eRessearch in Statistics

Applications of HKU Grid Point in Statistics

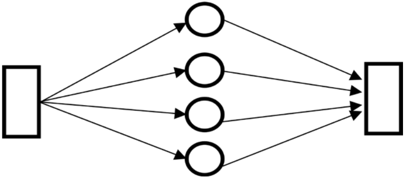
Wai-Keung Li

Department of Statistics and Actuarial Science
The University of Hong Kong

Aug 27, 2010

Embarrassingly parallel

- Many statistical computations are embarrassingly parallel
- All sub-tasks are completely independent



Embarrassingly parallel (Cont.)

Examples:

- Pure Monte Carlo Simulation
- Bootstrap and Jackknife
- Cross Validation
- Optimization with multiple starts

Advantages of using Grid Point for such computations:

- Provides linear speedup with respect to number of processors.
- Speedup$\approx \frac{\text{time with 1 CPU}}{\text{time with many CPUs}}$.

Pure Monte Carlo Simulation

Example 1:

- For (i in $1 \cdots n$) {
- Generate dataset A_i
- Fit statistical model to dataset A_i
- Obtain results from model i

  }

Aggregate summaries and report the results.

Bootstrap and Jackknife

Example 2:

```
For ($i$ in $1 \cdots n$) {
    Generate dataset $A_i$ by resampling the original dataset in some sense.
    Fit statistical model to dataset $A_i$
    Obtain results from dataset $A_i$
}
```

Aggregate summaries and report the results (e.g. biases, MSE, quantiles, critical values, etc.)

Cross Validation

Example 3:

Partition data into v folds

- For (i in $1 \cdots v$) {
- Construct training dataset A_i and test dataset $\bar{A}_i$.
- Fit statistical model to dataset A_i
- Obtain results from test dataset $\bar{A}_i$

  }

Aggregate summaries and report the results (e.g. error rates and other performance indicators)

Optimization with multiple starts

Example 4:

- For (i in $1 \cdots m$) {
- Run the optimization algorithm with start i
- Save summaries for start i
  }

Choose the best from these m starts

An application of HKU Grid Point on cross validation

Separable two-dimensional linear discriminant analysis. Jianhua Zhao and Philip L.H. Yu, 2010.

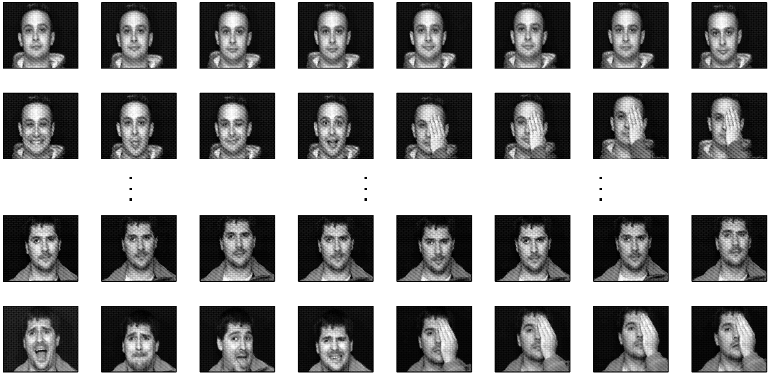


Image size: 96×72 pixels. Dimension reduction.
Find a low-dimensional representation that has the lowest misclassification rate. Commonly estimated by cross validation.

An application of HKU Grid Point on cross validation

- Compare ~ 5 methods for face recognition.
- Examine ~ 10 face image datasets (including large and small sample size)
- For each method and each dataset, consider ~ 4 different training parameters.
- For a given training parameter, perform 50 replications, each of which costs ~ 10 min.
- With 1 CPU, the estimated time is $5 \times 10 \times 4 \times 50 \times 10 \times \frac{1}{60} \times \frac{1}{24} \approx 70$ days.
- Using HKU Grid Point with 4 nodes, the time is ~ 2 days only.

Other potential applications of HKU Grid Point

- Finite distribution of parameter estimation in nonlinear times series.
- Statistical data mining.
- Estimation of risk measurement (e.g. value-at-risk) in financial risk management for a large investment portfolio.
- Spatial-temporal modeling in hydrology and climatology.

Conclusion

- Clearly the HKU Grid Point has greatly enhanced our ability to investigate problems that were previously deemed computationally too complicated and demanding. We are fortunate and grateful to have this powerful facility available.

Thank You!