universität wien

OSG

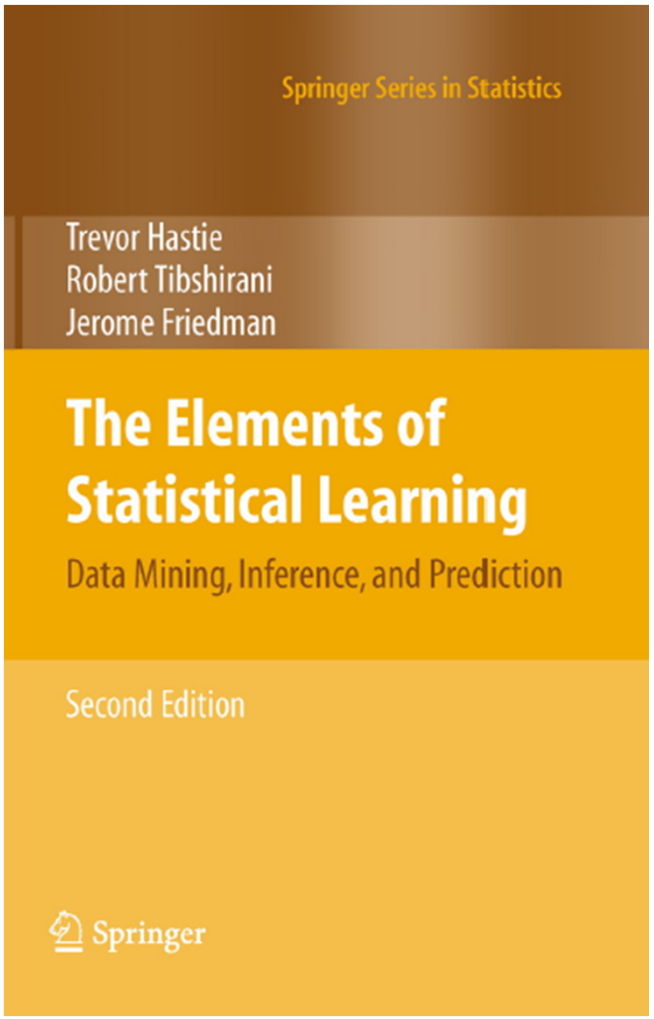
Anwendung von Ensemble Methoden für Klassifikationsaufgaben

Marcus Hudec
marcus.hudec@univie.ac.at

Österreichische Statistiktage 2011
Graz, 7.- 9. September 2011

Vorbemerkungen

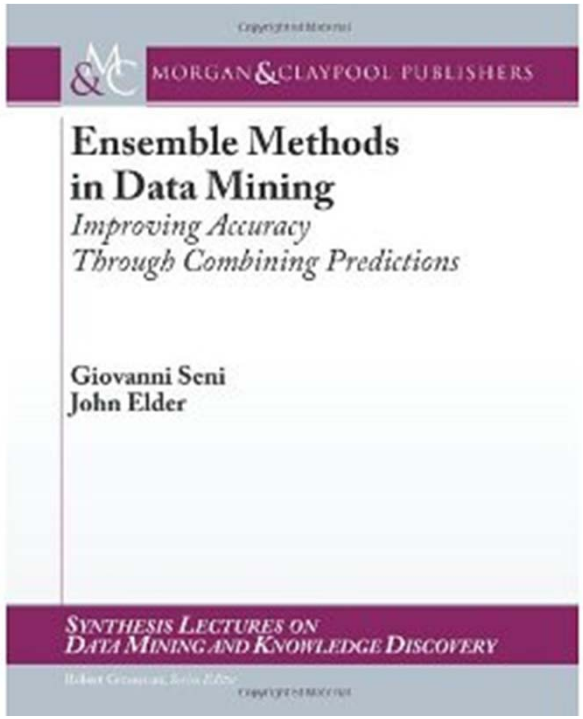
- Ensemble Methoden sind sicherlich eine der interessantesten Entwicklungen im Bereich der Angewandten Statistik der letzten 10 Jahre



Chapter	What's new
1. Introduction	
2. Overview of Supervised Learning	
3. Linear Methods for Regression	LAR algorithm and generalizations of the lasso
4. Linear Methods for Classification	Lasso path for logistic regression
5. Basis Expansions and Regularization	Additional illustrations of RKHS
6. Kernel Smoothing Methods	
7. Model Assessment and Selection	Strengths and pitfalls of cross-validation
8. Model Inference and Averaging	
9. Additive Models, Trees, and Related Methods	
10. Boosting and Additive Trees	New example from ecology; some material split off to Chapter 16.
11. Neural Networks	Bayesian neural nets and the NIPS 2003 challenge
12. Support Vector Machines and Flexible Discriminants	Path algorithm for SVM classifier
13. Prototype Methods and Nearest-Neighbors	
14. Unsupervised Learning	Spectral clustering, kernel PCA, sparse PCA, non-negative matrix factorization archetypal analysis, nonlinear dimension reduction, Google page rank algorithm, a direct approach to ICA
15. Random Forests	New
16. Ensemble Learning	New
17. Undirected Graphical Models	New
18. High-Dimensional Problems	New

Vorbemerkungen

- Ensemble Methoden sind eine der interessantesten Entwicklungen im Bereich der Angewandten Statistik der letzten 10 Jahre
- Ensemble Methoden sind von hoher Relevanz in der industriellen Anwendung im Kontext des Predictive Analytics (Data Mining)



- Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. Expert Systems with Applications, 36(3, Part 1), 4626-4636.
- Weiyun, Y., Xiu, L., Yaya, X., & Johnson, E. (2008, 13-15 July 2008). Preventing customer churn by using random forests modeling. Paper presented at the IEEE International Conference on Information Reuse and Integration, 2008.

Vorbemerkungen

- Ensemble Methoden sind eine der interessantesten Entwicklungen im Bereich der Angewandten Statistik der letzten 10 Jahre
- Ensemble Methoden sind von hoher Relevanz in der industriellen Anwendung im Kontext des Predictive Analytics (Data Mining)
- Ensemble Methoden kombinieren multiple Modelle zu einem komplexen Gesamtmodell, das eine höhere Prognosegüte (Trennschärfe) aufweist als die einzelnen Komponenten
- Ensemble Methoden sind für solche Fragestellungen vielversprechend, wo Prognosegüte (Trennschärfe) wichtiger ist als eine einfache Modellinterpretation

Ein neuer Modellierungsansatz

- Ensemble Methoden erweitern den Werkzeugkoffer der Angewandten Statistik um eine zusätzliche Facette
- Ensemble Methoden repräsentieren eine neue innovative Herangehensweise an Fragestellungen der Angewandten Statistik
- In diesem Sinne stellen sie ein neues Paradigma dar

Strategien der Angewandten Statistik

- **Wahl der Modellierungs-Methodik**
- Teste, ob die zu Grunde liegenden Annahmen der gewählten Methode eine gute Koinzidenz mit den Eigenschaften des konkreten Datensatzes aufweisen (Diagnostische Tools; Transformationen; Wechsel der Schätzmethodik (Robuste Methoden; Regularization; Shrinkage))
- **Modell Selektion**
- Man passt mehrere Modelle an die Daten an und wählt jenes aus, das sich in Bezug auf eine Zielfunktion optimal verhält (Wahl der Prädiktoren; Interaktionseffekte; Modellsegmentierung)

Strategie bei Ensemble Methoden

- Für Aufgaben des Predictive Modellings existiert heute ein reichhaltiges Methodenspektrum
- Die Qualität dieser Methoden hängt in hohem Maße von den spezifischen Details einer Problemstellung ab. Eine allgemeine qualitative Reihung in Bezug auf Trennschärfe oder Prognosegenauigkeit ist praktisch nicht möglich.
- Für den Praktiker stellt sich die Frage: welche Methode soll ich zur Lösung meines konkreten Businessproblems anwenden?
- Idee: Verwende mehrere Methoden/Modelle und kombiniere die Vorhersagen zu einer Gesamtschätzung

Klassifikationsaufgaben

- Wir betrachten Stichpoben aus g verschiedenen Teilpopulationen
- Ausgangspunkt bildet ein klassifizierter Trainings-Datensatz
- Ziel von Klassifikationsaufgaben (supervised learning) ist es Entscheidungsregeln zu finden, die es ermöglichen die Gruppenzugehörigekit noch nicht klassifizierte Objekte möglichst exakt vorherzusagen

Ausgangsdaten

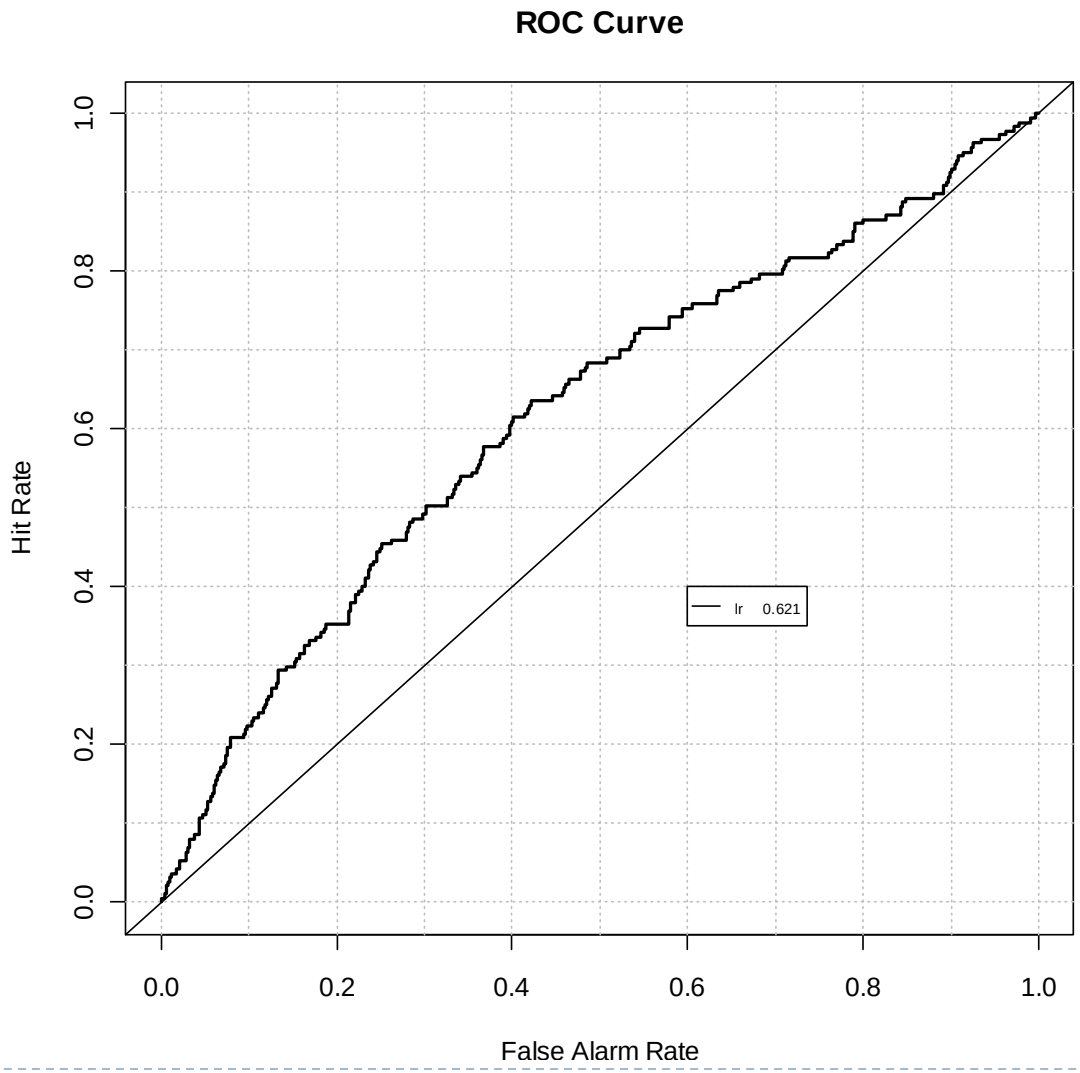
Trainings-Datensatz mit bekannter Gruppenzugehörigkeit

Object	Group	X1	X2	...	Xp
1	1				
...	...				
N_1	1				
1	2				
...	...				
N_2	2				
...	...				
1	g				
...	...				
N_g	g				

Anwendungsbeispiel

- Bilanzrating im Kontext der Kreditrisikoanalyse
- g=2 binäre Zielvariable 0/1 Non-Default/Default
- Realer Datensatz:
  Teil des Portfolios eines österreichischen Kreditunternehmens
- N=3.927 Beobachtungen davon 358 Defaults
- 12 Bilanzindikatoren
  (X1, …, X12) z.B.: Umsatzrentabilität
- Vorgabe seitens Finanzaufsicht:
  Logistische Regression als „state of the art“-Methode

Diskriminationsgüte
Logistische Regression (Out of Sample)



Alternative Methoden

Out-of sample test

	AUC	Rang
Logistische Regression	62,1%	4
Lineare Diskriminanzanalyse	62,2%	3
Quadratische Diskriminanzanalyse	57,0%	7
Regularisierte Diskriminanzanalyse	63,4%	2
Support Vector Machine (linear)	47,0%	10
Support Vector Machine (gaussian)	58,9%	6
Naive Bayes (parametric)	60,6%	5
Naive Bayes (non-parametric)	63,6%	1
Nearest Neighbor	56,3%	9
Recursive Partitioning	56,9%	8

Naive Ensemble Methode (Methoden-Mix)

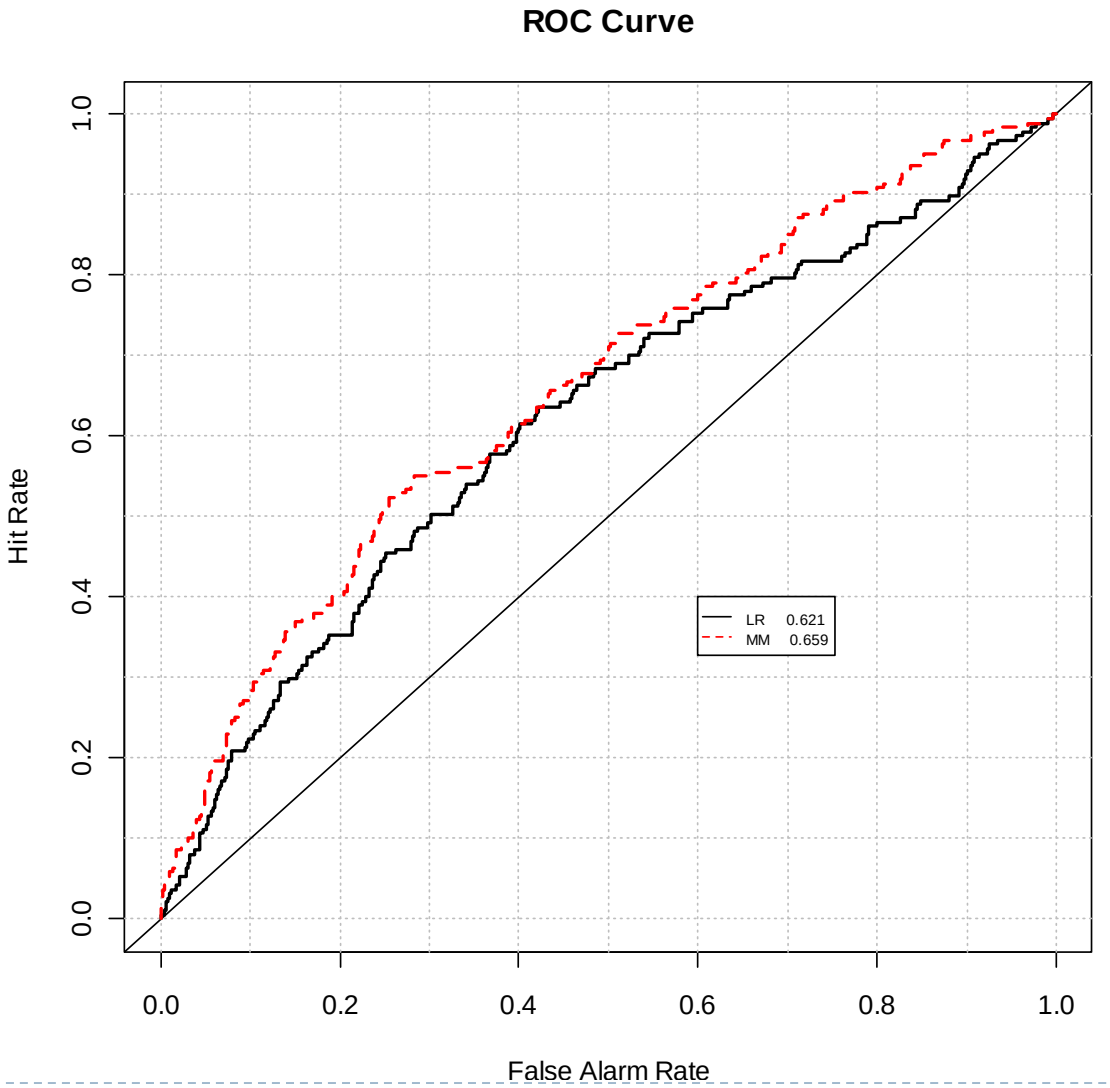
- Kombination der verschiedenen alternativen Schätzer aus den verschiedenen Methoden
- Übernehmen der mehrheitlichen Zuordnung (Majority Voting)
- Gewichtete Schätzung der posteriore Wahrscheinlichkeiten aus den verschiedenen Methoden

Methoden-Mix

Out-of sample test

	AUC	Rang
Logistische Regression	62,1%	5
Lineare Diskriminanzanalyse	62,2%	4
Quadratische Diskriminanzanalyse	57,0%	8
Regularisierte Diskriminanzanalyse	63,4%	3
Support Vector Machine (linear)	47,0%	11
Support Vector Machine (gaussian)	58,9%	7
Naive Bayes (parametric)	60,6%	6
Naive Bayes (non-parametric)	63,6%	2
Nearest Neighbor	56,3%	10
Recursive Partitioning	56,9%	9
Methoden-Mix	65,9%	1

Verbesserung durch den Methoden-Mix



Methoden-Mix

- Idee: durch die Kombination unterschiedlicher Methoden können die Defizite einzelner Methoden kompensiert werden
- Nachteil: es existiert keine Theorie hinter der Vorgangsweise
- Schätzung der Gewichte oft unter Multikollinearität
- Welche Methoden sollen berücksichtigt werden
- Summary: datenanalytische Vorgehensweise, die sich in der Praxis häufig bewährt aber theoretisch nicht fundiert werden kann

Theoriegeleitete Ansätze

- Bagging (Breiman 1996): Fit many large trees to bootstrap-resampled versions of the training data, and classify by majority voting
- Random Forest (Breiman 1999): Improvement of the bagging principle
- Boosting (Freund & Shapire, 1996): Fit many small trees to reweighted versions of the training data, and classify by weighted majority vote
- Modern Gradient Boosting (Friedman, 2001): Fit a sequence of small models to the residual of the previous model

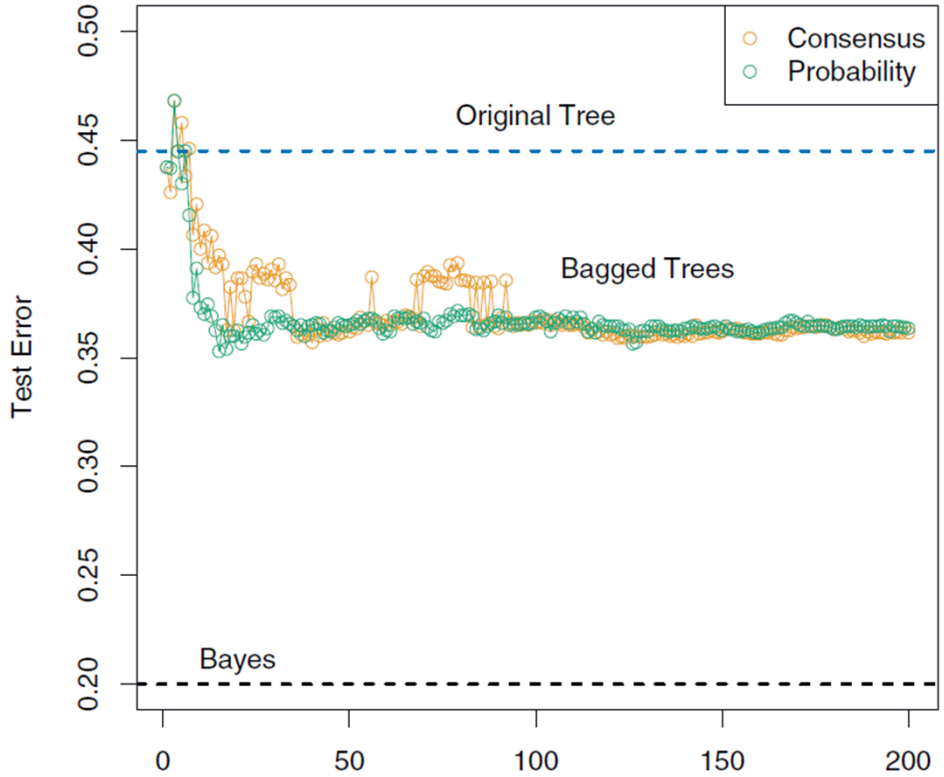
Bagging

- Anwendung des Bootstrap-Prinzips zur Verbesserung der Präzision der Schätzer
- Bootstrap Aggregation mittelt Vorhersagen, die aus einer Kollektion von Bootstrap-Samples gewonnen wird
- Durch die Mittelung kommt es zu einer Reduktion der Varianz der Schätzer, falls diese nichtlinear sind

$$\hat{f}_{\mathrm{bag}}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{*b}(x)$$

Bagging von Trees

- Bagging can dramatically reduce the variance of unstable procedures like trees, leading to improved prediction. (Hastie, Tibsharani & Friedman)



Random Forests

- Analog wie beim Bagging basiert die RF-Methode auf Bootstrap-Samples
- Für jede Bootstrap-Stichprobe wird ein eigener Klassifikationsbaum generiert, wobei bei der Bildung jedes Knoten immer nur eine Teilmenge vom Umfang m aus den M zur Verfügung stehenden Inputvariablen verwendet wird (m<<M)
  - Random Input Selection
  - Random Linear Combination

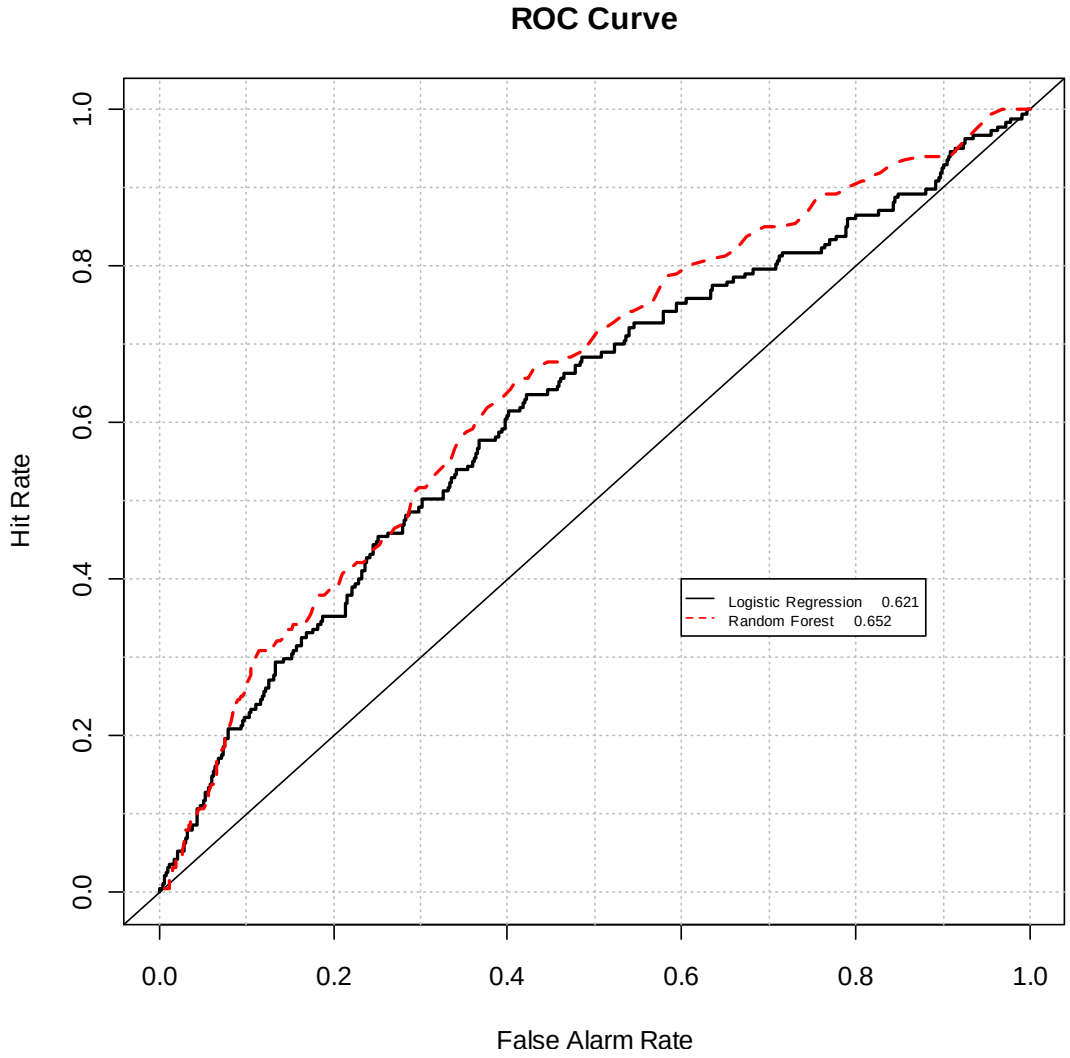
m Tuning-Parameter für Random Forests

- Die Präzision der resultierenden Schätzer hängt von zwei Aspekten ab:
- **Korrelation zwischen den Bäumen des RF**
  Mit wachsender Korrelation steigt die Fehlklassifikationsrate
- **Klassifikationsstärke der Bäume des RF**
  Je höher die Diskriminationsgüte der einzelnen Bäume desto geringer die Fehlklassifikationsrate
- Jede Veränderung von m hat einen direkten Einfluss auf die Korrelation
- Optimale Wahl mittels OOB-error Rate

Eigenschaften von Random Forests

- Breiman:
  “Best of the shelf-procedure for data mining”
- Random Forests laufen effizient auch über große Datensätze
- Random Forests können mit einer großen Zahl von Prädiktoren umgehen, ohne vorher einen Variablenselektionsprozess durchführen zu müssen
- Man erhält automatisch quantitative Indikatoren über die relative Bedeutung der einzelnen Variablen
- Random Forests liefern aus den OOB-Daten einen unverzerrten Schätzer für die Fehlerrate

Verbesserung mit Random Forest



Anwendungsbeispiel

Out-of sample test

	AUC	Rang
Logistische Regression	62,1%	6
Lineare Diskriminanzanalyse	62,2%	5
Quadratische Diskriminanzanalyse	57,0%	9
Regularisierte Diskriminanzanalyse	63,4%	4
Support Vector Machine (linear)	47,0%	12
Support Vector Machine (gaussian)	58,9%	8
Naive Bayes (parametric)	60,6%	7
Naive Bayes (non-parametric)	63,6%	3
Nearest Neighbor	56,3%	11
Recursive Partitioning	56,9%	10
Methoden-Mix	65,9%	1
Random Forest	65,2%	2

Boosting

- Boosting extrem mächtiges Konzept
- Ausgangspunkt: AdaBoost (Freund und Shapire1997)
- Boosting verwendet keine Bootstrap-Samples sondern basiert auf iterativen Modifikationen des Trainingsdatensatzes (perturbation sampling)
- Motivation: Boosting ist ein Algorithmus der die Ergebnisse von vielen “weak classifier” zu einem starken “committee classifier” kombiniert

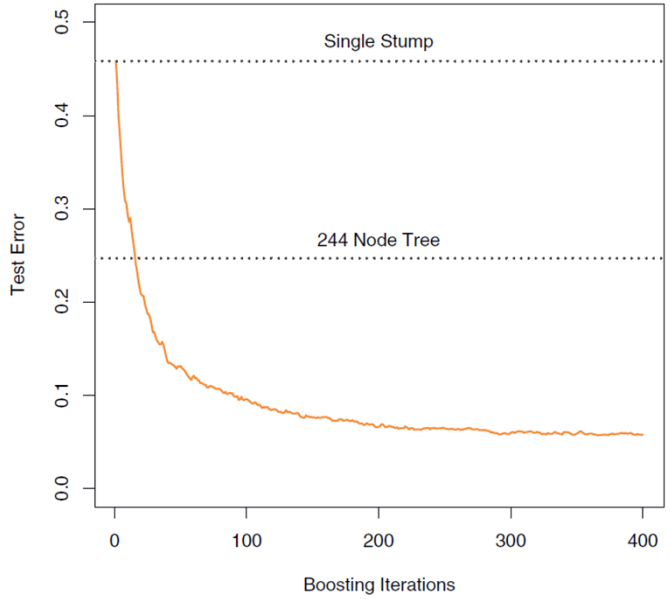
Boosting Algorithmus

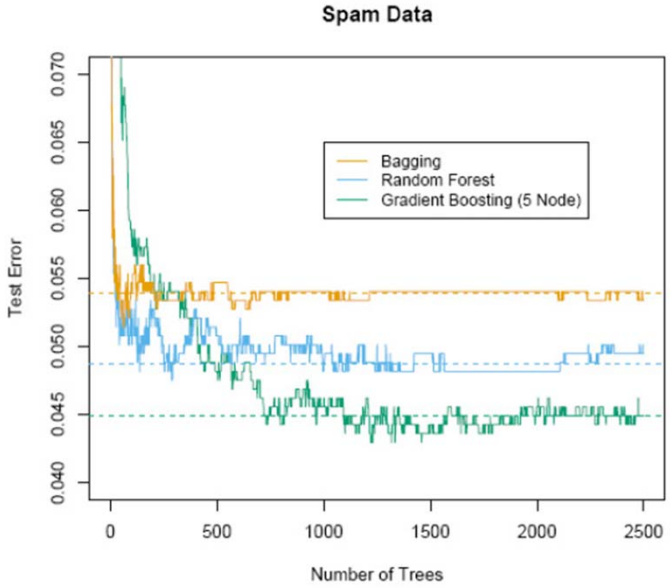
- Ein „weak classification algorithm” wird immer wieder auf den iterativ modifizierten Traininsgdatensatz angewandt, wodurch eine Sequenz von “weak classifiers” generiert wird.
- Sei die Gesamtlänge dieser Sequenz M und bezeichnen wir die “weak classifier” G_m so ergibt sich der finale Classifier nach dem Prinzip des Mehrheitsvotums:

$$G(x) = \text{sign}\left(\sum_{m=1}^{M} \alpha_m G_m(x)\right)$$

Boosting Algorithmus

- Die Modifikation des Trainingsdatensatzes wird durch eine Neugewichtung der Datensätze realisiert
- Diese Neugewichtung erfolgt dabei derart, dass fehlerhaft klassifizierte Daten ein höheres Gewicht in der Trainingsstichprobe erhalten





Verallgemeinerung

- Bei AdaBoost basiert die Bestimmung der modifizierten Fallgewichte auf einer exponentiellen Verlustfunktion, was sowohl die algorithmische Komplexität reduziert als auch eine theoretische Fundierung hat
- Eine Verallgemeinerung auf beliebige (differenzierbare) Verlustfunktionen geht auf Friedman (2001, 2002) zurück und basiert auf Methoden der numerischen Optimierung: **Gradient Boosting**

Anwendungsbeispiel

Out-of sample test

	AUC	Rang
Logistische Regression	62,1%	8
Lineare Diskriminanzanalyse	62,2%	7
Quadratische Diskriminanzanalyse	57,0%	11
Regularisierte Diskriminanzanalyse	63,4%	6
Support Vector Machine (linear)	47,0%	14
Support Vector Machine (gaussian)	58,9%	10
Naive Bayes (parametric)	60,6%	9
Naive Bayes (non-parametric)	63,6%	5
Nearest Neighbor	56,3%	13
Recursive Partitioning	56,9%	12
Methoden-Mix	65,9%	1
Random Forest	65,2%	2
ADA-Boost	64,1%	4
Gradienten-Boosting	64,2%	3

Theoriebasierte Ensemble-Modellierung

- Bagging (bootstrap aggregating)
- Random Forests (Bagging with subsets of variables)
- Boosting (put higher weights to wrong classified data points)
- Anwendung dieser Ensemble-Methoden folgt einem gemeinsamen Schema, das zwei Schritte umfasst:

  (1) Konstruktion einer Vielzahl von Modellen (base-learners)

  Bootstrap Samples, Restriktion auf Teilmengen der Variablen, Variation der Fallgewichte

  (2) Kombination der Schätzer

  Majority Voting, Weighted averaging

Generelle Theorie

- Theoretical Foundation: Friedman & Popescu 2003
- Die vorgestellten Ansätze können als additives Modell gesehen werden

$$G(x) = \alpha_0 + \sum_{m=1}^{M} \alpha_m T_m(x)$$

- $T_m(x)$ … Base-Learners (Basisfunktionen)
- Ensemble Learner sind also ein lineares Modell in einem hochdimensionalen Raum von abgeleiteten Variablen (vgl.: Neuronale Netze, Wavelets, Multivariate Adaptive Regression Splines)

Generelle Theorie

- Jeder Base-Learner T_m kann durch einen Parametervektor $\mathbf{p_m}$ charakterisiert werden (z.B. falls T_m ein Klassifikationsbaum ist, spezifiziert $\mathbf{p_m}$ die Splits, die den Baum konstituieren)
- Allgemeine Ensemble Learning Problem:

$$\min_{\{\alpha_m, p_m\}} \sum_{i=1}^{N} L\left(y_i, \alpha_0 + \sum_{m=1}^{M} \alpha_m T(x; p_m) \right)$$

Ensemble Learning

Konkretisierung der beiden Modellierungsschritte

- A finite dictionary $\mathcal{T}_L = \{T_1(x), T_2(x), \ldots, T_M(x)\}$ of basis functions is induced from the training data;
- A family of functions $f_\lambda(x)$ is built by fitting a lasso path in this dictionary:

$$\alpha(\lambda) = \arg\min_{\alpha} \sum_{i=1}^{N} L[y_i, \alpha_0 + \sum_{m=1}^{M} \alpha_m T_m(x_i)] + \lambda \sum_{m=1}^{M} |\alpha_m|.$$

Generelle Theorie

- Friedman & Popescu zeigen, dass die Aufgabenstellung des Ensemble-Learnings im wesentlichen der Lösung eines hochdimensionalen Integrals entspricht
- Solche Integrationsprobleme werden häufig mit Techniken der Monte Carlo Integration gelöst.
- Eine wichtige Basistechnik bildet dabei das Prinzip des „Importance Sampling“, welches vorsieht, dass “wichtige” Punkte des Definitionsbereichs mit einer höheren Wahrscheinlichkeit gesampelt werden.
- Dazu ist es notwendig über **p** eine Verteilung zu definieren: Wahrscheinlichkeit invers zum Erwarteten Verlust

ISLE-Algorithmus

- Genereller Algorithmus
- *Importance sampled learning ensemble* (ISLE)
- Bagging, Random Forest, Boosting können durch unterschiedliche Sampling-Schemata als Spezialfälle des allgemeinen ISLE-Algorithmus aufgefasst werden
- Verallgemeinerungen bzw. Varianten sind möglich

Ensemble Methods

- Ensemble Methods perform extremely well in a variety of problem domains, have desirable statistical properties, and are computationally scalable (parallelization)
- However ensembles are not so easy interpretable
- While this is negligible for predictive modelling tasks, this is a severe drawback in case of descriptive modelling tasks
- In the last years new types of summary statistics have been developed to interpret ensemble models

Interpretation of Ensemble Methods

- Importance Scores

  quantify the relative influence or contribution of each variable in predicting the response

- Interaction Statistic

  to answer the question which variables are involved in interactions with other variables

- Partial dependence plots

  to understand the nature of the dependence of response on influential inputs

- Measuring Model Complexity

  Generalized degrees of freedom (Ye 1998): the better a model can match an arbitrary change of response the more complex is it

Ausblick

- Noch in Arbeit:
  Simulationsstudie um das Verhalten dieser innovativen Methodenkonzepte in praxisrelevanten Situationen zu analysieren
- Thematik: Eignung als “Off-the-Shelf” Procedure für Predictive Analytics
- Fragestellung: Wie reagiert die Diskriminationsgüte im Vergleich zu anderen Methoden
  - Verhalten in Sparsity-Situationen
  - Auswirkung von fehlenden Werten
  - Robustheit gegenüber Outlier bei den Prädiktoren bzw. falsch klassifizierten Datensätzen im Trainingsdatensatz