Computer Hardware Data Set Analysis

12/31/2020

1 Introduction

1.1 Used data set

This project will draw conclusions about computer hardware industry using Computer Hardware Data Set from UCI (Data Description). This data set contains information about the hardware's vendor, characteristics and performance.It also contains predicted performance by a model used in an article, which will be ignored.

1.2 Questions that might get answers

1- What's the best vendor?

2- What's the hardware component with the greatest impact on performance?

2 Data

2.1 Insights to answer questions

Using this data, regression model will be fitted. According to the coefficient of the model, conclusions will be drawn.

2.2 Notes about the data

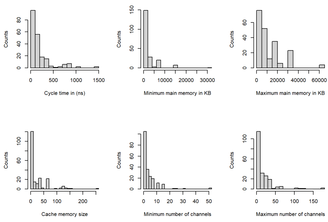
After exploring the data set, the following notes are found:

1. In few cases, same features results different performance with relatively high variance (e.g., ranges from 269 to 172), this will decrease the prediction accuracy of regression model, due to missing features that might explain this variations. Even if the fitted model isn't very accurate for prediction, it's parameters can be used to get insight of features impact on performance.
2. The way of computing performance wasn't clarified, also there are no priors about the model, that's why non-informative priors will be used.

2.3 Exploratory analysis

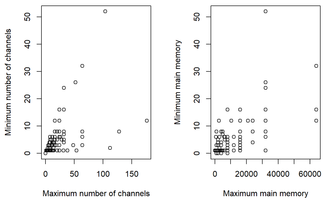
2.3.1 Histograms of data features

The purpose of those plot is getting insight of dynamic ranges of variables in the data set.



2.3.2 Plots between min and max values for both channels and main memory.

The purpose of this plot is exploring if there any correlation between those variable, which will impact the conclusion if exists.



As observed, correlation between variables isn't obvious. ACcordingly, it's not recommended to omit any of them.

3 Model

3.1 Hierarchical model

Symbol	Name
MYCT	Machine cycle time in nanoseconds
MMIN	Minimum main memory in kilobytes
MMAX	Maximum main memory in kilobytes
CACH	Cache memory
CHMIN	Minimum number of channels
CHMAX	Maximum number of channels

$$y(i) \sim \text{Pois}(\lambda(i))$$

$$log(\lambda(i)) = [\alpha(j) \quad \beta(1,j) \quad \beta(2,j) \quad \beta(3,j) \quad \beta(4,j) \quad \beta(5,j) \quad \beta(6,j)] \begin{bmatrix} 1 \\ MYCT(i) \\ MMIN(i) \\ MMAX(i) \\ CACH(i) \\ CHMIN(i) \\ CHMAX(i) \end{bmatrix}$$

$$\alpha(j) \sim N(\mu_1, \sigma_1^2)$$

$$\beta(1-6,j) \sim N(\mu_2, \sigma_2^2)$$

$$\mu_1 \sim N(0, 10^6)$$

$$\frac{1}{\sigma_1^2} \sim \Gamma(\frac{1}{2}, \frac{10}{2})$$

$$\mu_2 \sim N(0, 10^6)$$

$$\frac{1}{\sigma_2^2} \sim \Gamma(\frac{1}{2}, \frac{10}{2})$$

Where,

$$i = 1, 2, \ldots, n_{data}$$

$$j = 1, 2, \ldots, n_{vendors}$$

As observed, non-informative priors are used, since information about performance's scaling with components' upgrade is missing.

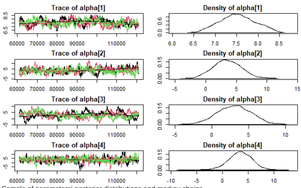
3.2 How to find answers using this model ?

In this model, vendors have different constant and coefficients, which indicates different quality. Comparing the coefficients will lead to fair comparison between vendors.

Also, finding the highest coefficient for most of the vendors models will indicate the most effective component in the machine, but this will arise the need of normalizing data to get fair comparison between components

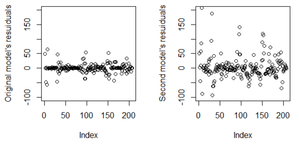
3.3 Model testing

1. The model parameters had high correlations, that's why it was need for a long burn in cycles and also large number of samples.



Sample of parameters' posterior distributions and markov chains

2. To justify the use of the specified model, it's required to compare it's residuals to the residuals of model which uses different constant for each vendor but same coefficients for all vendors (e.g.,$\beta(i)$ instead of $\beta(i,j)$, where i and j are data's index and vendo's index respectively).



As observed, most of the residuals of the original model are within acceptable range, which it's not the case for the second model. Also the residuals in both of them seems uncorrelated.

3 Results and conclusions

Using data from posterior distribution for all coefficients, will give answers to the specified questions.

3.1 What's the best vendor at each components of hardware?

To find answer to this question, it's required to find maximum value of posterior mean for coefficients assigned to each vendor (for e.g, maximum value in $[\beta_{\text{mean}}(1,1), \beta_{\text{mean}}(1,2), \ldots, \beta_{\text{mean}}(1, n_{\text{vendors}})]$, will result the best vendor in hardware's first component).

Since the model contains 7 parameters, that's how the result looks like.

```
## [1] "amdahl"    "amdahl"    "burroughs" "amdahl"    "c.r.d"     "c.r.d"
## [7] "apollo"
```

Which corresponds to following parameters ["constant","MYCT","MMIN","MMAX","CACH","CHMIN","CHMAX"]. The "constant" can be interpreted as the brand name finger print, in other words, it's the value added to the performance regardless of the hardware's specification. Accordingly it can be considered that "amdhl" is the best vendor among others.

3.2 What's the most effective component on hardware's performance?

To find an answer, its required to find maximum coefficient for each vendor, the most frequent component in all vendor will probably be the most effective. But before that, to get fair comparsion, we have to normalize the data values and re-fit the model.

As observed, the results are totally logical. The most valuable component for most of vendors is the cache memory, the second most valuable property is the number of channels, and the less important component is the normal memory, noting that $n_{vendors}$=29.

```
## max_comp
##  CACH CHMAX  MMAX  MMIN CHMIN
##    10     7     7     3     2
```