Ensemble Learning

An Introduction

with examples

by Dr Brad Morantz
bradscientist@machine-cognition.com

Contents

- What it is in simple terms
- Some examples in every day life
- Why implement on a computer
- Computer Implementations
- Random forests
- Other programs
- Pros
- Cons
- References

Definition

- Process by which multiple models, such as classifiers or experts, are strategically generated and combined to solve a particular problem. (note: experts can be biological organisms)
- Uses multiple models to get as good or better performance than the constituent models.
- A contradiction to Occam's Razor?

Long-term Goal

- Make a high quality decision
  - Based upon accurate determination
    - Classification
    - Pattern recognition
    - Identification
  - Modeling and simulation
    - Forecasts
    - Expected results
  - Facts
    - Knowledge
    - Intelligence

Other Uses

- Improve
  - Classification
  - Prediction
  - Function approximation
  - Model performance
- Select features
- Data Fusion
- Incremental Learning

Ensemble Learning

- When the need is for accurate
  - Classification
  - Pattern recognition
  - Identification

Simple Explanation

- A decision maker that listens to other decision makers
- A classifier that has other classifiers as its input

Second Stage Classification

- Simplest is voting
  - Each first stage votes
  - Answer with most votes wins
- Pattern recognition
  - Can be many differences in first stage
  - Second stage can learn patterns
  - More complex but can be more accurate
  - Requires supervised training

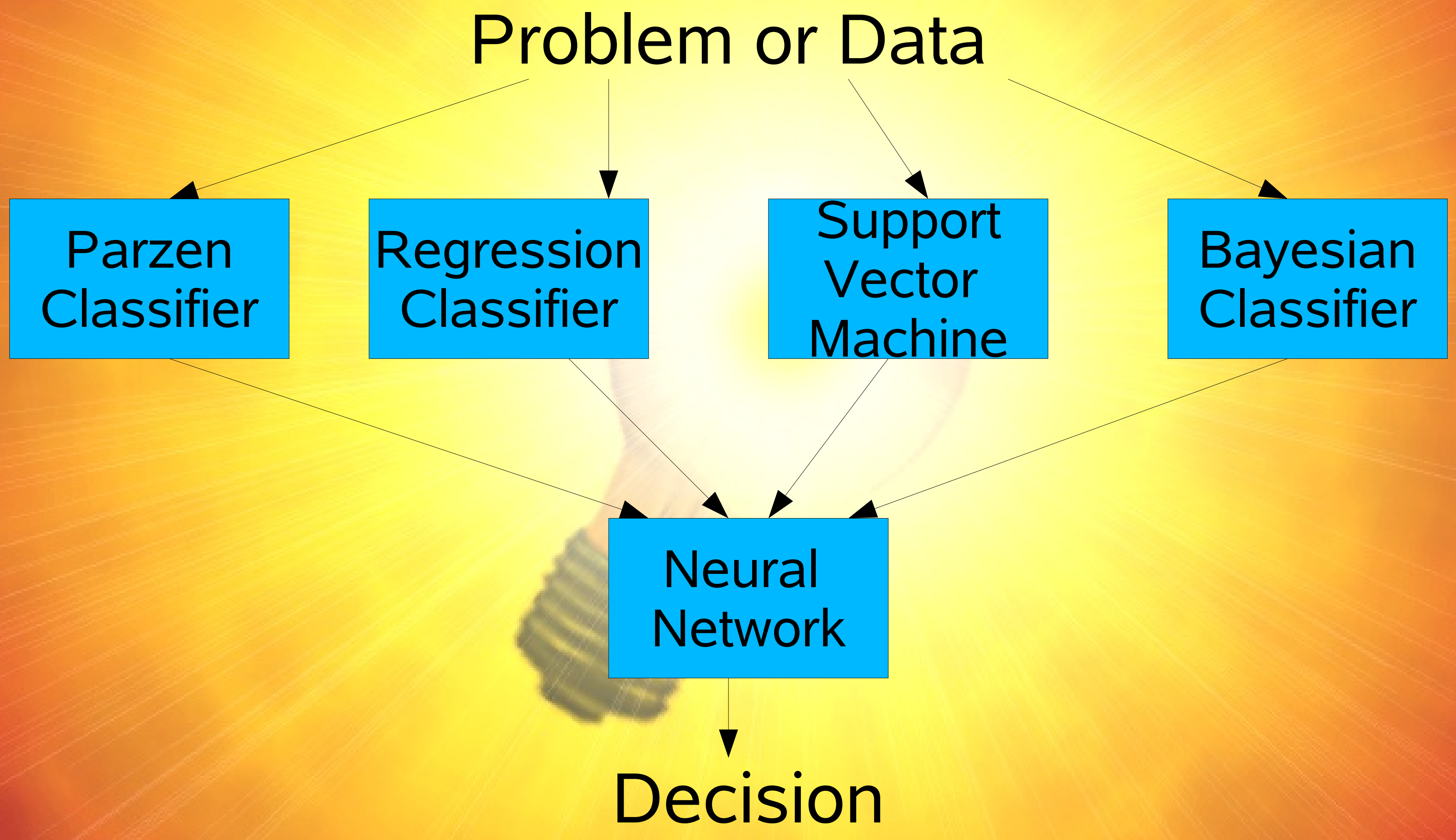
Example

- Government
  - Each cabinet minister gives their conclusion
  - Prime Minister listens to each of them
  - Based upon that information, PM makes decision
- Company
  - Each department head gives his report
  - CEO/President makes decision based on inputs
- Computer
  - Many classifiers feeding into master classifier
  - Master finding pattern of results, makes decision

Human Version

Each person giving their answer on the problem/situation

Mathematical Version



What is Happening?

- Each 'first stage'
  - Uses its expertise and methods
  - In its area of solution space
  - Generates its own answer based on above
- 'Master' classifier
  - Learns area of classification of each first stage
  - Learns patterns of results
  - Generates answer that is as good or better than any single first stage could.

Probabilities

- If had 6 first stages, each with P(error) = 0.1
- If used majority voting
  - Would require at least 4 to be wrong to get wrong classification
  - Assuming independence among classifiers
    - P(wrong decision) = $_6C_4 * 0.1^4 = 1.5 \times 10^{-3}$
    - This increases accuracy greatly
- To reduce correlation models should be diverse as possible

Complex Decision Boundary

- Can be achieved with diversity of individual classifiers in first stage
- Each defines a piece of the boundary
- Each operates in a different section of variable space
- First stage classifiers are independent from each other

Why Implement on a Computer

- Computers are fast
- Can implement multiple first stage classifiers
  - This is just yelling “multi-processing”
- Can handle large matrices of data
- Supervised training can be done

Computer Implementations

- Collection of classifier programs
  - Use shell or Perl script to put together
- Use modular or object oriented programming
- Multiprocessing
  - Assign processors to each first stage
  - Do all first stage at same time
  - If enough processors, then assign to matrix manipulations

Random Forests (TM)

- Leo Breiman & Adele Cutler
- Collection of decision trees
  - Controlled variation
  - Different subset of input variables (features) for each first stage classifier – Randomly chosen
  - Trees are not pruned
- Selects by mode (voting) of class

Random Forests (TM)

Advantages

- Works w/many variables or features
- Estimates the importance of variables/features
- Good with missing data
- This is an experimental way of determining variable interaction

Disadvantages

- Prone to over-fitting in noisy data
- Not as good as other methods on problems with many variables/features

Random Multinomial Logit*

- Like a Random Forest, but using Logit in lieu of decision trees
- Good for multi-class output
- This alleviates the problem of too many dimensions in MNL where it becomes too computationally intensive
- Works with bootstrapping** and bagging**

* Logit is SAS logistic regression, also called Multnomial Logistic Regression (MNL)

** Explained in later slides

Random Naïve Bayes

- Naïve Bayes classifier assumes variables to be conditionally independent
- First stage is a group of naïve Bayes classifiers
  - Random feature selection

Boosting

- AT&T Labs
  - Singer, Shapire, & Freund
    - AdaBoost (Adaptive Boosting)
- Weighted training set
  - Weighted to help reduce error
    - Weight missed observations heavier
  - Iterative process
- Final result is weighted majority of hypotheses

Strengths and Advantages

- Better accuracy
- Good application for multiprocessing
  - Save time while increasing accuracy
- Works with bootstrapping or bagging
  - Random sampling with replenishment
- Works with large data sets
  - Split data among first stage classifiers

Bagging

- Term for Bootstrap Aggregating
- Meta Algorithm to improve classification and regression
- Reduces variance and helps avoid overfitting
- A training method where:
  - Bootstrapping: Training dataset is multiply resampled with replacement
  - Aggregating: combining models
    - Averaging for regression
    - Voting for classification

References

- Morantz, B, www.machine-cognition.com
- Polikar, R, *Ensemble Learning*, Scholarpedia, 4(1):2776, 2009
- Russell, S.J. & Norvig, P., *Artificial Intelligence: A Modern Approach,* Second Edition, Prentice-Hall: 2003
- Prinze & Van den Poel, *Random Forests for Multiclass Classification: Random Multinomial Logit*, Expert Systems with Applications, vol 34, #3, April 2008
- IEEE Computational Intelligence Society www.ieee-cis.org