https://www.slideshare.net/ColleenFarrelly/machine-learning-by-analogy-59094152

Machine Learning by Analogy (Updated)

Colleen M. Farrelly

Overview of Problem

- Many machine learning methods exist in the literature and in industry.
  - What works well for one problem may not work well for the next problem.
  - In addition to poor model fit, an incorrect application of methods can lead to incorrect inference.
    - Implications for data-driven business decisions.
    - Low future confidence in data science and its results.
    - Lower quality software products.
- Understanding the intuition and mathematics behind these methods can ameliorate these problems.
  - This talk focuses on building intuition.
  - Links to theoretical papers underlying each method.

P L A N

- Régressions paramétriques et extensions
- Régressions semi paramétriques et extensions
- Régressions non paramétriques via méthodes d’optimisation
- Combinaisons de méthodes supervisées et non supervisées
- Apprentissage non supervisé
- Séries chronologiques

Multiple Regression

- Total variance of a normally-distributed outcome as cookie jar.
- Error as empty space.
- Predictors accounting for pieces of the total variance as cookies.
  - Based on relationship to predictor and to each other.
    - Cookies accounting for the same piece of variance as those smooshed together.
- Many statistical assumptions need to be met.

www.zelcoviacookies.com

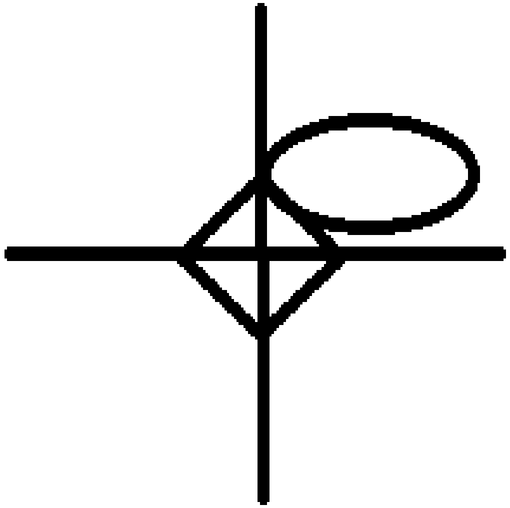
Generalized Linear Models

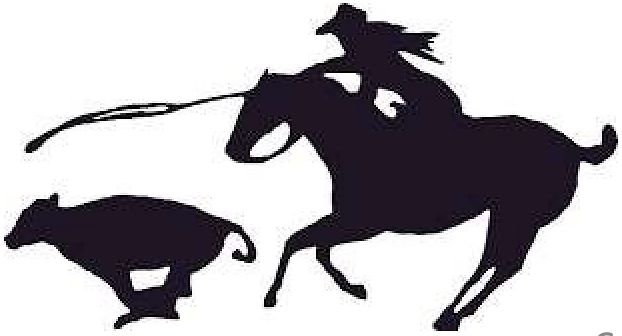
tomstock.photoshelter.com

- Extends multiple regression.
  - Many types of outcome distributions.
  - Transform an outcome variable to create a linear relationship with predictors.
- Sort of like silly putty stretching the outcome variable in the data space.
- Does not work on high-dimensional data where predictors > observations.
- Only certain outcome distributions.
  - Exponential family as example.

LASSO, Ridge Regession, and Elastic Net

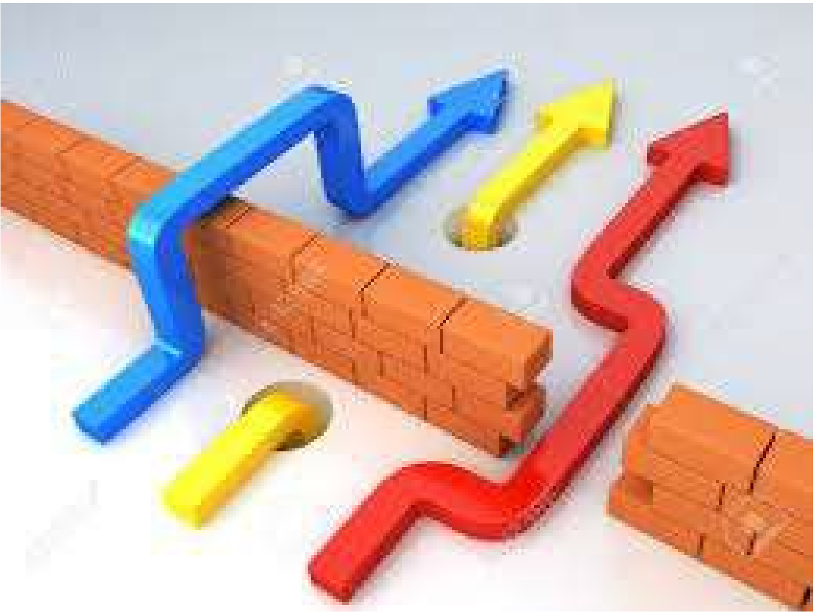
- Impose size and/or variable overlap constraints (penalties) on generalized linear models.
  - Elastic net as hybrid of these constraints.
  - Can handle large numbers of predictors.
- Reduce the number of predictors.
  - Shrink some predictor estimates to 0.
  - Examine sets of similar predictors.
- Similar to eating irrelevant cookies in the regression cookie jar or a cowboy at the origin roping coefficients that get too close





Homotopy-Based LASSO/LARS

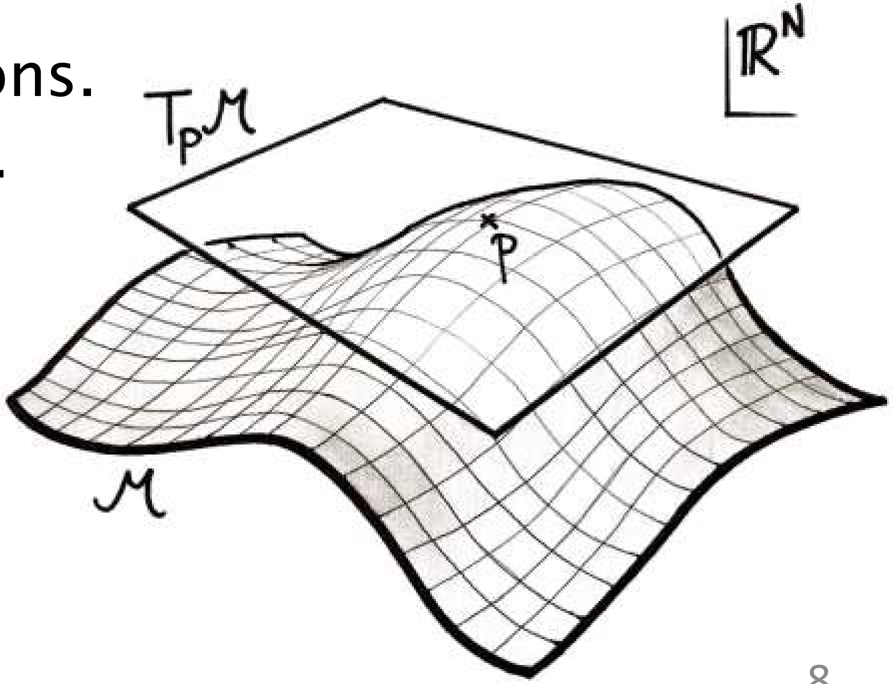
- Homotopy as path equivalence
  - Intrinsic property of topological spaces (such as data manifolds)



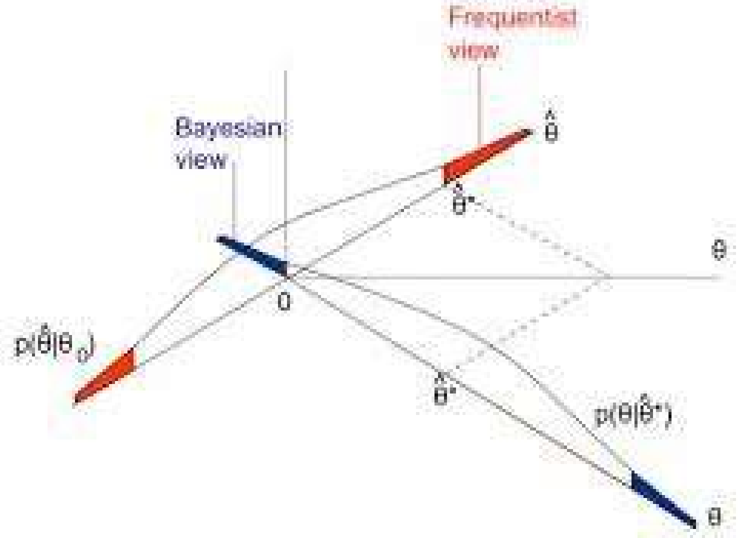
- Homotopy arrow example
  - Red and blue arrows can be deformed into each other by wiggling and stretching the line path with anchors at start and finish of line
  - Yellow arrow crosses holes and would need to backtrack or break to the surface to freely wiggle into the blue or red line
- Homotopy method in LASSO/LARS wiggles an easy regression path into an optimal regression path
  - Avoids obstacles that can trap other regression estimators (peaks, valleys, saddles…)

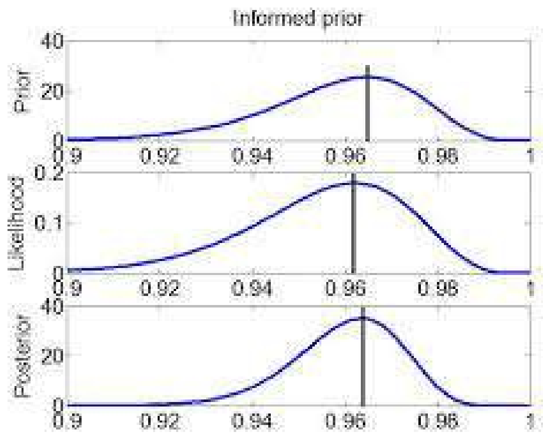
Differential Geometry and Regression

- Instead of fitting model to data, fit model to tangent space (what isn't the data).
  - Deals with collinearity, as parallel vectors share a tangent space (only one selected of collinear group).
  - LASSO and LARS extensions.
  - Rao scoring for selection.
    - Effect estimates (angles).
    - Model selection criteria.
      - Information criteria.
      - Deviance scoring.



Bayesian Regression Models





- Fit all possible models and figure out likelihood of each given observed data.
  - Based on Bayes' Theorem and conditional probability.
  - Instead of giving likelihood that data came from a specific parameterized population (univariate), figure out likelihood of set of data coming from sets of population (multivariate).
  - Can select naïve prior (no assumptions on model or population) or make an informed guess (assumptions about population or important factors in model).
  - Combine multiple models according to their likelihoods into a blended model given data.

CHAPTER 2

Semi-Parametric Extensions of Regression

Spline Models

- Extends a generalized linear models by estimating unknown (nonlinear) functions of an outcome on intervals of a predictor.

- Use of "knots" to break function into intervals.
- Similar to a downhill skier.
  - Slope as a function of outcome.
  - Skier as spline.
  - Flags as knots anchoring the skier as he travels down the slope.

www.dearsportsfan.com

MARS

- Multivariate adaptive regression splines as an extension of spline models to multivariate

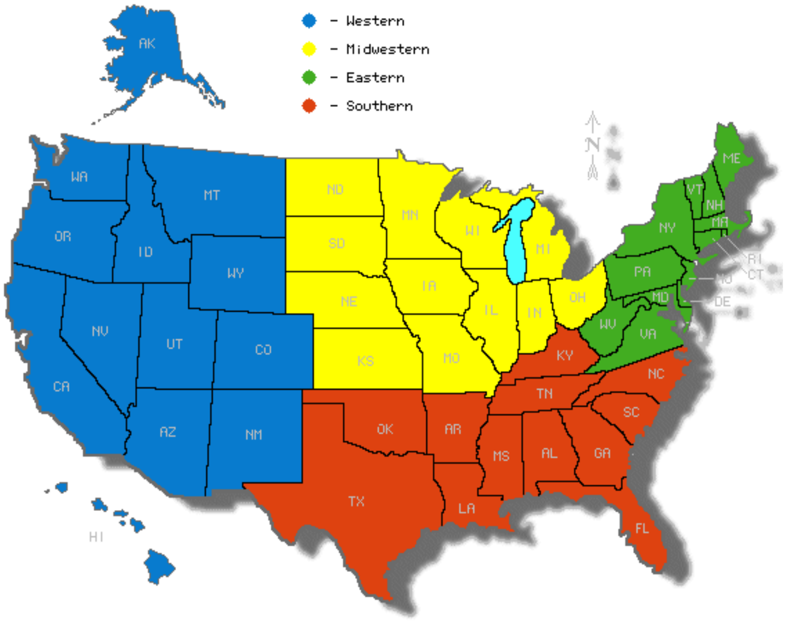
  data
  - Knots placed according to multivariate structure and splines fit between knots
  - Much like fixing a rope to the peaks and valleys of a mountain range, where the rope has enough slack to hug the terrain between its fixed points

Generalized Additive Models

www.filigreeinn.com

- Extends spline models to the relationships of many predictors to an outcome.
  - Also allows for a “silly-putty” transformation of the outcome variable.
- Like multiple skiers on many courses and mountains to estimate many relationships in the dataset.

Piece-Wise Regression

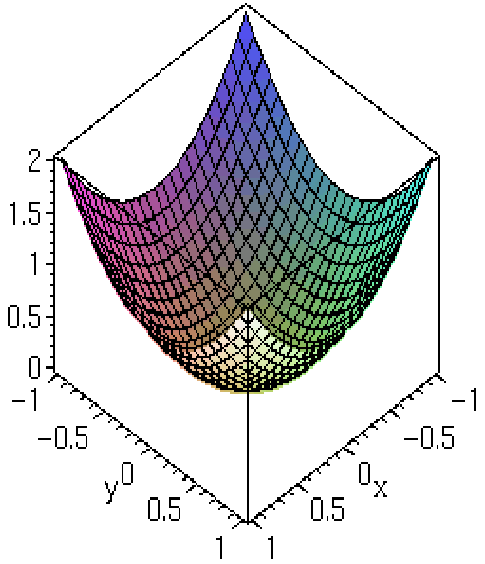


www.pinterest.com

- Chop data into partitions and then fit multiple regression models to each partition.
- Divide-and-conquer approach.
- Examples:
  - Multivariate adaptive regression splines
  - Regression trees
  - Morse-Smale regression

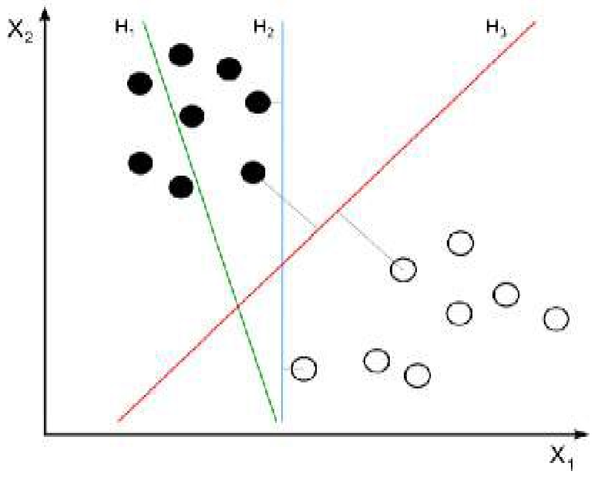
Morse–Smale Regression

- Based on a branch of math called **topology.**
  - Study of changes in function behavior on shapes.
  - Used to classify similarities/ differences between shapes.
  - Data clouds turned into discrete shape combinations (simplices).
- Use these principles to partition data and fit elastic net models to each piece.
  - Break data into multiple toy sets.
  - Analyze sets for underlying properties of each toy.
- Useful visual output for additional data mining.



Support Vector Regression

- Linear or non-linear support beams used to separate group data for classification or map data for kernel- based regression.



http://en.wikipedia.org/wiki/Support_vector_machine

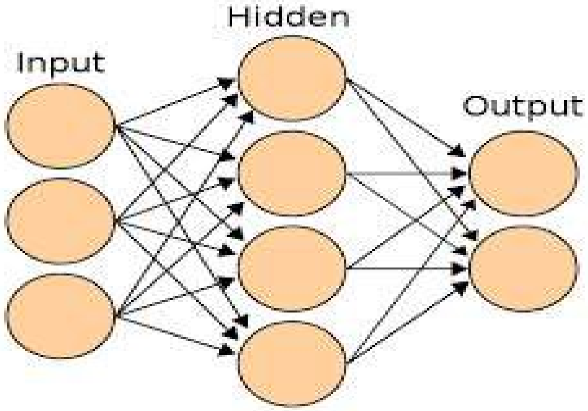
- Much like scaffolding and support beams separating and holding up parts of a high rise.

leadertom.en.ec21.com

Neural Networks

www.alz.org

- Based on processing complex, nonlinear information the way the human brain does via a series of feature mappings.

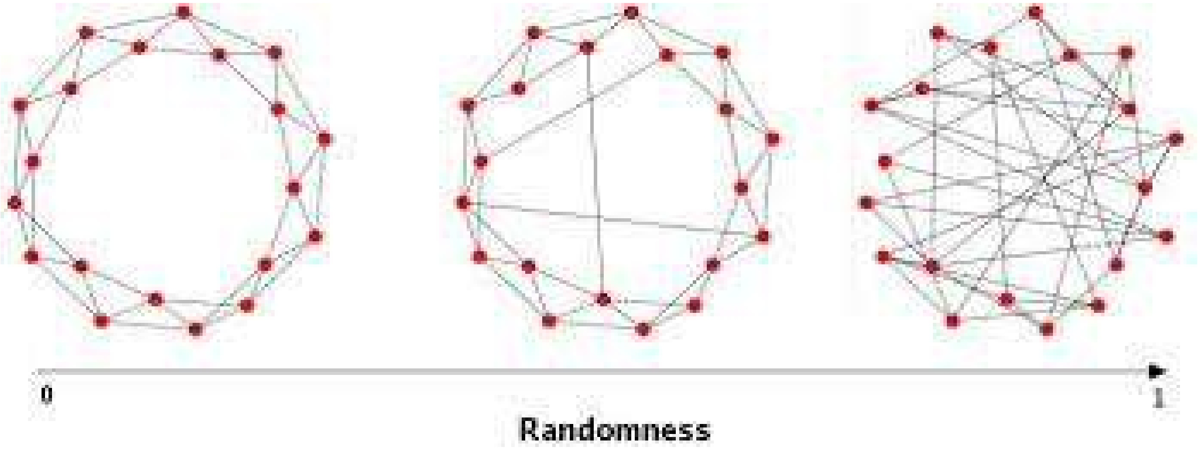


Arrows denote mapping functions, which take one topological space to another

colah.github.io

Extreme Learning Machines

- Type of shallow, wide neural network.
- Reduces framework to a penalized linear algebra problem, rather than iterative training (much faster to solve).
- Based on random mappings.
- Shown to converge to correct classification/regression (universal approximation property—may require unreasonably wide networks).
- Semi-supervised learning extensions.



Deep Learning

- Added layers in neural network to solve width problem in single-layer networks for universal approximation.
- More effective in learning features of the data.
- Like sifting data with multiple sifters to distill finer and finer pieces of the data.
- Computationally intensive and requires architecture design and optimization.

Tensor Flow

- Recent extension of deep learning framework from spreadsheet data to multiple simultaneous spreadsheets
  - Spreadsheets may be of the same dimension or different dimensions
  - Could process multiple or hierarchical networks via adjacency matrices
- Like sifting through data with multiple inputs of varying sizes and textures

CHAPTER 3

Nonparametric Regression via Optimization Methods

Single Decision Tree Models

tvtropes.org

- Classifies data according to optimal partitioning of data (visualized as a high-dimensional cube).
- Slices data into squares, cubes, and hypercubes (4+ dimensional cubes).
  - Like finding the best place to cut through the data with a sword.
- Option to prune tree for better model (glue some of the pieces back together).
- Notoriously unstable.
- Optimization possible to arrive at best possible trees (genetic algorithms).

Gradient Descent Methods

- Optimization technique.
  - Minimize a loss function in regression.
- Accomplished by choosing a variable per step that achieves largest minimization.
- Climber trying to descend a mountain.
  - Minimize height above sea level per step.
- Directions (N, S, E, W) as a collection of variables.
- Climber rappels down cliff according to these rules.

www.chinatravelca.com

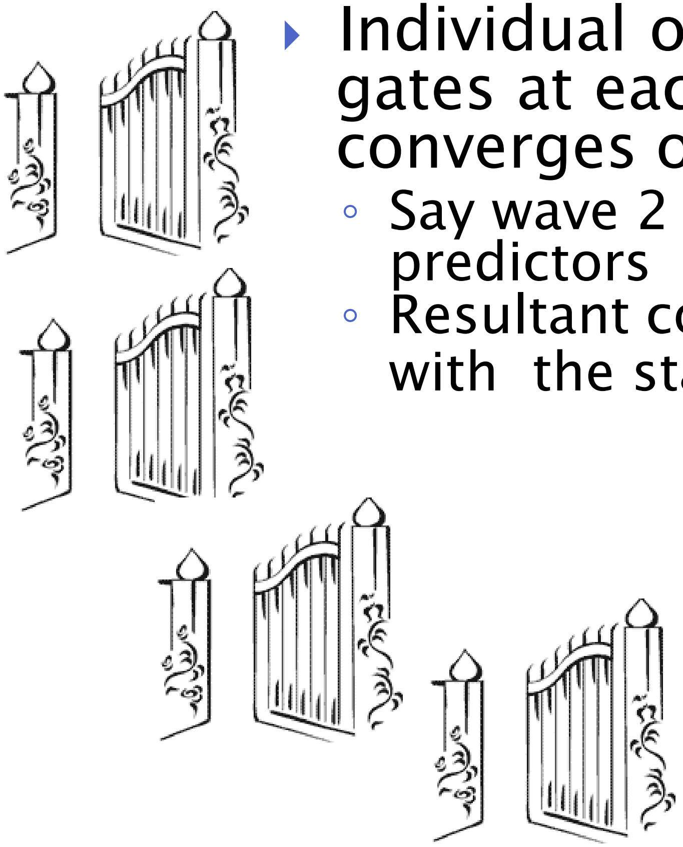
Genetic Algorithms

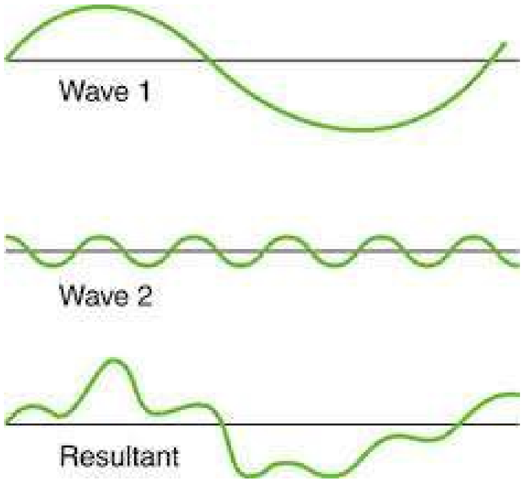
wallpapers.brothersoft.com

- Optimization helpers.
  - Find global maximum or minimum by searching many areas simultaneously.
  - Can impose restrictions.
- Teams of mountain climbers trying to find the summit.
- Restrictions as climber supplies, hours of daylight left to find summit…
- Genetic algorithm as population of mountain climbers each climbing on his/her own.

Quantum Evolutionary Models

- One climber exploring multiple routes simultaneously.
- Individual optimized through a series of gates at each update until superposition converges on one state
  - Say wave 2 is the optimal combination of predictors
  - Resultant combination slowly flattens to wave 2, with the states of wave 1disappearing

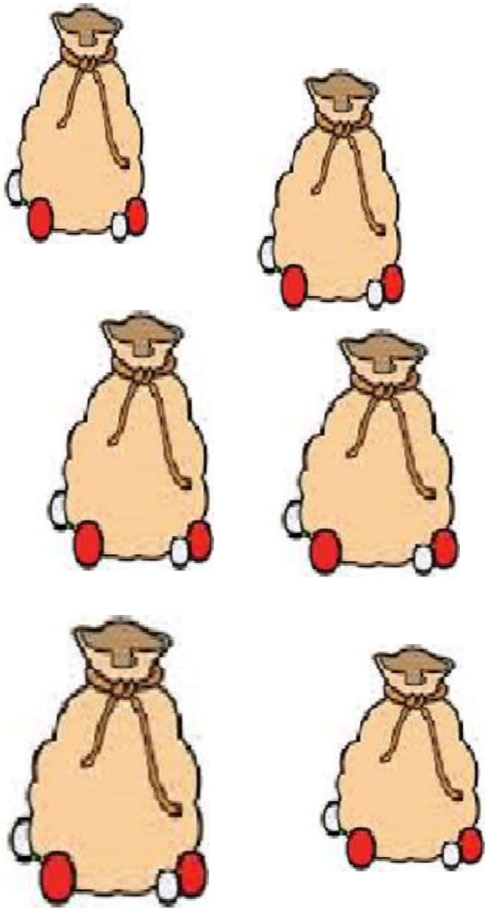




http://philschatz.com/physics-book/contents/m42249.html

Ensemble Methods

- Combine multiple models of the same type (ex. trees) for better prediction.
  - Single models unstable.
    - Equally-good estimates from different models.
  - Use bootstrapping.
    - Multiple “marble draws” followed by model creation.
  - Creates diversity of features.
  - Creates models with different biases and error.



Boosted Regression Models

play.google.com

- Uses gradient descent algorithm to model an outcome based on predictors (linear, tree, spline…).
  - Proceeds in steps, adding variables and adjusting weights according to the algorithm.
- Like putting together a puzzle.
  - Algorithm first focuses on most important parts of the picture (such as the Mona Lisa's eyes).
  - Then adds nuances that help identify other missing pieces (hands, landscape…).

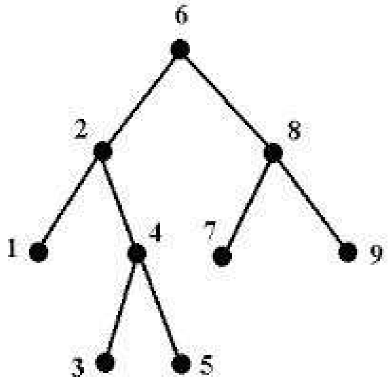
XGBoost

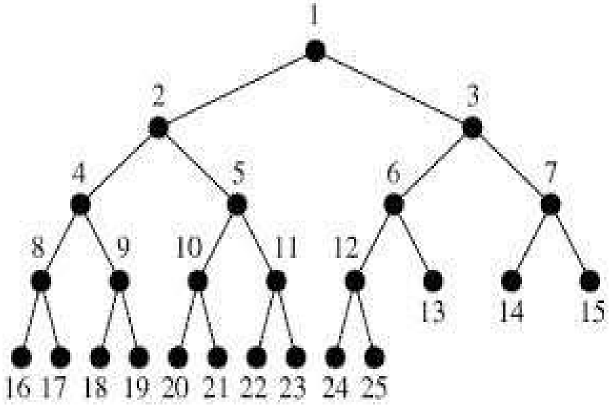
- Adds a penalty term to boosted regression
  - Can be formulated as LASSO/ridge regression/elastic net with constraints
  - Also leverages hardware to further speed up boosting ensemble

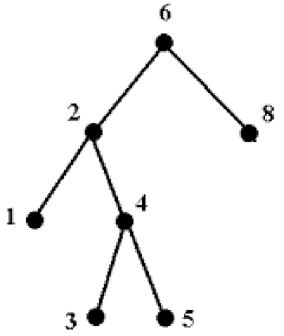
play.google.com

Random Forests: Multiple Trees

- Single tree models poor predictors.
- Grow a forest of them for a strong predictor (another ensemble method).
- Algorithm:
  - Takes random sample of data.
  - Builds one tree.
  - Repeats until forest grown.
  - Averages across forest to identify strong predictors.







CHAPTER 4

Combining Supervised and Unsupervised Methods

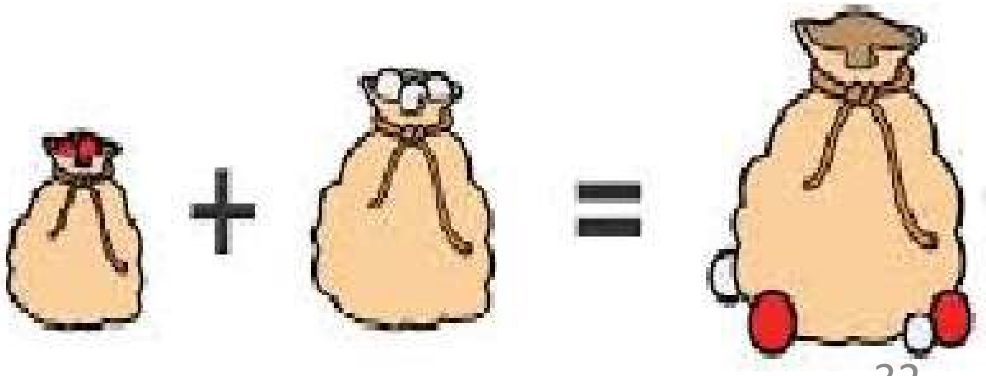
Dimension Reduction

marmaladeandmileposts.com

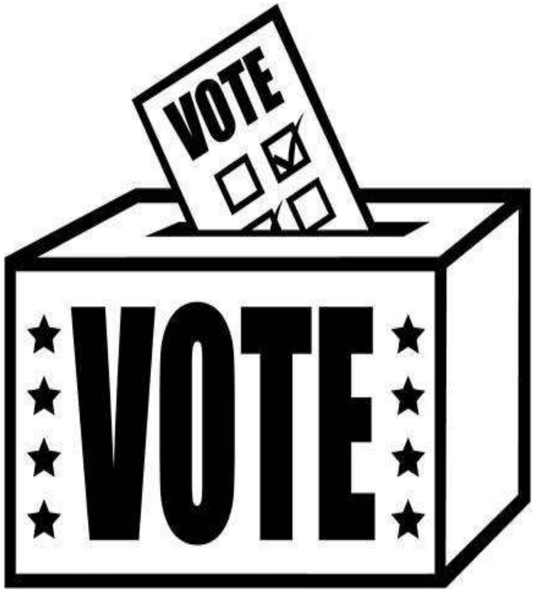
- Principle Component Analysis (PCA)
  - Variance partitioning to combine pieces of variables into new linear subspace.
  - Smashing cookies by kind and combining into a flat hybrid cookie in previous regression model.
- Manifold Learning
  - PCA-like algorithms that combine pieces into a new nonlinear subspace.
  - Non-flat cookie combining.
- Useful as pre-processing step for prediction models.
  - Reduce dimension.
  - Obtain uncorrelated, non-overlapping variables (bases).

Random Forest as Extension Example

- Balanced sampling for low-frequency predictors.
  - Stratified samples (i.e. sample from bag of mostly white marbles and few red marbles with constraint that 1/5th of draws must be red marbles).
- Dimension reduction/mapping pre-processing
  - Principle component, manifold learning…
  - Hybrid of neural network methods and tree models.



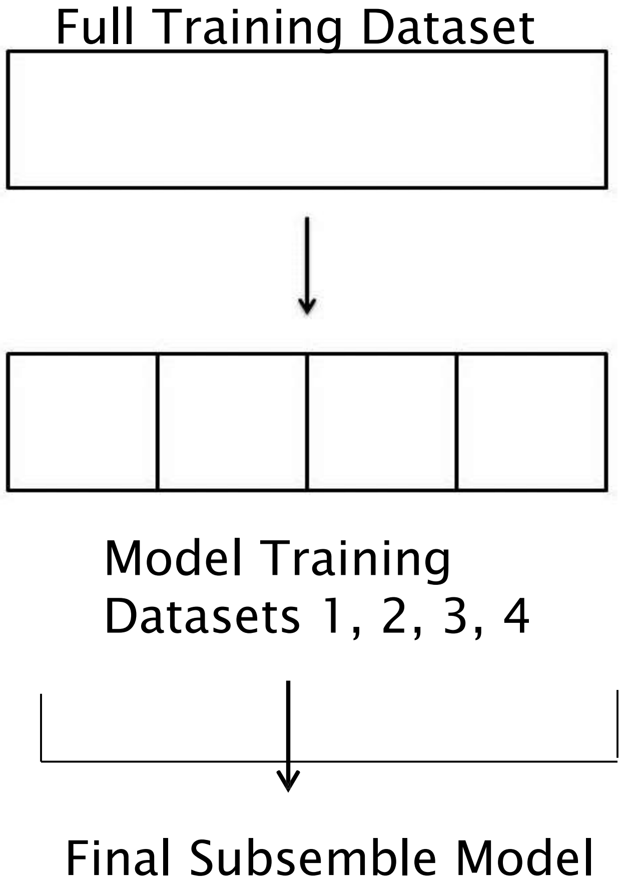
Superlearners



- Aggregation of multiple types of models.
  - Like a small town election.
  - Different people have different views of the politics and care about different issues.
- Different modeling methods capture different pieces of the data and vote in different pieces.
  - Leverage strengths, minimize weaknesses
  - Diversity of methods to better explore underlying data geometry
- Avoids multiple testing issues.

Subsembles

- Subsembles
  - Partition data into training sets
    - Randomly selected or through partition algorithm
  - Train a model on each data partition
  - Combine into final weighted prediction model
- This is similar to national elections.
  - Each elector in the electoral college learns for whom his constituents voted.
  - The final electoral college pools these individual votes.

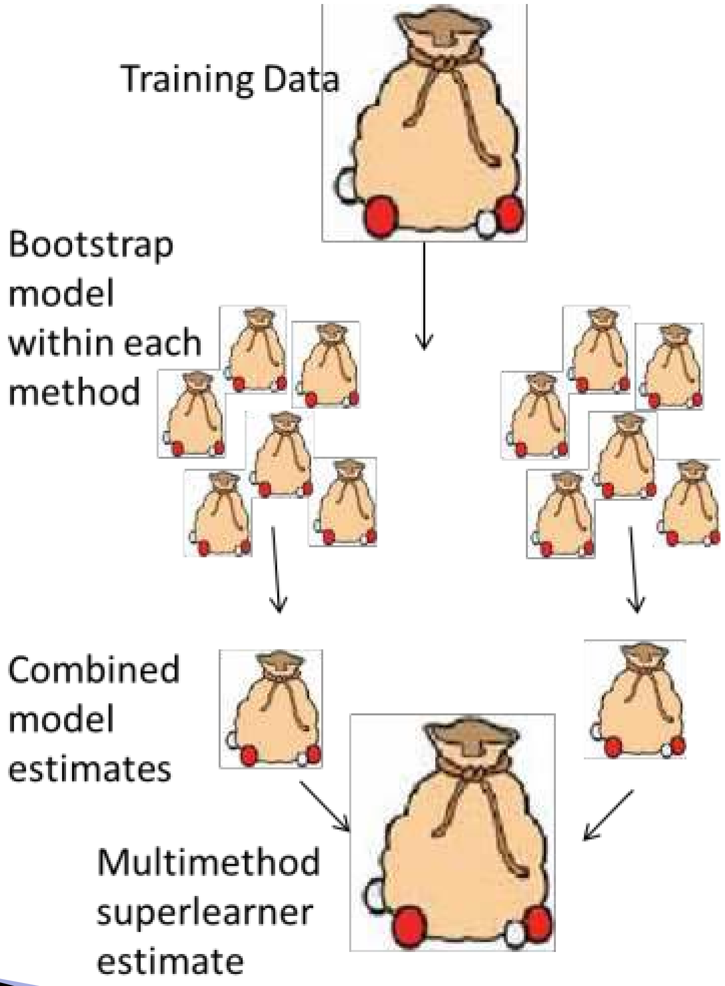


Full Training Dataset

Model Training Datasets 1, 2, 3, 4

Final Subsemble Model

Supersemble


Training Data
Bootstrap model within each method
Combined model estimates
Multimethod superlearner estimate

- Combines superlearning and subsembles for better prediction
  - Diversity improves subsemble method (better able to explore data geometry)
  - Bootstrapping improves superlearner pieces (more diversity within each method)
  - Preliminary empirical evidence shows efficacy of combination.

CHAPTER 5

Unsupervised Learning

K Nearest Neighbors

- Classification of a data point based on the classification of its nearest neighboring points.
- Like a junior high lunchroom and student clicks.
  - Students at the same table tend to be more similar to each other than to a distant table.

www.wrestlecrap.com

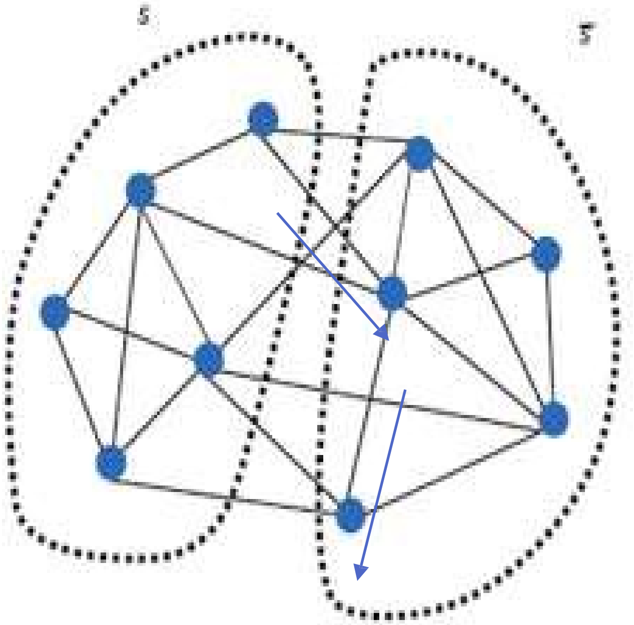
K-Means Clustering

krazymommakreations.blogspot.com

- Iteratively separating groups within a dataset based on similarity.
- Like untangling tangled balls of yarn into multiple, single-colored balls (ideally).
  - Each step untangles the mess a little further.
  - Few stopping guidelines (when it looks separated).

Graph-Based Techniques

- Hybrid of supervised and unsupervised learning (groupings and prediction).
- Uses graphical representation of data and its relationships.
- Algorithms with connections to topology, differential geometry, and Markov chains.
- Useful in combination with other machine learning methods to provide extra insight (ex. spectral clustering).



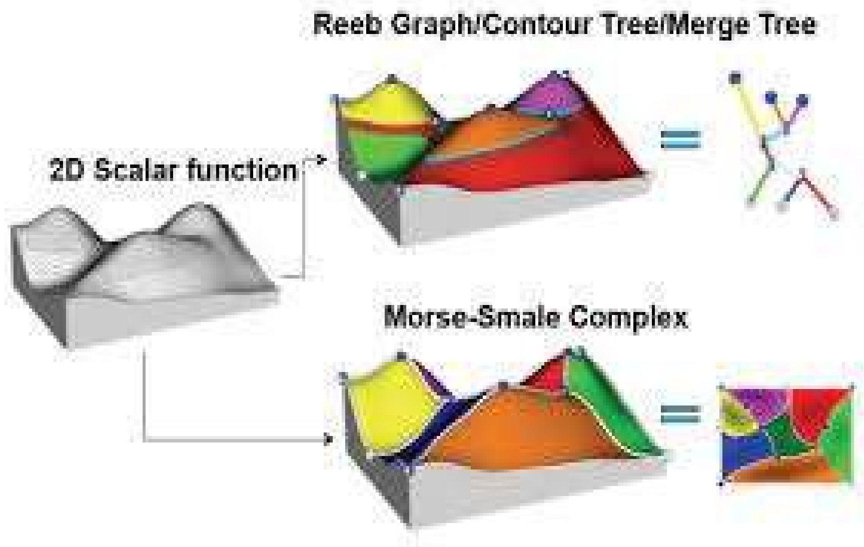
Spectral Clustering

- K-means algorithm with weighting and dimension reduction components of similarity measure.
  - Simplify balls of string to warm colors and cool colors before untangling.
- Can be reformulated as a graph clustering problem.
  - Partition subcomponents of a graph based on flow equations.

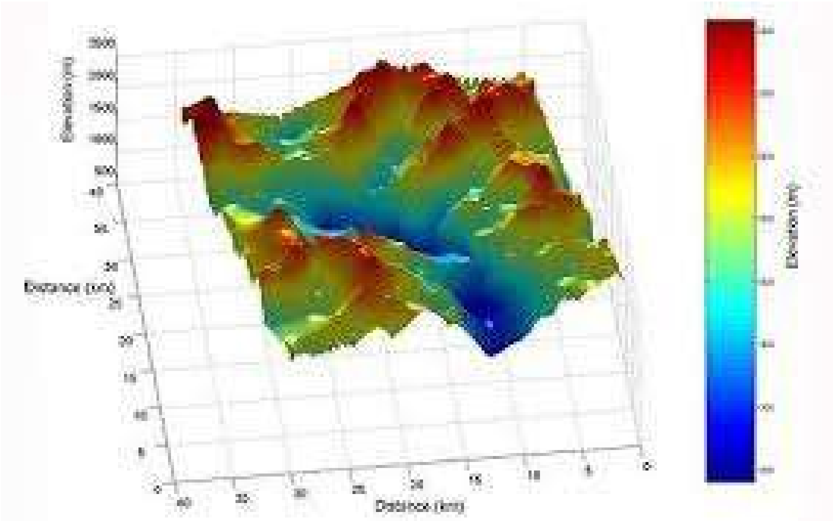
www.simplepastimes.com

Morse-Smale Mode Clustering

- Multivariate technique similar to mode or density clustering.
  - Find peaks and valleys in data according to an input function on the data (level set slices)—much like a watershed on mountains.

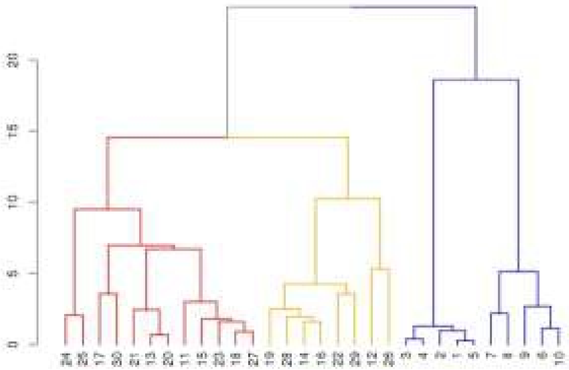


  - Separate data based on shared peaks and valleys across slices (shared multivariate density/gradient).
  - Many nice theoretical developments on validity and convergence.



Mapper Clustering

- Topological clustering.
  - Define distance metric.
  - Slice multidimensional dataset with Morse function.
  - Examine function behavior across slice.
  - Cluster function behavior.
  - Iterate through multiple slices to obtain hierarchy of function behavior.
- Much like examining the behavior of multiple objects across a flip book.
  - Nesting
  - Cluster overlap

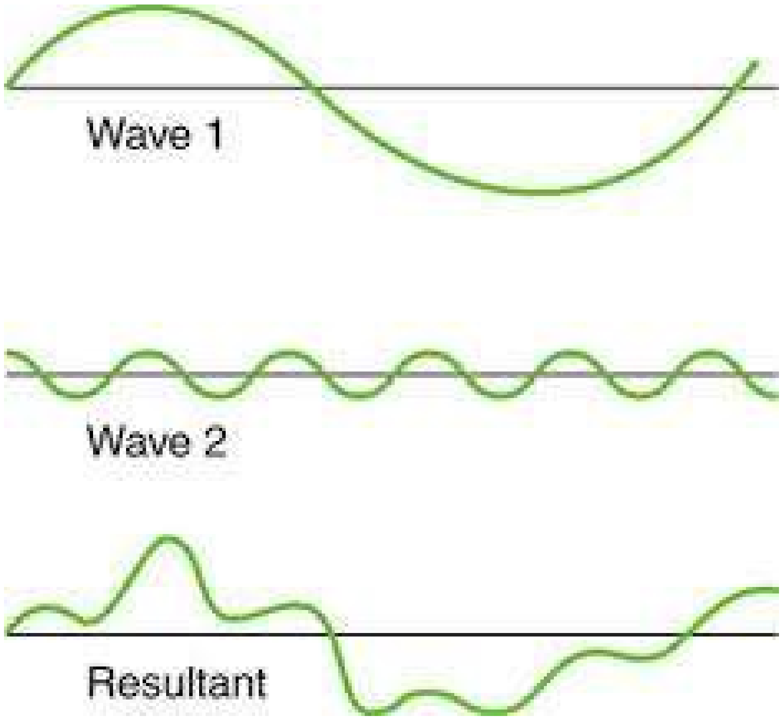


Filtered functions then used to create various resolutions of a modified Reeb graph summary of topology.

CHAPTER 6

Time Series Forecasting

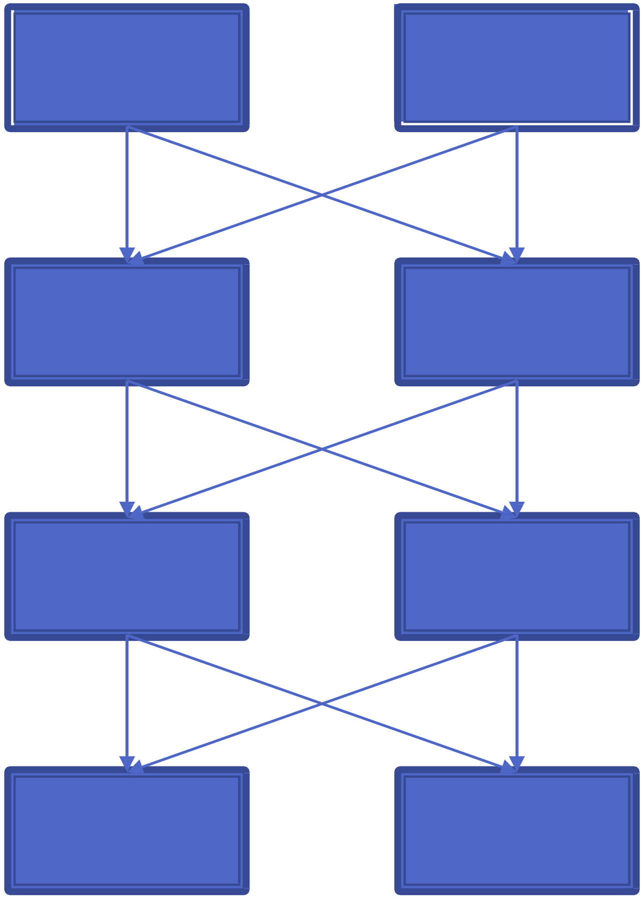
ARIMA Models



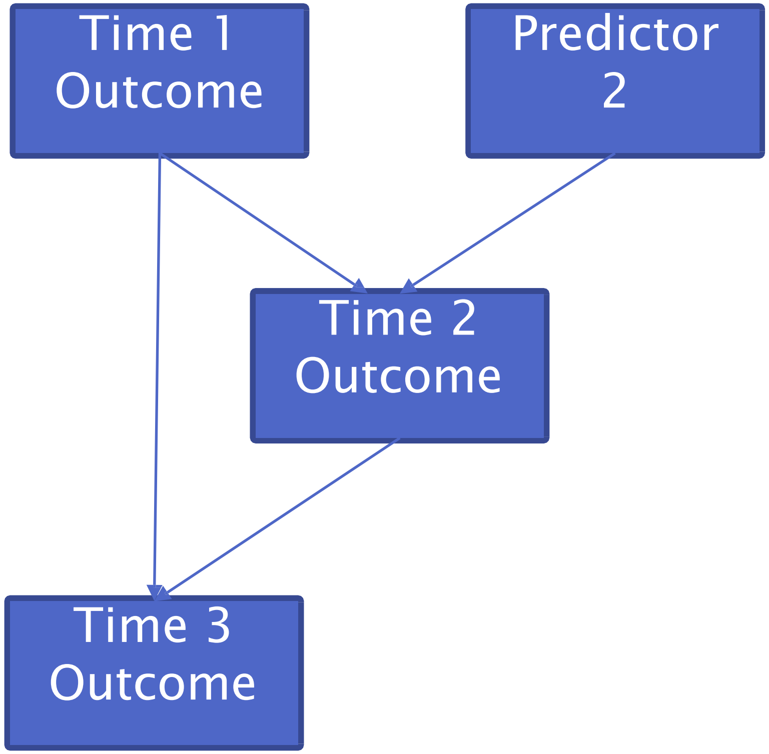
- Similar to decomposing superposed states
  - Seasonal trends
  - Yearly trends
  - Trend averages
  - Dependencies on previous time point
- Knit individual forecasted pieces into a complete forecast by superposing these individual forecasts
- Several extensions to neural networks, time-lagged machine learning models…

Structural Equation Models (SEM)

- A time-series method incorporating predictors
    - Constant predictors at initial time point
    - Varying predictors at multiple time points
- Creates a sort of correlation web between predictors and time points
    - Can handle multiple time lags and multivariate outcomes
    - Can handle any GLM outcome links
- Related to partial differential equations of dynamic systems



Bayesian Networks

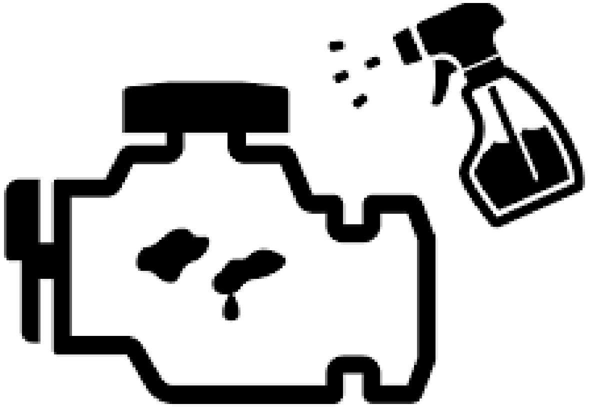


- Data-based mining for SEM

  relationships/time-lag components
  - Leverages conditional probability between predictors to find dependencies
  - Does not require a priori model formulation like SEM
- Peeking at data to create a dependency web over time or predictors/outcome
- Can be validated by a follow-up SEM based on network structure

Singular Spectrum Analysis

- Technically:
  - Matrix decomposition (similar to PCA/manifold learning)
  - Followed by spectral methods
  - Cleaning of time-lagged covariance matrix
  - Reconstruction with simple forecast
- Kind of like deconstructing, cleaning, a rebuilding a car engine



Extensible Markov Models

Like remembering classes of chess board configurations across games played

- Combines k-nearest neighbors-based clustering with memoryless state changes converging to a transition distribution (weighted directed graph)
  - Reduce an observation to a pattern
  - Remember patterns seen (across time or space)
  - Match new observations to this set of patterns
  - Computationally more feasible than k-means clustering

Conclusions

- Many machine learning methods exist today, and many more are being developed every day.
- Methods come with certain assumptions that must be met.
  - Breaking assumptions can lead to poor model fit or incorrect inference.
  - Matching a method to a problem not only can help with better prediction and inference; it can also lead to faster computation times and better presentations to clients.
- Development depends on problem-matching and deep understanding of the mathematics behind the methods used.

Machine Learning by Analogy by Colleen Farrelly

https://www.slideshare.net/ColleenFarrelly/machine-learning-by-analogy-59094152

NOTES

2. } Many machine learning methods exist in the literature and in industry. ◦ What works well for one problem may not work well for the next problem. ◦ In addition to poor model fit, an incorrect application of methods can lead to incorrect inference. □ Implications for data-driven business decisions. □ Low future confidence in data science and its results. □ Lower quality software products. } Understanding the intuition and mathematics behind these methods can ameliorate these problems. ◦ This talk focuses on building intuition. ◦ Links to theoretical papers underlying each method.

3. } Total variance of a normally- distributed outcome as cookie jar. } Error as empty space. } Predictors accounting for pieces of the total variance as cookies. ◦ Based on relationship to predictor and to each other. □ Cookies accounting for the same piece of variance as those smooshed together. } Many statistical assumptions need to be met. www.zelcoviacookies.com

4. } Extends multiple regression. ◦ Many types of outcome distributions. ◦ Transform an outcome variable to create a linear relationship with predictors. } Sort of like silly putty stretching the outcome variable in the data space. } Does not work on high- dimensional data where predictors > observations. } Only certain outcome distributions. ◦ Exponential family as example. tomstock.photoshelter.com

5. } Impose size and/or variable overlap constraints (penalties) on generalized linear models. ◦ Elastic net as hybrid of these constraints. ◦ Can handle large numbers of predictors. } Reduce the number of predictors. ◦ Shrink some predictor estimates to 0. ◦ Examine sets of similar predictors. } Similar to eating irrelevant cookies in the regression cookie jar or a cowboy at the origin roping coefficients that get too close

6. Homotopy arrow example ◦ Red and blue arrows can be deformed into each other by wiggling and stretching the line path with anchors at start and finish of line ◦ Yellow arrow crosses holes and would need to backtrack or break to the surface to freely wiggle into the blue or red line } Homotopy method in LASSO/LARS wiggles an easy regression path into an optimal regression path ◦ Avoids obstacles that can trap other regression estimators (peaks, valleys, saddles…) } Homotopy as path equivalence ◦ Intrinsic property of topological spaces (such as data manifolds)

7. } Instead of fitting model to data, fit model to tangent space (what isn't the data). ◦ Deals with collinearity, as parallel vectors share a tangent space (only one selected of collinear group). ◦ LASSO and LARS extensions. ◦ Rao scoring for selection. □ Effect estimates (angles). □ Model selection criteria. □ Information criteria. □ Deviance scoring.

8. } Fit all possible models and figure out likelihood of each given observed data. ◦ Based on Bayes' Theorem and conditional probability. ◦ Instead of giving likelihood that data came from a specific parameterized population (univariate), figure out likelihood of set of data coming from sets of population (multivariate). ◦ Can select naïve prior (no assumptions on model or population) or make an informed guess (assumptions about population or important factors in model). ◦ Combine multiple models according to their likelihoods into a blended model given data.

9. Semi-Parametric Extensions of Regression

10. } Extends a generalized linear models by estimating unknown (nonlinear) functions of an outcome on intervals of a predictor. } Use of "knots" to break function into intervals. } Similar to a downhill skier. ◦ Slope as a function of outcome. ◦ Skier as spline. ◦ Flags as knots anchoring the skier as he travels down the slope. www.dearsportsfan.com

11. } Multivariate adaptive regression splines as an extension of spline models to multivariate data ◦ Knots placed according to multivariate structure and splines fit between knots ◦ Much like fixing a rope to the peaks and valleys of a mountain range, where the rope has enough slack to hug the terrain between its fixed points

12. } Extends spline models to the relationships of many predictors to an outcome. ◦ Also allows for a “silly- putty” transformation of the outcome variable. } Like multiple skiers on many courses and mountains to estimate many relationships in the dataset. www.filigreeinn.com

13. } Chop data into partitions and then fit multiple regression models to each partition. } Divide-and-conquer approach. } Examples: ◦ Multivariate adaptive regression splines ◦ Regression trees ◦ Morse-Smale regression www.pinterest.com

14. } Based on a branch of math called topology. ◦ Study of changes in function behavior on shapes. ◦ Used to classify similarities/ differences between shapes. ◦ Data clouds turned into discrete shape combinations (simplices). } Use these principles to partition data and fit elastic net models to each piece. ◦ Break data into multiple toy sets. ◦ Analyze sets for underlying properties of each toy. } Useful visual output for additional data mining.

15. } Linear or non-linear support beams used to separate group data for classification or map data for kernel- based regression. } Much like scaffolding and support beams separating and holding up parts of a high rise. http://en.wikipedia.org/wiki/Support_vector_machine leadertom.en.ec21.com

16. } Based on processing complex, nonlinear information the way the human brain does via a series of feature mappings. colah.github.io www.alz.org Arrows denote mapping functions, which take one topological space to another

17. } Type of shallow, wide neural network. } Reduces framework to a penalized linear algebra problem, rather than iterative training (much faster to solve). } Based on random mappings. } Shown to converge to correct classification/regression (universal approximation property—may require unreasonably wide networks). } Semi-supervised learning extensions.

18. } Added layers in neural network to solve width problem in single-layer networks for universal approximation. } More effective in learning features of the data. } Like sifting data with multiple sifters to distill finer and finer pieces of the data. } Computationally intensive and requires architecture design and optimization.

19. } Recent extension of deep learning framework from spreadsheet data to multiple simultaneous spreadsheets ◦ Spreadsheets may be of the same dimension or different dimensions ◦ Could process multiple or hierarchical networks via adjacency matrices } Like sifting through data with multiple inputs of varying sizes and textures

20. CHAPTER 3

21. } Classifies data according to optimal partitioning of data (visualized as a high- dimensional cube). } Slices data into squares, cubes, and hypercubes (4+ dimensional cubes). ◦ Like finding the best place to cut through the data with a sword. } Option to prune tree for better model (glue some of the pieces back together). } Notoriously unstable. } Optimization possible to arrive at best possible trees (genetic algorithms). tvtropes.org

22. } Optimization technique. ◦ Minimize a loss function in regression. } Accomplished by choosing a variable per step that achieves largest minimization. } Climber trying to descend a mountain. ◦ Minimize height above sea level per step. } Directions (N, S, E, W) as a collection of variables. } Climber rappels down cliff according to these rules. www.chinatravelca.com

23. } Optimization helpers. ◦ Find global maximum or minimum by searching many areas simultaneously. ◦ Can impose restrictions. } Teams of mountain climbers trying to find the summit. } Restrictions as climber supplies, hours of daylight left to find summit… } Genetic algorithm as population of mountain climbers each climbing on his/her own. wallpapers.brothersoft.com

24. } One climber exploring multiple routes simultaneously. } Individual optimized through a series of gates at each update until superposition converges on one state ◦ Say wave 2 is the optimal combination of predictors ◦ Resultant combination slowly flattens to wave 2, with the states of wave 1disappearing http://philschatz.com/physics -book/contents/m42249.html

25. } Combine multiple models of the same type (ex. trees) for better prediction. ◦ Single models unstable. □ Equally-good estimates from different models. ◦ Use bootstrapping. □ Multiple “marble draws” followed by model creation. ◦ Creates diversity of features. ◦ Creates models with different biases and error.

26. } Uses gradient descent algorithm to model an outcome based on predictors (linear, tree, spline…). ◦ Proceeds in steps, adding variables and adjusting weights according to the algorithm. } Like putting together a puzzle. ◦ Algorithm first focuses on most important parts of the picture (such as the Mona Lisa’s eyes). ◦ Then adds nuances that help identify other missing pieces (hands, landscape…). play.google.com

27. } Adds a penalty term to boosted regression ◦ Can be formulated as LASSO/ridge regression/elastic net with constraints ◦ Also leverages hardware to further speed up boosting ensemble play.google.com

28. } Single tree models poor predictors. } Grow a forest of them for a strong predictor (another ensemble method). } Algorithm: ◦ Takes random sample of data. ◦ Builds one tree. ◦ Repeats until forest grown. ◦ Averages across forest to identify strong predictors.

29. CHAPTER 4

30. } Principle Component Analysis (PCA) ◦ Variance partitioning to combine pieces of variables into new linear subspace. ◦ Smashing cookies by kind and combining into a flat hybrid cookie in previous regression model. } Manifold Learning ◦ PCA-like algorithms that combine pieces into a new nonlinear subspace. ◦ Non-flat cookie combining. } Useful as pre-processing step for prediction models. ◦ Reduce dimension. ◦ Obtain uncorrelated, non- overlapping variables (bases). marmaladeandmileposts. com

31. } Balanced sampling for low-frequency predictors. ◦ Stratified samples (i.e. sample from bag of mostly white marbles and few red marbles with constraint that 1/5th of draws must be red marbles). } Dimension reduction/mapping pre-processing ◦ Principle component, manifold learning… ◦ Hybrid of neural network methods and tree models.

32. } Aggregation of multiple types of models. ◦ Like a small town election. ◦ Different people have different views of the politics and care about different issues. } Different modeling methods capture different pieces of the data and vote in different pieces. ◦ Leverage strengths, minimize weaknesses ◦ Diversity of methods to better explore underlying data geometry } Avoids multiple testing issues.

33. } Subsembles ◦ Partition data into training sets □ Randomly selected or through partition algorithm ◦ Train a model on each data partition ◦ Combine into final weighted prediction model } This is similar to national elections. ◦ Each elector in the electoral college learns for whom his constituents voted. ◦ The final electoral college pools these individual votes. Full Training Dataset Model Training Datasets 1, 2, 3, 4 Final Subsemble Model

34. } Combines superlearning and subsembles for better prediction ◦ Diversity improves subsemble method (better able to explore data geometry) ◦ Bootstrapping improves superlearner pieces (more diversity within each method) ◦ Preliminary empirical evidence shows efficacy of combination.

35. CHAPTER 5

36. } Classification of a data point based on the classification of its nearest neighboring points. } Like a junior high lunchroom and student clicks. ◦ Students at the same table tend to be more similar to each other than to a distant table. www.wrestlecrap.com

37. } Iteratively separating groups within a dataset based on similarity. } Like untangling tangled balls of yarn into multiple, single- colored balls (ideally). ◦ Each step untangles the mess a little further. ◦ Few stopping guidelines (when it looks separated). krazymommakreations.blogspot.com

38. } Hybrid of supervised and unsupervised learning (groupings and prediction). } Uses graphical representation of data and its relationships. } Algorithms with connections to topology, differential geometry, and Markov chains. } Useful in combination with other machine learning methods to provide extra insight (ex. spectral clustering).

39. } K-means algorithm with weighting and dimension reduction components of similarity measure. ◦ Simplify balls of string to warm colors and cool colors before untangling. } Can be reformulated as a graph clustering problem. ◦ Partition subcomponents of a graph based on flow equations. www.simplepastimes.com

40. } Multivariate technique similar to mode or density clustering. ◦ Find peaks and valleys in data according to an input function on the data (level set slices)— much like a watershed on mountains. ◦ Separate data based on shared peaks and valleys across slices (shared multivariate density/gradient). ◦ Many nice theoretical developments on validity and convergence.

41. } Topological clustering. ◦ Define distance metric. ◦ Slice multidimensional dataset with Morse function. ◦ Examine function behavior across slice. ◦ Cluster function behavior. ◦ Iterate through multiple slices to obtain hierarchy of function behavior. } Much like examining the behavior of multiple objects across a flip book. ◦ Nesting ◦ Cluster overlap Filtered functions then used to create various resolutions of a modified Reeb graph summary of topology.

42. **Time Series Forecasting**

43. } Similar to decomposing superposed states ◦ Seasonal trends ◦ Yearly trends ◦ Trend averages ◦ Dependencies on previous time point } Knit individual forecasted pieces into a complete forecast by superposing these individual forecasts } Several extensions to neural networks, time- lagged machine learning models…

44. } A time-series method incorporating predictors ◦ Constant predictors at initial time point ◦ Varying predictors at multiple time points } Creates a sort of correlation web between predictors and time points ◦ Can handle multiple time lags and multivariate outcomes ◦ Can handle any GLM outcome links } Related to partial differential equations of dynamic systems

45. } Data-based mining for SEM relationships/time-lag components ◦ Leverages conditional probability between predictors to find dependencies ◦ Does not require a priori model formulation like SEM } Peeking at data to create a dependency web over time or predictors/outcome } Can be validated by a follow-up SEM based on network structure Time 1 Outcome Predictor 2 Time 2 Outcome Time 3 Outcome

46. } Technically: ◦ Matrix decomposition (similar to PCA/manifold learning) ◦ Followed by spectral methods ◦ Cleaning of time-lagged covariance matrix ◦ Reconstruction with simple forecast } Kind of like deconstructing, cleaning, a rebuilding a car engine

47. } Combines k-nearest neighbors-based clustering with memoryless state changes converging to a transition distribution (weighted directed graph) ◦ Reduce an observation to a pattern ◦ Remember patterns seen (across time or space) ◦ Match new observations to this set of patterns ◦ Computationally more feasible than k-means clustering Like remembering classes of chess board configurations across games played

48. } Many machine learning methods exist today, and many more are being developed every day. } Methods come with certain assumptions that must be met. ◦ Breaking assumptions can lead to poor model fit or incorrect inference. ◦ Matching a method to a problem not only can help with better prediction and inference; it can also lead to faster computation times and better presentations to clients. } Development depends on problem-matching and deep understanding of the mathematics behind the methods used.

49. } Parametric Regression: ◦ Draper, N. R., Smith, H., & Pownell, E. (1966). Applied regression analysis (Vol. 3). New York: Wiley. ◦ McCullagh, P. (1984). Generalized linear models. European Journal of Operational Research, 16(3), 285-292. ◦ Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2), 301-320. ◦ Augugliaro, L., Mineo, A. M., & Wit, E. C. (2013). Differential geometric least angle regression: a differential geometric approach to sparse generalized linear models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 75(3), 471-498. ◦ Raftery, A. E., Madigan, D., & Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. Journal of the American Statistical Association, 92(437), 179-191. ◦ Osborne, M. R., & Turlach, B. A. (2011). A homotopy algorithm for the quantile regression lasso and related piecewise linear problems. Journal of Computational and Graphical Statistics, 20(4), 972-987. ◦ Drori, I., & Donoho, D. L. (2006, May). Solution of l 1 minimization problems by LARS/homotopy methods. In Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on (Vol. 3, pp. III-III). IEEE.

50. } Semi-Parametric Regression: ◦ Marsh, L. C., & Cormier, D. R. (2001). Spline regression models (Vol. 137). Sage. ◦ Hastie, T., & Tibshirani, R. (1986). Generalized additive models. Statistical science, 297-310. ◦ McZgee, V. E., & Carleton, W. T. (1970). Piecewise regression. Journal of the American Statistical Association, 65(331), 1109-1124. ◦ Gerber, S., Rübel, O., Bremer, P. T., Pascucci, V., & Whitaker, R. T. (2013). Morse– Smale Regression. Journal of Computational and Graphical Statistics, 22(1), 193- 214. ◦ Hearst, M. A., Dumais, S. T., Osman, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. Intelligent Systems and their Applications, IEEE, 13(4), 18-28. ◦ Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. Neural networks, 2(5), 359-366. ◦ Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105). ◦ Huang, G. B., Wang, D. H., & Lan, Y. (2011). Extreme learning machines: a survey. International Journal of Machine Learning and Cybernetics, 2(2), 107-122. ◦ Friedman, J. H. (1991). Multivariate adaptive regression splines. The annals of statistics, 1-67. ◦ Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., ... & Ghemawat, S. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467.

51. } Nonparametric Regression: ◦ Buntine, W. (1992). Learning classification trees. Statistics and computing, 2(2), 63-73. ◦ Friedman, J. H. (2002). Stochastic gradient boosting. Computational Statistics & Data Analysis, 38(4), 367-378. ◦ Bäck, T. (1996). Evolutionary algorithms in theory and practice: evolution strategies, evolutionary programming, genetic algorithms. Oxford university press. ◦ Zhang, G. (2011). Quantum-inspired evolutionary algorithms: a survey and empirical study. Journal of Heuristics, 17(3), 303-351. ◦ Dietterich, T. G. (2000). Ensemble methods in machine learning. In Multiple classifier systems (pp. 1-15). Springer Berlin Heidelberg. ◦ Friedman, J. H. (2002). Stochastic gradient boosting. Computational Statistics & Data Analysis, 38(4), 367-378. ◦ Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32. ◦ Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794). ACM.

52. } Supervised with Unsupervised Methods: ◦ van der Maaten, L. J., Postma, E. O., & van den Herik, H. J. (2009). Dimensionality reduction: A comparative review. Journal of Machine Learning Research, 10(1-41), 66-71. ◦ Kuncheva, L. I., & Rodríguez, J. J. (2007). An experimental study on rotation forest ensembles. In Multiple Classifier Systems (pp. 459-468). Springer Berlin Heidelberg. ◦ van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super learner. Statistical applications in genetics and molecular biology, 6(1). ◦ Sapp, S. K. (2014). Subsemble: A Flexible Subset Ensemble Prediction Method (Doctoral dissertation, University of California, Berkeley).

53. } Unsupervised Methods: ◦ Fukunaga, K., & Narendra, P. M. (1975). A branch and bound algorithm for computing k-nearest neighbors. Computers, IEEE Transactions on, 100(7), 750-753. ◦ MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, No. 14, pp. 281-297). ◦ Chartrand, G., & Oellermann, O. R. (1993). Applied and algorithmic graph theory. ◦ Ng, A. Y., Jordan, M. I., & Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. Advances in neural information processing systems, 2, 849-856. ◦ Singh, G., Mémoli, F., & Carlsson, G. E. (2007, September). Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition. In SPBG (pp. 91- 100). ◦ Chen, Y. C., Genovese, C. R., & Wasserman, L. (2016). A comprehensive approach to mode clustering. Electronic Journal of Statistics, 10(1), 210-241.

54. } Time Series ◦ Wu, J. P., & Wei, S. (1989). Time series analysis. Hunan Science and Technology Press, ChangSha. ◦ Vautard, R., Yiou, P., & Ghil, M. (1992). Singular-spectrum analysis: A toolkit for short, noisy chaotic signals. Physica D: Nonlinear Phenomena, 58(1), 95-126. ◦ Dunham, M. H., Meng, Y., & Huang, J. (2004, November). Extensible markov model. In Data Mining, 2004. ICDM'04. Fourth IEEE International Conference on (pp. 371-374). IEEE. ◦ Ullman, J. B., & Bentler, P. M. (2003). Structural equation modeling. John Wiley & Sons, Inc.. ◦ Heckerman, D., Geiger, D., & Chickering, D. M. (1994, July). Learning Bayesian networks: The combination of knowledge and statistical data. In Proceedings of the Tenth international conference on Uncertainty in artificial intelligence (pp. 293-301). Morgan Kaufmann Publishers Inc..

References

Parametric Regression

- Draper, N. R., Smith, H., & Pownell, E. (1966). *Applied regression analysis* (Vol. 3). New York: Wiley.
- McCullagh, P. (1984). Generalized linear models. *European Journal of Operational Research*, *16*(3), 285-292.
- Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301-320.
- Augugliaro, L., Mineo, A. M., & Wit, E. C. (2013). Differential geometric least angle regression: a differential geometric approach to sparse generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *75*(3), 471-498.
- Raftery, A. E., Madigan, D., & Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, *92*(437), 179-191.
- Osborne, M. R., & Turlach, B. A. (2011). A homotopy algorithm for the quantile regression lasso and related piecewise linear problems. *Journal of Computational and Graphical Statistics*, *20*(4), 972-987.
- Drori, I., & Donoho, D. L. (2006, May). Solution of l 1 minimization problems by LARS/homotopy methods. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on* (Vol. 3, pp. III-III). IEEE.

References

Semi-Parametric Regression

- Marsh, L. C., & Cormier, D. R. (2001). *Spline regression models* (Vol. 137). Sage.
- Hastie, T., & Tibshirani, R. (1986). Generalized additive models. *Statistical science*, 297-310.
- McZgee, V. E., & Carleton, W. T. (1970). Piecewise regression. *Journal of the American Statistical Association*, *65*(331), 1109-1124.
- Gerber, S., Rübel, O., Bremer, P. T., Pascucci, V., & Whitaker, R. T. (2013). Morse-Smale Regression. *Journal of Computational and Graphical Statistics*, *22*(1), 193-214.
- Hearst, M. A., Dumais, S. T., Osman, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *Intelligent Systems and their Applications, IEEE*, *13*(4), 18-28.
- Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, *2*(5), 359-366.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).
- Huang, G. B., Wang, D. H., & Lan, Y. (2011). Extreme learning machines: a survey. *International Journal of Machine Learning and Cybernetics*, *2*(2), 107-122.
- Friedman, J. H. (1991). Multivariate adaptive regression splines. *The annals of statistics*, 1-67.
- Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., ... & Ghemawat, S. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.

References

Nonparametric Regression

- Buntine, W. (1992). Learning classification trees. *Statistics and computing*, *2*(2), 63-73.
- Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, *38*(4), 367-378.
- Bäck, T. (1996). *Evolutionary algorithms in theory and practice: evolution strategies, evolutionary programming, genetic algorithms*. Oxford university press.
- Zhang, G. (2011). Quantum-inspired evolutionary algorithms: a survey and empirical study. *Journal of Heuristics*, *17*(3), 303-351.
- Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Multiple classifier systems* (pp. 1-15). Springer Berlin Heidelberg.
- Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, *38*(4), 367-378.
- Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.
- Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794). ACM.

References

Supervised with Unsupervised Methods

- van der Maaten, L. J., Postma, E. O., & van den Herik, H. J. (2009). Dimensionality reduction: A comparative review. *Journal of Machine Learning Research*, *10*(1-41), 66-71.
- Kuncheva, L. I., & Rodríguez, J. J. (2007). An experimental study on rotation forest ensembles. In *Multiple Classifier Systems* (pp. 459-468). Springer Berlin Heidelberg.
- van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super learner. *Statistical applications in genetics and molecular biology*, *6*(1).
- Sapp, S. K. (2014). *Subsemble: A Flexible Subset Ensemble Prediction Method* (Doctoral dissertation, University of California, Berkeley).

References

Unsupervised Methods

- Fukunaga, K., & Narendra, P. M. (1975). A branch and bound algorithm for computing k-nearest neighbors. *Computers, IEEE Transactions on*, *100*(7), 750-753.
- MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, No. 14, pp. 281-297).
- Chartrand, G., & Oellermann, O. R. (1993). Applied and algorithmic graph theory.
- Ng, A. Y., Jordan, M. I., & Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, *2*, 849-856.
- Singh, G., Mémoli, F., & Carlsson, G. E. (2007, September). Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition. In *SPBG* (pp. 91- 100).
- Chen, Y. C., Genovese, C. R., & Wasserman, L. (2016). A comprehensive approach to mode clustering. *Electronic Journal of Statistics*, *10*(1), 210-241.

References

Time Series

- Wu, J. P., & Wei, S. (1989). *Time series analysis*. Hunan Science and Technology Press, ChangSha.
- Vautard, R., Yiou, P., & Ghil, M. (1992). Singular-spectrum analysis: A toolkit for short, noisy chaotic signals. *Physica D: Nonlinear Phenomena*, *58*(1), 95-126.
- Dunham, M. H., Meng, Y., & Huang, J. (2004, November). Extensible markov model. In *Data Mining, 2004. ICDM'04. Fourth IEEE International Conference on* (pp. 371-374). IEEE.
- Ullman, J. B., & Bentler, P. M. (2003). *Structural equation modeling*. John Wiley & Sons, Inc..
- Heckerman, D., Geiger, D., & Chickering, D. M. (1994, July). Learning Bayesian networks: The combination of knowledge and statistical data. In *Proceedings of the Tenth international conference on Uncertainty in artificial intelligence* (pp. 293-301). Morgan Kaufmann Publishers Inc..