Beslutsträd och andra prediktiva modeller

THE POWER TO KNOW.

Mathias Lanner Sas Institute

Agenda

- Introduktion till Prediktiva modeller
- Beslutsträd
- Pruning
- Regressioner
- Neurala Nätverk
- Utvärdering av modeller

Predictive Modeling Applications

- Database marketing
- Financial risk management
- Fraud detection
- Process monitoring
- Pattern detection

Predictive Modeling Training Data

Training Data

case 1:	*inputs*	*target*
case 2:	*inputs*	*target*
case 3:	*inputs*	*target*
case 4:	*inputs*	*target*
case 5:	*inputs*	*target*

Numeric or categorical values

Predictive Modeling Score Data

Training Data

case 1:	*inputs*	*target*
case 2:	*inputs*	*target*
case 3:	*inputs*	*target*
case 4:	*inputs*	*target*
case 5:	*inputs*	*target*

Score Data

case 1:	*inputs*	*?*
case 2:	*inputs*	*?*
case 3:	*inputs*	*?*
case 4:	*inputs*	*?*
case 5:	*inputs*	*?*

Only input values known

Predictions

Training Data

case 1:	*inputs*	*target*
case 2:	*inputs*	*target*
case 3:	*inputs*	*target*
case 4:	*inputs*	*target*
case 5:	*inputs*	*target*

Predictions

Score Data

case 1:	*inputs*	*?*
case 2:	*inputs*	*?*
case 3:	*inputs*	*?*
case 4:	*inputs*	*?*
case 5:	*inputs*	*?*

Predictions

Training Data

case 1:	*inputs*	*target*
case 2:	*inputs*	*target*
case 3:	*inputs*	*target*
case 4:	*inputs*	*target*
case 5:	*inputs*	*target*

Predictions

prediction
prediction
prediction
prediction
prediction

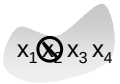
Score Data

case 1:	*inputs*	*?*
case 2:	*inputs*	*?*
case 3:	*inputs*	*?*
case 4:	*inputs*	*?*
case 5:	*inputs*	*?*

prediction
prediction
prediction
prediction
prediction

Predictive Modeling Essentials

Predict new cases

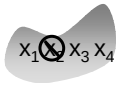


Select useful inputs

Optimize complexity

Predictive Modeling Essentials

Predict new cases



Select useful inputs

Optimize complexity

Three Prediction Types

Training Data			*Predictions*
case 1:	*inputs*	*target*	*prediction*
case 2:	*inputs*	*target*	*prediction*
case 3:	*inputs*	*target*	*prediction*
case 4:	*inputs*	*target*	*prediction*
case 5:	*inputs*	*target*	*prediction*

- **Decisions**
- **Rankings**
- **Estimates**

Decision Predictions

Training Data			*Decisions*
case 1:	*inputs*	*target*	*primary*
case 2:	*inputs*	*target*	*secondary*
case 3:	*inputs*	*target*	*tertiary*
case 4:	*inputs*	*target*	*primary*
case 5:	*inputs*	*target*	*secondary*

Trained model uses input measurements to make best decision for each case.

Ranking Predictions

Training Data			*Rankings*
case 1:	*inputs*	*target*	*720*
case 2:	*inputs*	*target*	*520*
case 3:	*inputs*	*target*	*620*
case 4:	*inputs*	*target*	*580*
case 5:	*inputs*	*target*	*470*

Trained model uses input measurements to optimally rank each case.

Estimate Predictions

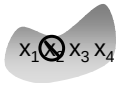
Training Data	*Estimates*
case 1: inputs target	*0.65*
case 2: inputs target	*0.33*
case 3: inputs target	*0.54*
case 4: inputs target	*0.47*
case 5: inputs target	*0.28*

Trained model uses input measurements to optimally estimate target value.

Model Essentials – Predict Review

Predict new cases

Decide, rank, estimate



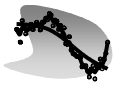
Select useful inputs

Optimize complexity

Model Essentials – Select Review

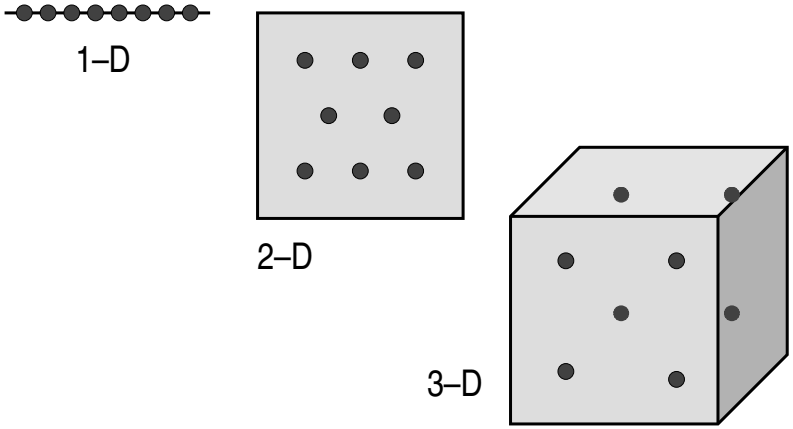
Predict new cases

Select useful inputs



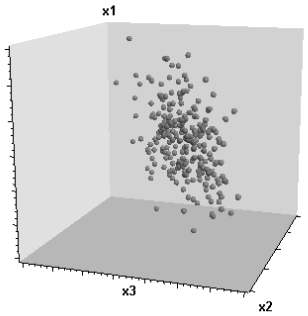
Optimize complexity

Curse of Dimensionality

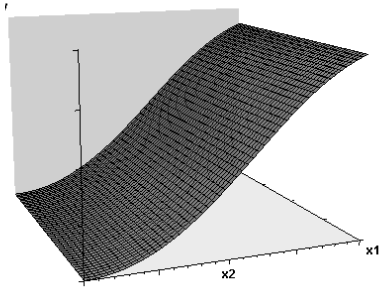


Input Selection

Redundancy

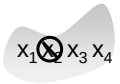


Irrelevancy

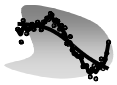


Model Essentials – Select Review

Predict new cases



Select useful inputs



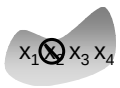
Optimize complexity

Decide, rank, estimate

Eradicate redundancies irrelevancies

Model Essentials – Optimize

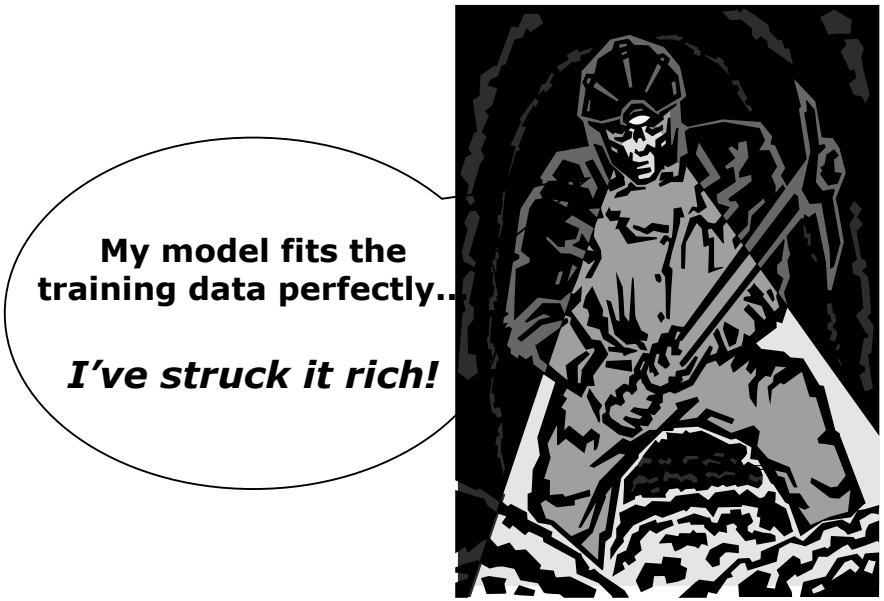
Predict new cases



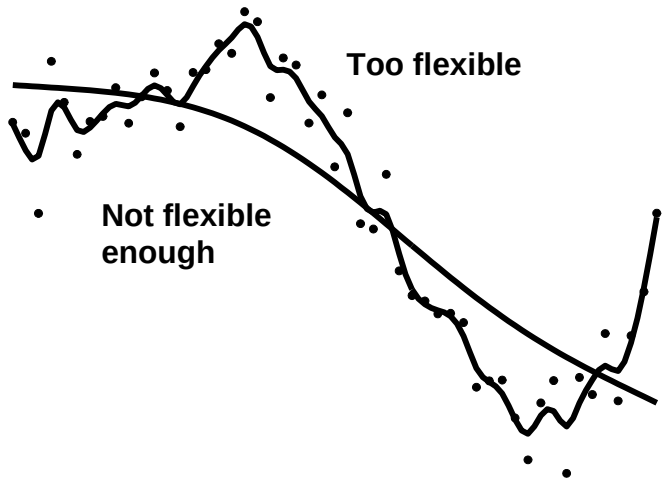
Select useful inputs

Optimize complexity

Fool's Gold



Model Complexity

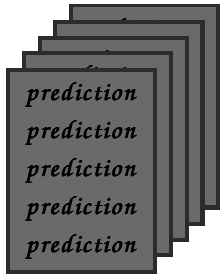


Data Splitting

Training Data Role

Training Data

case 1:	*inputs*	*target*
case 2:	*inputs*	*target*
case 3:	*inputs*	*target*
case 4:	*inputs*	*target*
case 5:	*inputs*	*target*



Training data gives sequence of predictive models with increasing complexity.

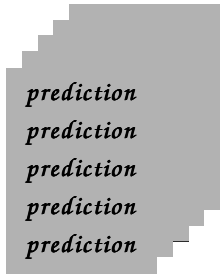
Validation Data

case 1:	*inputs*	*target*
case 2:	*inputs*	*target*
case 3:	*inputs*	*target*
case 4:	*inputs*	*target*
case 5:	*inputs*	*target*

Validation Data Role

Training Data

case 1:	*inputs*	*target*
case 2:	*inputs*	*target*
case 3:	*inputs*	*target*
case 4:	*inputs*	*target*
case 5:	*inputs*	*target*



Validation Data

case 1:	*inputs*	*target*
case 2:	*inputs*	*target*
case 3:	*inputs*	*target*
case 4:	*inputs*	*target*
case 5:	*inputs*	*target*

Validation data helps select best model from sequence

Validation Data Role

Training Data

case 1:	*inputs*	*target*
case 2:	*inputs*	*target*
case 3:	*inputs*	*target*
case 4:	*inputs*	*target*
case 5:	*inputs*	*target*

prediction
prediction
prediction
prediction
prediction

Validation Data

case 1:	*inputs*	*target*
case 2:	*inputs*	*target*
case 3:	*inputs*	*target*
case 4:	*inputs*	*target*
case 5:	*inputs*	*target*

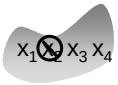
prediction
prediction
prediction
prediction
prediction

Validation data helps select best model from Sequence.

Model Essentials – Optimize

Predict new cases

Decide, rank, estimate



Select useful inputs

Eradicate redundancies irrelevancies

Optimize complexity

Tune models with validation data

Agenda

- Introduktion till Prediktiva modeller
- **Beslutsträd**
- Pruning
- Regressioner
- Neurala Nätverk
- Utvärdering av modeller

Predictive Modeling Tools

Primary

Decision Tree

Regression

Neural Network

Specialty

Dmine Regression

MBR

AutoNeural

Rule Induction

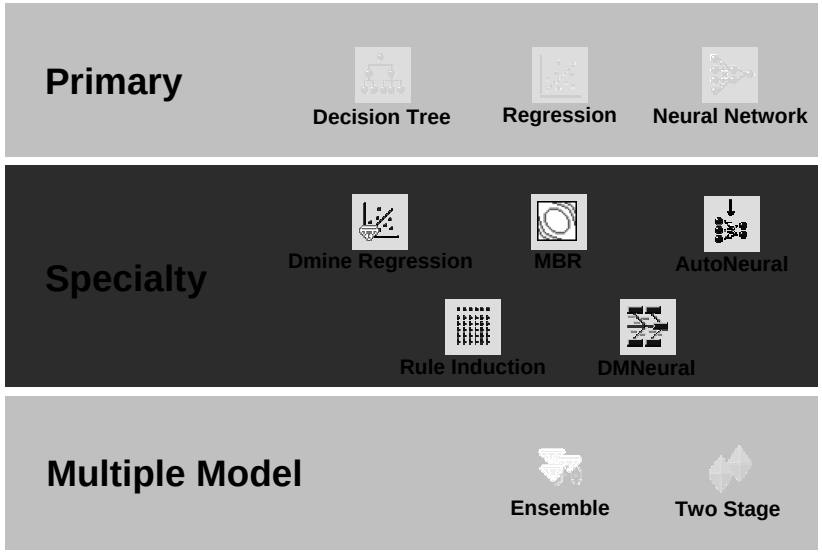
DMNeural

Multiple Model

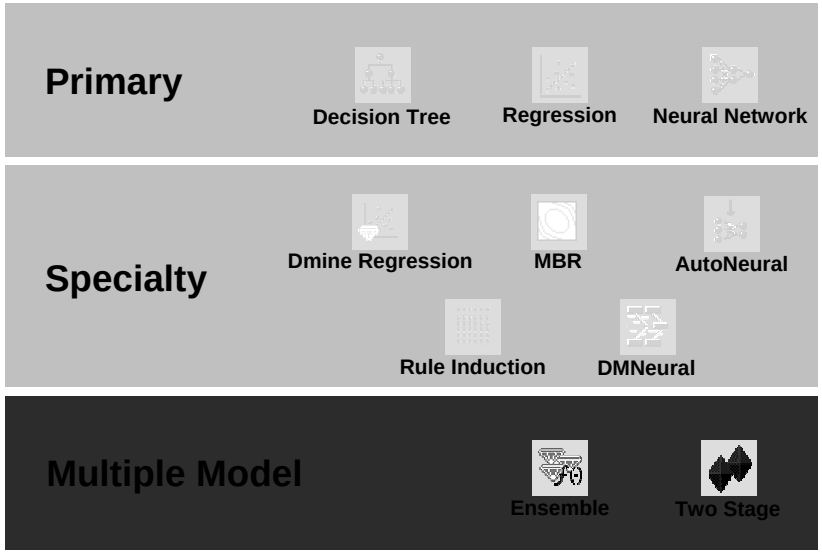
Ensemble

Two Stage

Predictive Modeling Tools



Predictive Modeling Tools



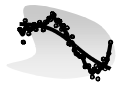
Model Essentials – Decision Trees

Predict new cases

Prediction rules

Select useful inputs

Split search



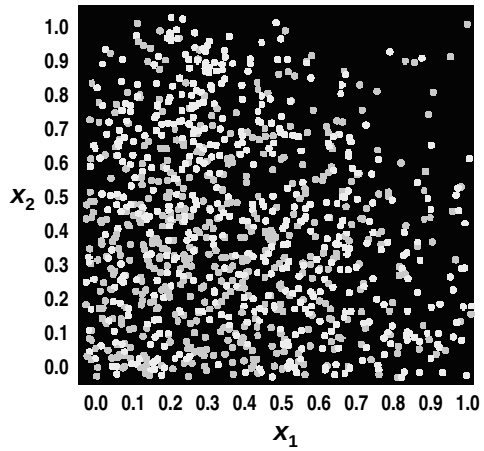
Optimize complexity

Pruning

Simple Prediction Illustration

Analysis goal:

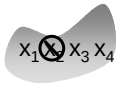
Predict the color of a dot based on its location in a scatter plot.



Model Essentials – Decision Trees

Predict new cases

Prediction rules



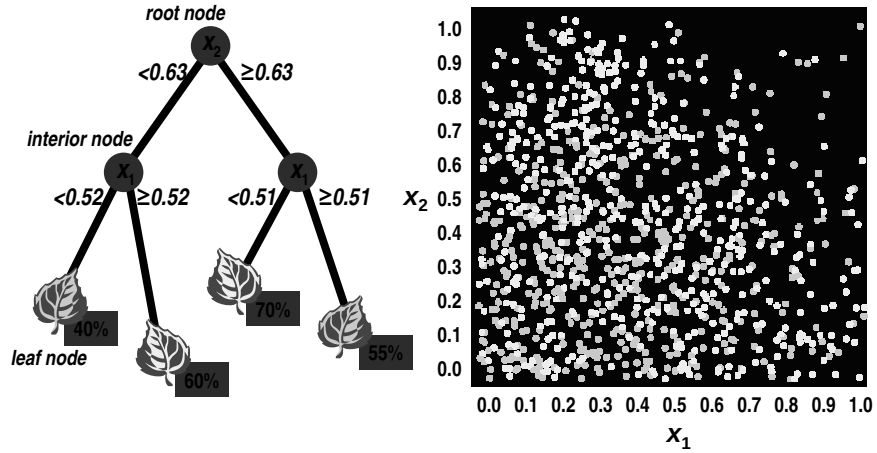
Select useful inputs

Split search

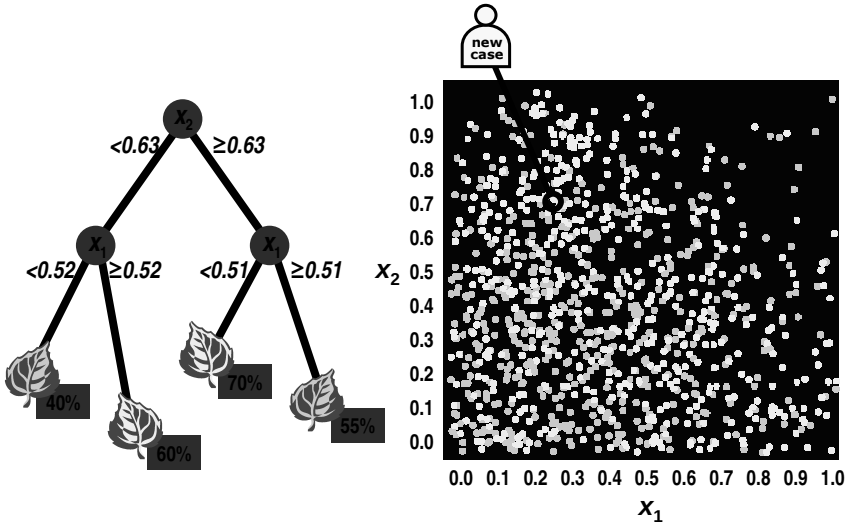
Optimize complexity

Pruning

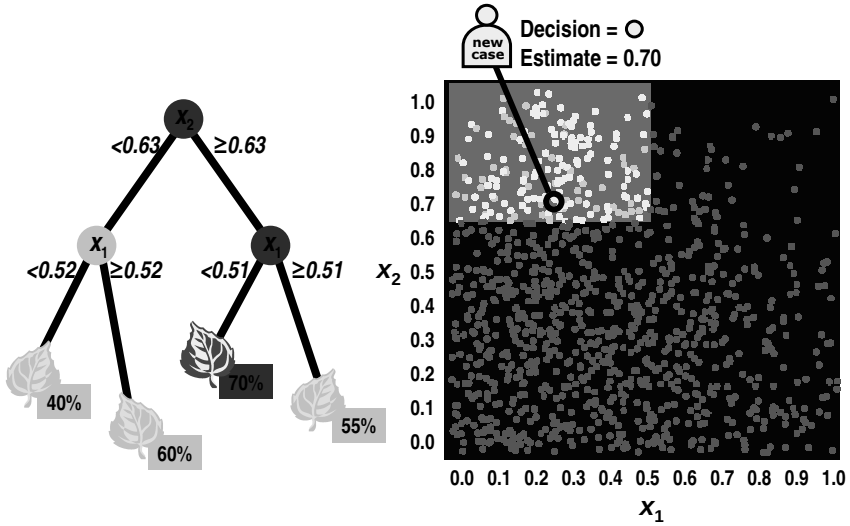
Decision Tree Prediction Rules



Decision Tree Prediction Rules



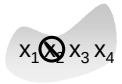
Decision Tree Prediction Rules



Model Essentials – Decision Trees

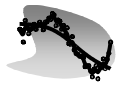
Predict new cases

Prediction rules



Select useful inputs

Split search



Optimize complexity

Pruning

Demo Beslutsträd

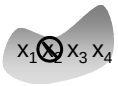
Agenda

- Introduktion till Prediktiva modeller
- **Beslutsträd**
- **Pruning**
- **Regressioner**
- **Neurala Nätverk**
- **Utvärdering av modeller**

Model Essentials – Decision Trees

Predict new cases

Prediction rules



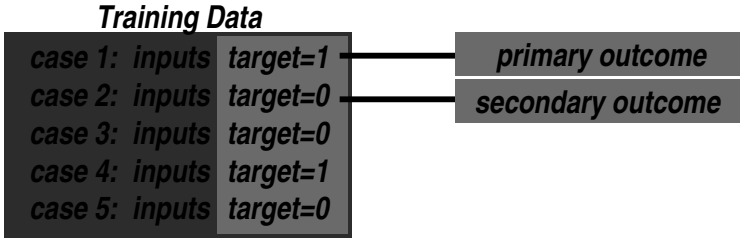
Select useful inputs

Split search

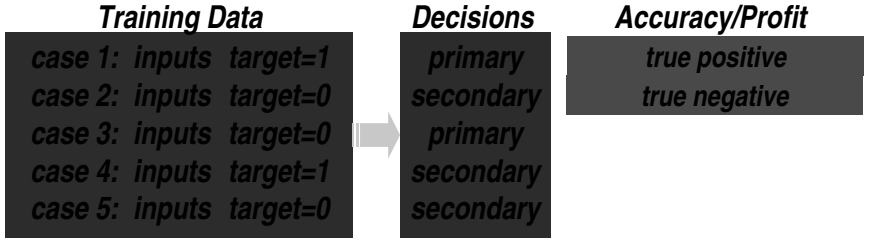
Optimize complexity

Pruning

Binary Targets

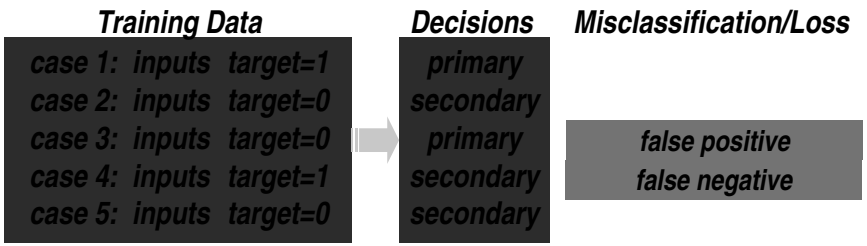


Decision Assessment



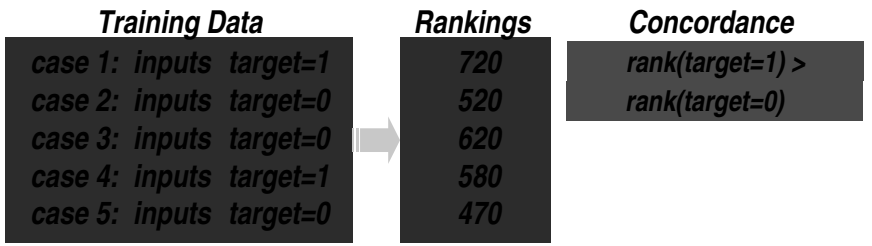
Focus on correct decisions

Decision Assessment (for Pessimists)



Focus on incorrect decisions

Ranking Assessment



Focus on correct ordering

Ranking Assessment (for Pessimists)

Training Data	Rankings	Discordance
case 1: inputs target=1	720	rank(target=1) < rank(target=0)
case 2: inputs target=0	520	
case 3: inputs target=0	620	
case 4: inputs target=1	580	
case 5: inputs target=0	470	

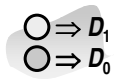
Focus on incorrect ordering

Estimate Assessment (only Pessimistic!)

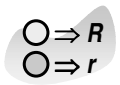
Training Data	Estimates	Squared Error
case 1: inputs target=1	0.65	$(target-estimate)^2$
case 2: inputs target=0	0.33	
case 3: inputs target=0	0.54	
case 4: inputs target=1	0.47	
case 5: inputs target=0	0.28	

Focus on incorrect estimation

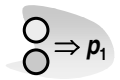
Predictive Modeling Assessments



Decisions — Accuracy, Misclassification

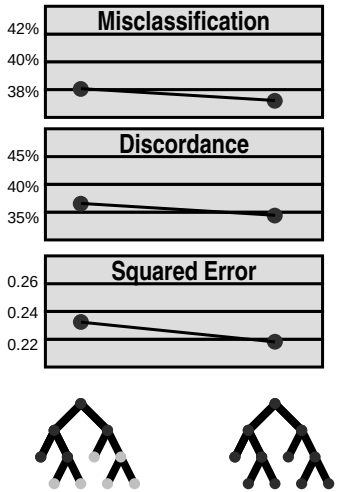


Rankings — Concordance, Discordance



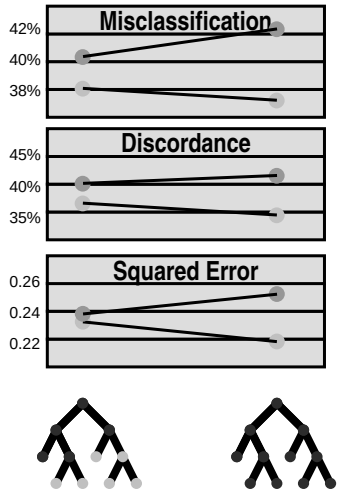
Estimates — Squared Error

Optimistic Assessment, Pessimistic Stats



Increasing leaf count *improves* assessment measures on training data.

Unbiased Assessment

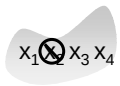


Increasing leaf count *might worsen* assessment measures on validation data.

Demo på Pruning

Model Essentials – Regressions

Predict new cases



Select useful inputs

Optimize complexity

Prediction formula

Sequential selection

Optimal sequence model

Model Essentials – Regressions

Predict new cases

Select useful inputs

Optimize complexity

Prediction formula

Sequential selection

Optimal sequence model

Linear Regression Prediction Formula

$$\hat{y} = \hat{w}_0 + \hat{w}_1 x_1 + \hat{w}_2 x_2$$

prediction estimate

intercept estimate

parameter estimate

input measurement

Choose intercept and parameter estimates to *minimize*.

squared error function

$$\sum_{\text{training data}} (y_i - \hat{y}_i)^2$$

Logistic Regression Prediction Formula

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{w}_0 + \hat{w}_1 x_1 + \hat{w}_2 x_2$$

logit scores

Choose intercept and parameter estimates to *maximize*.

log-likelihood function

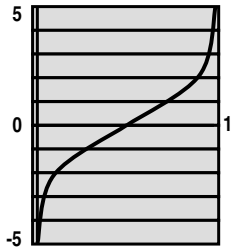
$$\sum_{\substack{\text{primary outcome} \\ \text{training cases}}} \log(\hat{p}_i) + \sum_{\substack{\text{secondary outcome} \\ \text{training cases}}} \log(1-\hat{p}_i)$$

Logit Link Function

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{w}_0 + \hat{w}_1 x_1 + \hat{w}_2 x_2$$

logit scores

logit link function



Δx_i	consequence	
1	$\Rightarrow odds \times \exp(w_i)$	*odds ratio*
doubling amount $\frac{0.69}{w_i}$	$\Rightarrow odds \times 2$	

Model interpretation

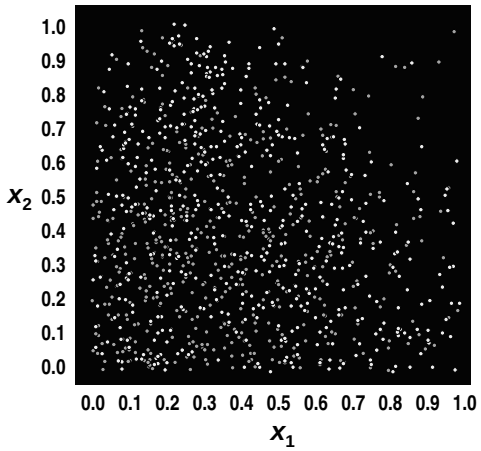
Simple Prediction Illustration – Regressions

logit equation

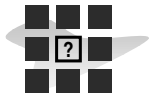
$$\text{logit}(\hat{p}) = \hat{w}_0 + \hat{w}_1 x_1 + \hat{w}_2 x_2$$

$$\hat{p} = \frac{1}{1 + e^{-\text{logit}(\hat{p})}}$$

logistic equation



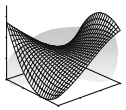
Beyond the Prediction Formula



Missing values

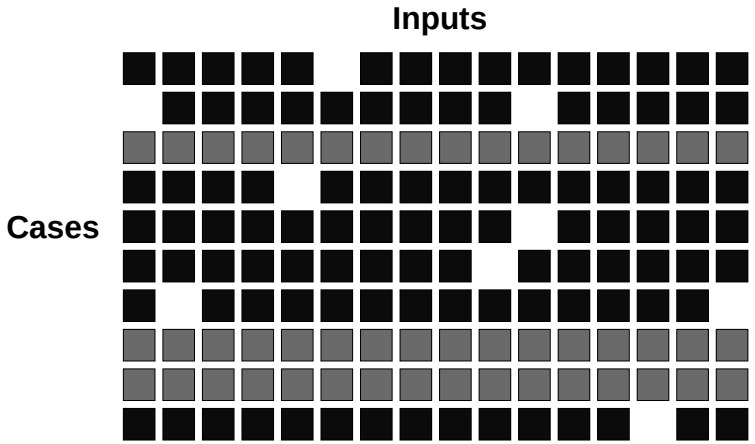
Extreme or unusual values

Non-numeric inputs

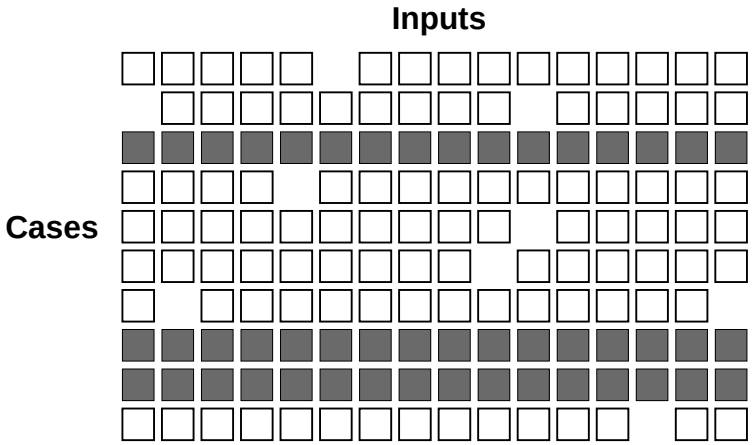


Nonlinearity and Non-additivity

Missing Values and Regression Modeling



Missing Values and Regression Modeling



Missing Values and the Prediction Formula

Prediction Formula:

$$\text{logit}(p) = -2.1 + 0.072x_1 - 0.89x_2 - 1.24x_3$$

New Case:

$$(x_1, x_2, x_3) = (2, ?, -1)$$

Predicted Value:

$$\text{logit}(p) = -2.1 + 0.144x_1 - 0.8\ ? + 1.24$$

Missing Value Causes

Not applicable

No match

Non-disclosure

Missing Value Remedies

Not applicable

No match

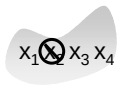
Non-disclosure

Synthetic distribution

Estimation
$x_i = f(x_1, \ldots, x_p)$

Model Essentials – Regressions

Predict new cases



Select useful inputs

Optimize complexity

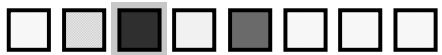
Prediction formula

Sequential selection

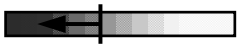
Optimal sequence model

Sequential Selection – Forward

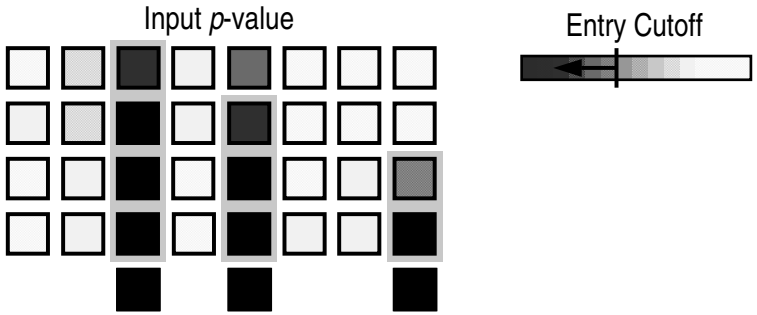
Input *p*-value



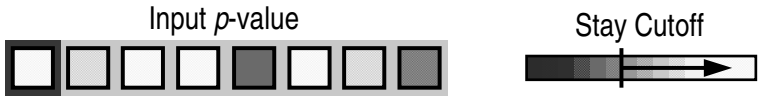
Entry Cutoff



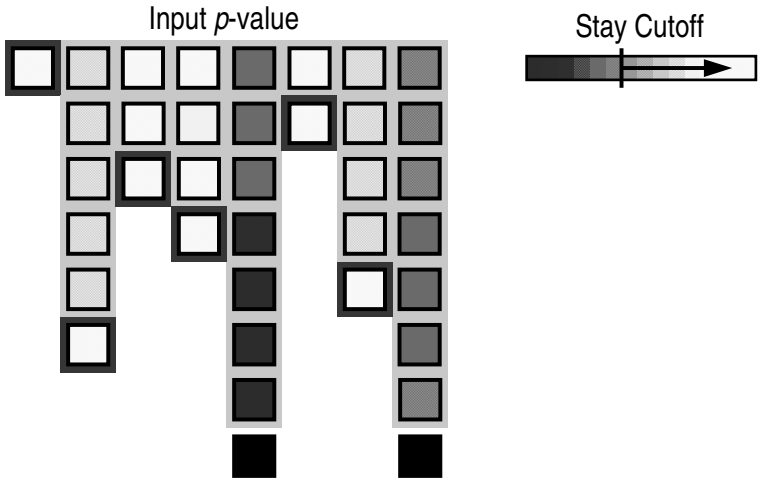
Sequential Selection – Forward



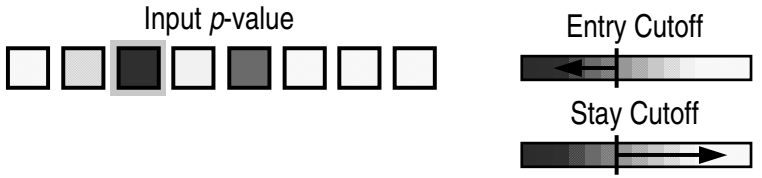
Sequential Selection – Backward



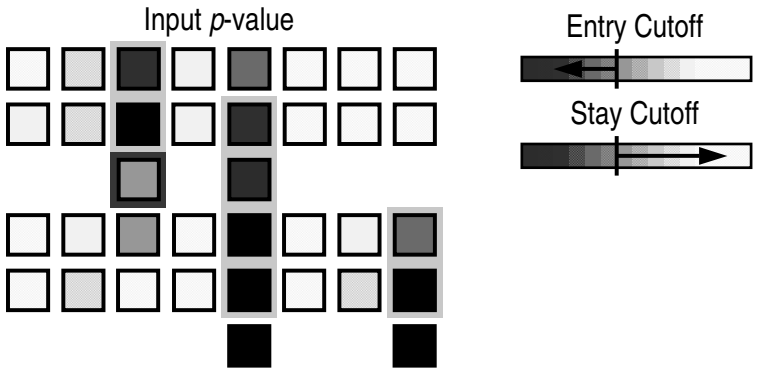
Sequential Selection – Backward



Sequential Selection – Stepwise

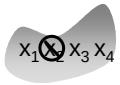


Sequential Selection – Stepwise



Model Essentials – Regressions

Predict new cases



Select useful inputs

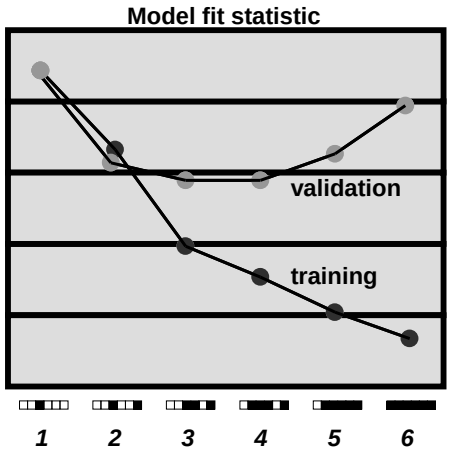
Optimize complexity

Prediction formula

Sequential selection

Optimal sequence model

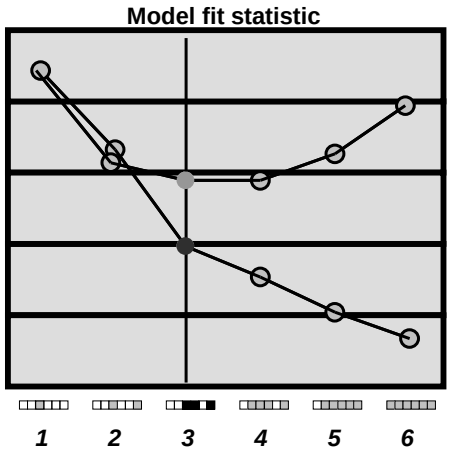
Model Fit versus Complexity



Evaluate each sequence step.

Choose simplest optimal model.

Select Model with Optimal Validation Fit



Evaluate each sequence step.

Choose simplest optimal model.

Agenda

- Introduktion till Prediktiva modeller
- Beslutsträd
- Pruning
- Regressioner
- **Neurala Nätverk**
- Utvärdering av modeller

Model Essentials – Neural Networks

Predict new cases	Prediction formula
Select useful inputs	None
Optimize complexity	Stopped training

Model Essentials – Neural Networks

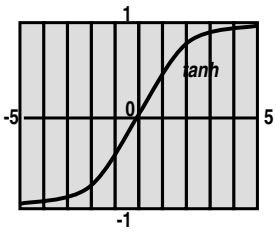
Predict new cases	Prediction formula
Select useful inputs	None
Optimize complexity	Stopped training

Neural Network Prediction Formula

hidden unit

prediction estimate $\hat{y} = \hat{w}_{00} + \hat{w}_{01} H_1 + \hat{w}_{02} H_2 + \hat{w}_{03} H_3$

bias estimate *weight estimate*



$H_1 = \tanh(\hat{w}_{10} + \hat{w}_{11} x_1 + \hat{w}_{12} x_2)$

$H_2 = \tanh(\hat{w}_{20} + \hat{w}_{21} x_1 + \hat{w}_{22} x_2)$

$H_3 = \tanh(\hat{w}_{30} + \hat{w}_{31} x_1 + \hat{w}_{32} x_2)$

activation function

Neural Network Binary Prediction Formula

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{w}_{00} + \hat{w}_{01} H_1 + \hat{w}_{02} H_2 + \hat{w}_{03} H_3$$

logit link function

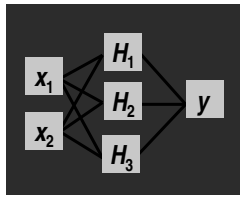
$$H_1 = \tanh(\hat{w}_{10} + \hat{w}_{11} x_1 + \hat{w}_{12} x_2)$$

$$H_2 = \tanh(\hat{w}_{20} + \hat{w}_{21} x_1 + \hat{w}_{22} x_2)$$

$$H_3 = \tanh(\hat{w}_{30} + \hat{w}_{31} x_1 + \hat{w}_{32} x_2)$$

Neural Network Diagram

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{w}_{00} + \hat{w}_{01} H_1 + \hat{w}_{02} H_2 + \hat{w}_{03} H_3$$



$$H_1 = \tanh(\hat{w}_{10} + \hat{w}_{11} x_1 + \hat{w}_{12} x_2)$$

$$H_2 = \tanh(\hat{w}_{20} + \hat{w}_{21} x_1 + \hat{w}_{22} x_2)$$

$$H_3 = \tanh(\hat{w}_{30} + \hat{w}_{31} x_1 + \hat{w}_{32} x_2)$$

Prediction Illustration – Neural Networks

logit equation

$$\text{logit}(\hat{p}) = \hat{w}_{00} + \hat{w}_{01} H_1 + \hat{w}_{02} H_2 + \hat{w}_{03} H_3$$

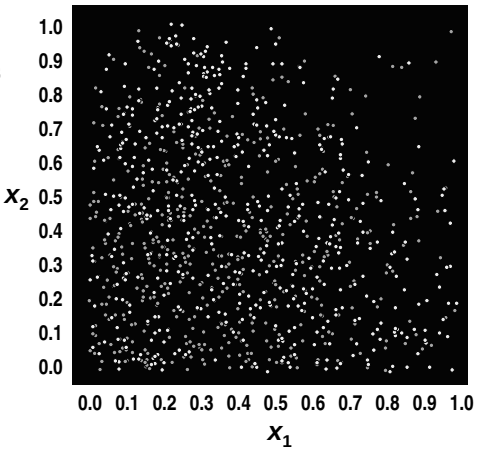
$$H_1 = \tanh(\hat{w}_{10} + \hat{w}_{11} x_1 + \hat{w}_{12} x_2)$$

$$H_2 = \tanh(\hat{w}_{20} + \hat{w}_{21} x_1 + \hat{w}_{22} x_2)$$

$$H_3 = \tanh(\hat{w}_{30} + \hat{w}_{31} x_1 + \hat{w}_{32} x_2)$$

$$\hat{p} = \frac{1}{1 + e^{-\text{logit}(\hat{p})}}$$

logistic equation



Agenda

- Introduktion till Prediktiva modeller
- Beslutsträd
- Pruning
- Regressioner
- Neurala Nätverk
- Utvärdering av modeller

Assessment Types

The Model Comparison tool provides

- Summary statistics
- Statistical graphics

Summary Statistics Summary

Prediction Type	*Statistic*
Decisions	Accuracy / Misclassification Profit / Loss KS-statistic
1,2,3,... Rankings	ROC Index (concordance) Gini coefficient
$\hat{p} \approx E(Y)$ Estimates	Average squared error SBC / Likelihood

Summary Statistics Summary

Prediction Type	*Statistic*
Decisions	Accuracy / Misclassification Profit / Loss KS-statistic
1,2,3,... Rankings	ROC Index (concordance) Gini coefficient
$\hat{p} \approx E(Y)$ Estimates	Average squared error SBC / Likelihood

Summary Statistics Summary

Prediction Type	*Statistic*
Decisions	Accuracy / Misclassification Profit / Loss KS-statistic
1,2,3,... Rankings	ROC Index (concordance) Gini coefficient
$\hat{p} \approx E(Y)$ Estimates	Average squared error SBC / Likelihood

Statistical Graphics Summary

Prediction Type	*Statistic*
 **Decisions**	
1,2,3,… **Rankings**	**Sensitivity charts** **Response rate charts**
$\hat{p} \approx E(Y)$ **Estimates**	