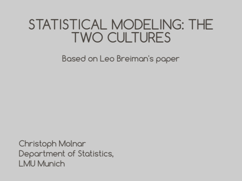
STATISTICAL MODELING: THE TWO CULTURES

Based on Leo Breiman's paper

Christoph Molnar
Department of Statistics,
LMU Munich



STATISTICAL MODELING: THE TWO CULTURES

Based on Leo Breiman's paper

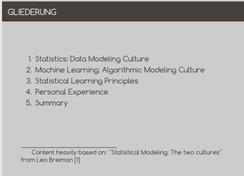
Christoph Molnar
Department of Statistics,
LMU Munich

Abstract:
This presentation compares two cultures of statistical modeling: the data modeling culture, which assumes a stochastical process that produced the data. This culture is associated with traditional statistics. The other culture is called algorithmic modeling culture, which can be reduced to optimisation of a loss function with an algorithm. This culture is associated with Machine Learning. It is argued to use algorithmic modeling more often in statistics.

GLIEDERUNG

1. Statistics: Data Modeling Culture
2. Machine Learning: Algorithmic Modeling Culture
3. Statistical Learning Principles
4. Personal Experience
5. Summary

Content heavily based on: ``Statistical Modeling: The two cultures" from Leo Breiman [1]



GLIEDERUNG

1. Statistics: Data Modeling Culture
2. Machine Learning: Algorithmic Modeling Culture
3. Statistical Learning Principles
4. Personal Experience
5. Summary

Content heavily based on: ``Statistical Modeling: The two cultures" from Leo Breiman [1]

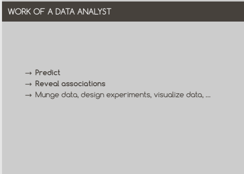
This presentation is based on ``Statistical Modeling: The two cultures" from Leo Breiman [1].

The first segment introduces the data modeling culture and analogously the second segment explains the algorithmic modeling culture together with the presentation of three algorithms. The part about statistical learning principals presents aspects which help to compare both of cultures. Personal experiences in both cultures are addressed. The conclusion summarizes the message of the paper [1].

DATA MODELING CULTURE

WORK OF A DATA ANALYST

- → **Predict**
- → **Reveal associations**
- → Munge data, design experiments, visualize data, ...

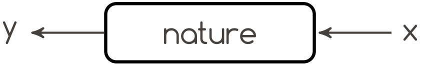


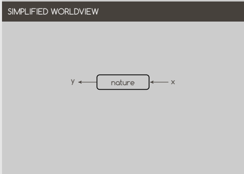
WORK OF A DATA ANALYST

- → **Predict**
- → **Reveal associations**
- → Munge data, design experiments, visualize data, ...

The work of a data analyst is very diverse. This presentation focuses on the modeling of data, which can be reduced to two targets: Learn a model to predict the outcome for new covariates and get a better understanding about the relationship between covariates and outcome.

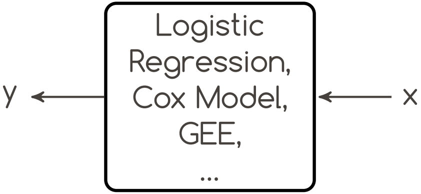
SIMPLIFIED WORLDVIEW




SIMPLIFIED WORLDVIEW

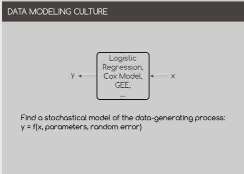
In a strongly simplified world an arbitrary outcome y is produced by the nature given the covariates x. The knowledge about the natures true mechanisms range between entirely unkwown and established (scientific) explanations of the mechanism. One example: Outcome y is the rent for appartments and covariates x are size, number of bathrooms and location.

DATA MODELING CULTURE



y ← Logistic Regression, Cox Model, GEE, ... ← x

Find a stochastical model of the data-generating process:
y = f(x, parameters, random error)



DATA MODELING CULTURE

Find a stochastical model of the data-generating process:
$y = f(x, \text{parameters}, \text{random error})$

The direct modeling of the mechanism in the ``box" is labeled »Data Modeling Culture« by Leo Breiman. In this culture a stochastical model for the data- generating process is assumed. A common formulation of these model is: y is a function of x with corresponding weights and a random error. For example: Given the covariates size, number of bathrooms and location, the rent of appartments is normal distributed.

TYPICAL ASSUMPTIONS AND RESTRICTIONS

→ Specific stochastical model that generated the data

→ Distribution of residuals

→ Linearities (e.g. linear predictor)

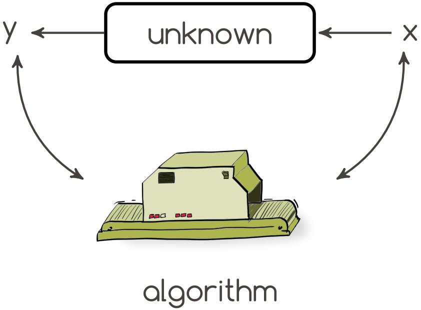
→ Manual specification of interactions

PROBLEMS

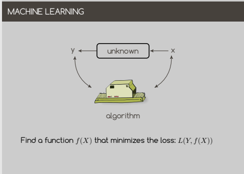
- → Conclusions about model, not about nature
- → Assumptions often violated
- → Often no model evaluation
- → ⇒ can lead to irrelevant theory and questionable statistical conclusions
- → Focus not on prediction
- → Data models fail in areas like image and speech recognition

ALGORITHMIC MODELING CULTURE

MACHINE LEARNING



Find a function $f(X)$ that minimizes the loss: $L(Y, f(X))$



MACHINE LEARNING

Find a function $f(X)$ that minimizes the loss: $L(Y, f(X))$

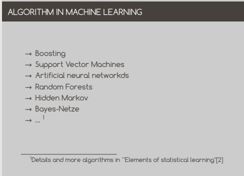
In the »Algorithmic Modeling Culture«, the true mechanism is treated as unkown. It is not the target to find the true data-generating mechanism but to use an algorithm that imitates the mechanism as good as possible. Modeling is reduced to a mere problem of function optimization: Given the covariates x, outcome y and a loss function find a function f(x) which minimizes the loss for the prediction of the outcome. This culture is lived in the machine learning area.

Summary: The data modeling culture tries to find the true data-generating mechanism, the algorithmic modeling culture tries to imitate the true mechanism as good as possible.

ALGORITHM IN MACHINE LEARNING

→ Boosting
→ Support Vector Machines
→ Artificial neural networkds
→ Random Forests
→ Hidden Markov
→ Bayes-Netze
→ ...[1]

[1] Details and more algorithms in ``Elements of statistical learning"[2]



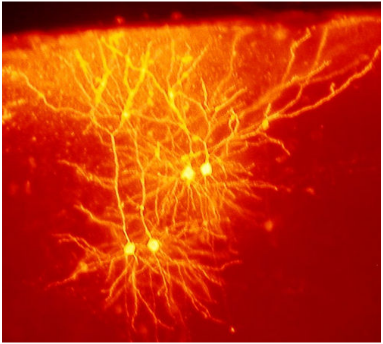
ALGORITHM IN MACHINE LEARNING

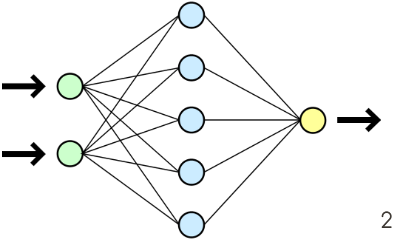
- → Boosting
- → Support Vector Machines
- → Artificial neural networkds
- → Random Forests
- → Hidden Markov
- → Bayes-Netze
- → ...[1]

[1] Details and more algorithms in ``Elements of statistical learning"[2]

The algorithms used in machine learning are motivated differently. Three algorithms are presented in short.

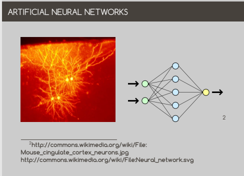
ARTIFICIAL NEURAL NETWORKS





[2]

[2] http://commons.wikimedia.org/wiki/File:Mouse_cingulate_cortex_neurons.jpg
http://commons.wikimedia.org/wiki/File:Neural_network.svg



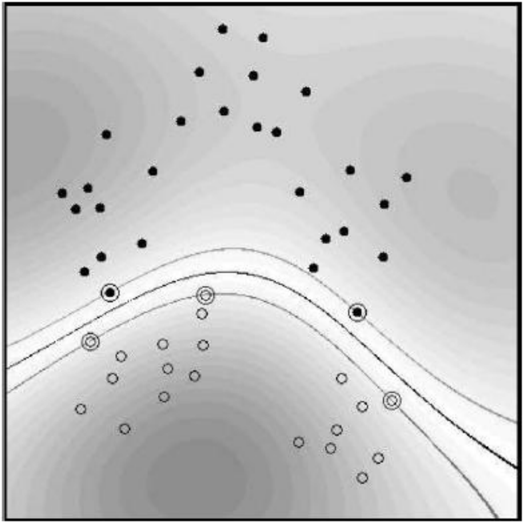
ARTIFICIAL NEURAL NETWORKS

[2]

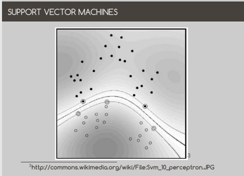
[2] http://commons.wikimedia.org/wiki/File: Mouse_cingulate_cortex_neurons.jpg
http://commons.wikimedia.org/wiki/File:Neural_network.svg

Artificial neural networks are used in classification and regression. They are inspired by the brain, which consists of a network of brain cells (neurons). Mathematically artifical neural networks are a concatenation of weighted functions. An exemplary application is image processing.

SUPPORT VECTOR MACHINES


3

[3] http://commons.wikimedia.org/wiki/File:Svm_10_perceptron.JPG



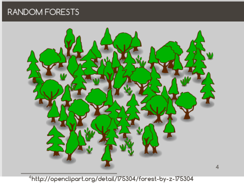
SUPPORT VECTOR MACHINES

[3] http://commons.wikimedia.org/wiki/File:Svm_10_perceptron.JPG

Support Vector Machines (SVM) were originally a classification method (regression is also possible). SVMs try to draw a border between to classes in the covariate space. The distance between the border and the class points is maximized. They use a mathematical trick to implicitly project the covariates in a space with higher dimensions (yes, it sounds a bit crazy) in order to achieve class separation. Text classification is an exemplary usage.

RANDOM FORESTS

4



RANDOM FORESTS

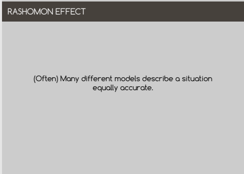
[4] http://openclipart.org/detail/175304/forest-by-z-175304

Random Forests™(invented by Leo Breiman) are used for regression und classification. A Random Forest consists of many decision trees. Two random mechanisms are used to train different trees on the data. Results from all trees are averaged for the prediction.

STATISTICAL LEARNING PRINCIPLES

RASHOMON EFFECT

(Often) Many different models describe a situation equally accurate.



RASHOMON EFFECT

(Often) Many different models describe a situation equally accurate.

Rashomon is a japanese movie in which four witnessess tell different versions of a crime. All versions account for the facts but they contradict each other. In terms of statistical learning this means, that often different models (e.g. same y but different covariates) can be equally accurate. Each model has a different interpretation which makes it difficult to find the true mechanism in the data modeling cultures. In the algorithmic modeling culture the Rashomon effect is exploited by aggregating over many models. Random forests use this effect by aggregating over many trees. It is also common to average the predictions of different algorithms.

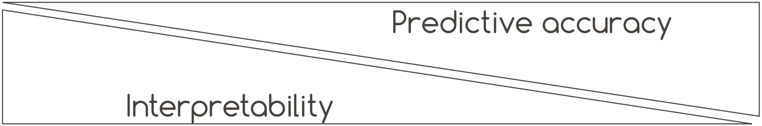
DIMENSIONALITY OF THE DATA

→ The higher the dimensionality of the data (# covariates) the more difficult is the separation of signal and noise

→ Common practice in data modeling: variable selection (by expert selection or data driven) and reduction of dimensionality (e.g. PCA)

→ Common practice in algorithmic modeling: Engineering of new features (covariates) to increase predictive accuracy; algorithms robust for many covariates

PREDICTION VS. INTERPRETATION



Predictive accuracy

Interpretability

Tree

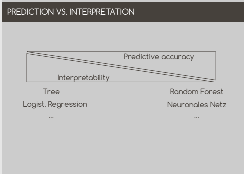
Logist. Regression

...

Random Forest

Neuronales Netz

...



PREDICTION VS. INTERPRETATION

Predictive accuracy

Interpretability

Tree
Logist. Regression
...

Random Forest
Neuronales Netz
...

There is a tradeoff between interpretability and predictive accuracy: the models that are good in prediction are often complex and models that are easy to interpret are often bad predictors. See for example trees and Random Forests: A single decision tree is very intuitive and easy to read for non-professionals, but they are instable and give weak predictions. A complex aggregation of decision trees (Random Forest) has an excellent prediction accuracy, but it is impossible to interpret the model structure.

GOODNESS OF MODEL

→ Data modeling culture: Goodness of fit often based on model assumptions (e.g. AIC) and calculated on training data.

→ Algorithmic modeling culture: Evaluation of predictive accuracy with an extra test set or cross validation.

How good is a statistical model if the predictive accuracy is weak? Is it legit to interpretate parameters and p-values?

PERSONAL EXPERIENCES

STATISTICAL CONSULTING

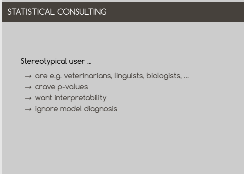
Stereotypical user ...

→ are e.g. veterinarians, linguists, biologists, ...

→ crave p-values

→ want interpretability

→ ignore model diagnosis



STATISTICAL CONSULTING

Stereotypical user ...

→ are e.g. veterinarians, linguists, biologists, ...
→ crave p-values
→ want interpretability
→ ignore model diagnosis

From my experience in the statistical consulting unit of our university, most user want to test their scientific hypthesis with models. They want models which are easy to interpret regarding their questions. Thus it is more important to have a model that gives parameters associated with covariates and p-values than to have a model that predicts the data well.

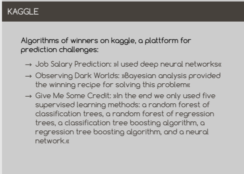
KAGGLE

Algorithms of winners on kaggle, a plattform for prediction challenges:

→ Job Salary Prediction: »I used deep neural networks«

→ Observing Dark Worlds: »Bayesian analysis provided the winning recipe for solving this problem«

→ Give Me Some Credit: »In the end we only used five supervised learning methods: a random forest of classification trees, a random forest of regression trees, a classification tree boosting algorithm, a regression tree boosting algorithm, and a neural network.«



KAGGLE

Algorithms of winners on kaggle, a platform for prediction challenges:

→ Job Salary Prediction: »I used deep neural networks«
→ Observing Dark Worlds: »Bayesian analysis provided the winning recipe for solving this problem«
→ Give Me Some Credit: »In the end we only used five supervised learning methods: a random forest of classification trees, a random forest of regression trees, a classification tree boosting algorithm, a regression tree boosting algorithm, and a neural network.«

Kaggle (**http://www.kaggle.com**) is a platform for prediction challenges. Data are provided and the participant with the best prediction on a test set wins the challenge. To generate insights about the mechanisms in the data is secondary because the prediction is all that counts to win. That's why the algorithmic modeling culture is superior in this field.

CONCLUSION

Data analysts should:

→ Use predictive accuracy to evaluate models

→ Seek the best model

→ Add Machine Learning to their toolbox

FURTHER LITERATURE

L. Breiman
Statistical modeling: The two cultures (with comments and a rejoinder by the author)
Institute of Mathematical Statistics, 2001.

T. Hastie, R. Tibshirani and J. Friedman
The elements of statistical learning
Springer New York, 2001

»All models are wrong, but some are useful« (G. Box)