the technique Bimodal Distribution Removal can be applied to detect outliers and remove them. Moreover, in order to evaluate the performance, Holdout and Leave-One-Out validations are used. With the changes of parameters in BDR, the variance of errors becomes smaller and the performance tend to be better. However, there is no significant and stable accuracy increase after applying BDR to simple neural network. The reason can be that the number of data set is too small and the data points is imbalanced.

When applying Genetic Algorithm for feature selection, it can be seen that the accuracy increase. There is trade-off between accuracy and time-consuming. Bigger size of populations and generations tend to evolved to higher fitness (accuracy), however, it will take much time to compute. The balance between them should be investigated in the future. Moreover, testing accuracy may not be a good fitness value, since the original data set is too small and the divisions of training and testing set have significant influence on accuracy. Although BDR can help to remove some outliers, the fitness (accuracy) still be affected. The accuracy does not simply depend on ANN training effect and input features.

BDR plays important part with ANN in evaluation function in GA, however, after feature selection, excessive pattern removal can lead to lower accuracy. After feature selection, some outliers do not exist in the data set. The particular parameter settings need more efforts to evaluate and analyze.

For more accurate classifications of L1/L2 readers, more training patterns are needed. In this data set, only 24 records about L2 readers. The number of records about L1 and the number of records about L2 readers should also be balanced. For future work, more practices are needed. Parameter settings are vital. Some settings may be successful in some scenarios while in other datasets, appropriate modifications are needed. More work should be done to investigate and recorded for a successful model. Other techniques about outliers detections and removal can also be investigated. These methods will be applied, practiced and adapted in more scenarios in the future.

5 Reference

1. Blickle, T., & Thiele, L. (1995). A Comparison of Selection Schemes Used in Evolutionary Algorithms. from https://www.mitpressjournals.org/doi/abs/10.1162/evco.1996.4.4.361

2. Brill, F., Brown, D., & Martin, W. (1992). Fast genetic selection of features for neural network classifiers - IEEE Journals & Magazine. from https://ieeexplore.ieee.org/document/125874/

3. Copeland, L., & Gedeon, T. (2015). Visual Distractions Effects on Reading in Digital Environments: A Comparison of First and Second English Language Readers. In *Proceedings of the Annual Meeting of the Australian Special Interest Group for Computer Human Interaction* (pp. 506-516).

4. LeCun, Y., Denker, J. S., & Solla, S. A. (1990). Optimal brain damage. In *Advances in neural information processing systems* (pp. 598-605).

5. Reddi, S., Kale, S., & Kumar, S. (2019). On the Convergence of Adam and Beyond. from https://arxiv.org/abs/1904.09237

6. Slade P., Gedeon T.D. (1993) Bimodal distribution removal. In: Mira J., Cabestany J., Prieto A. (eds) New Trends in Neural Computation. IWANN 1993. Lecture Notes in Computer Science, vol 686. Springer, Berlin, Heidelberg

7. Xu, S. & Chen, L. (2008). A novel approach for determining the optimal number of hidden layer neurons for FNN's and its application in data mining, in *Proceedings of the 5th International Conference on Information Technology and Applications (ICITA '08), pp*. 683–686

8. Vehtari, A., Gelman, A. & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Stat Comput 27, 1413–1432. https://doi-org.virtual.anu.edu.au/10.1007/s11222-016-9696-4