Data Mining Techniques for Prognosis in Pancreatic Cancer

Masters Thesis Presentation

Stuart Floyd
AIRG
April 26, 2007

Advisors: Professor Carolina Ruiz,
Professor Sergio Alvarez (Boston College)

UMass Collaborators: Professor Jennifer Tseng,
Professor Giles Whalen

Motivation

- Hospitals have databases of patient level data
- Data has been useful for finding overall trends
- Use data to build models of best treatment for individual patients?

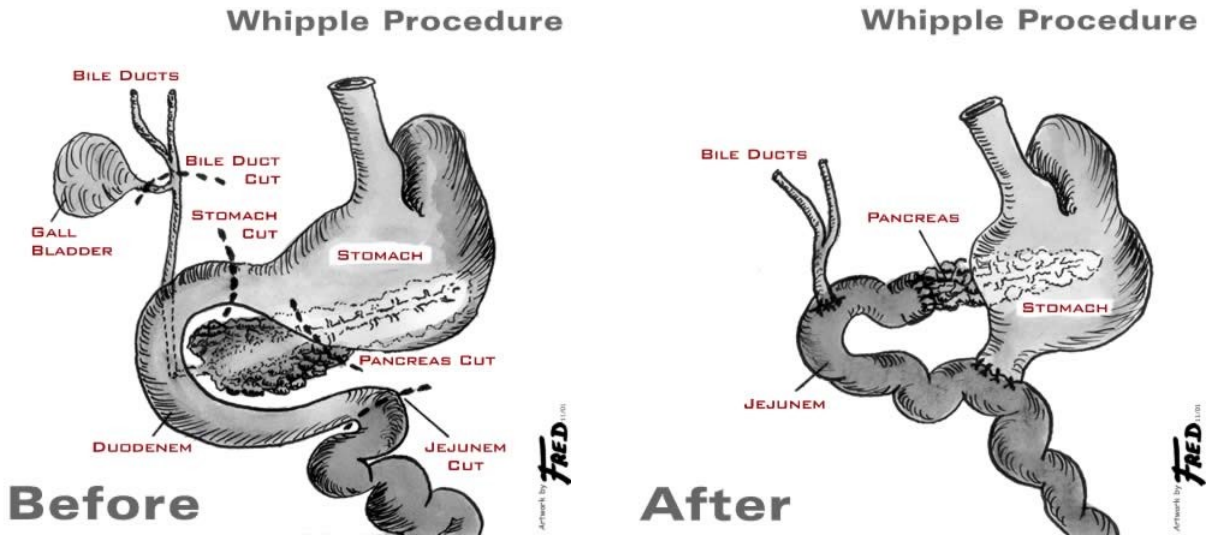
Motivation:
Study of Pancreatic Cancer

- 2.5% of cancer cases in US
- 4th largest killer among cancers in US

Source: http://www.cancer.org/downloads/STT/CAFF2007PWSecured.pdf

Motivation:
Modeling Expected Survival

- Surgical Removal of Pancreatic Tumor:
  - Best way to increase chance of curing the cancer

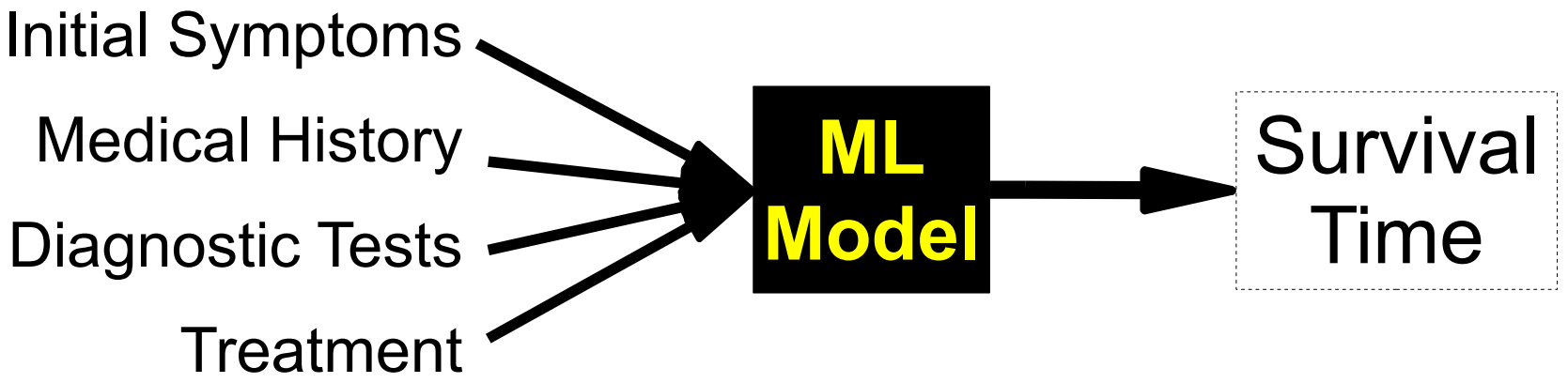


- Only remove tumor if patient's expected survival time greater then recovery time

Whipple Pictures Source: http://www.ddc.musc.edu/ddc_pub/patientInfo/surgeries/pancreatic/page06.htm

Problem Statement

Want to investigate the best techniques to construct models that are able to reliably predict the expected survival of a patient with pancreatic cancer.



Challenges

- 60 patients in dataset
  - Patients first seen between April 2002 and December 2005 with date of death information
- 189 Total Attributes for Each Patient
  - 23 Initial Symptoms
  - 28 Medical History
  - 8 Lab Scores
  - 48 Imaging Scores
  - 82 Treatment/Post Op

Prior Work:
John Hayward's Thesis

Targets Investigated:
- Tumor Size
- T-staging, N-staging
- Vascular Involvement
- Tumor Histology and Malignancy
- Survival Rates
- ECOG Score at 6-month, 9-month and 12-month intervals

Machine Learning Algorithms:
- Majority Class
- OneR
- J48
- Locally Weighted Learning
- Artificial Neural Networks
- Bayesian Networks
- Naïve Bayes
- Logistic Regression

Evaluation Techniques:
- 10 Fold Cross Validation

Feature Selection:
- CfsSubsetEval

Meta-Learning Algorithms:
- Bagging
- Boosting
- Stacking

Prior Work:
John Hayward's Thesis

Targets Investigated:
- Tumor Size
- T-staging, N-staging
- Vascular Involvement
- Tumor Histology and Malignancy
- Survival Rates
- ECOG Score at 6-month, 9-month and 12-month intervals

Machine Learning Algorithms:
- Majority Class
- OneR
- J48
- Locally Weighted Learning
- Artificial Neural Networks
- Bayesian Networks
- Naïve Bayes
- Logistic Regression

Evaluation Techniques:
- 10 Fold Cross Validation

Feature Selection:
- CfsSubsetEval

Meta-Learning Algorithms:
- Bagging
- Boosting
- Stacking

Prior Work:
John Hayward's Thesis

Targets Investigated:
- Tumor Size
- T-staging, N-staging
- Vascular Involvement
- Tumor Histology and Malignancy
- Survival Rates
- ECOG Score at 6-month, 9-month and 12-month intervals

Machine Learning Algorithms:
- Majority Class
- OneR
- J48
- Locally Weighted Learning
- Artificial Neural Networks
- Bayesian Networks
- Naïve Bayes
- Logistic Regression

Evaluation Techniques:
- 10 Fold Cross Validation

Feature Selection:
- CfsSubsetEval

Meta-Learning Algorithms:
- Bagging
- Boosting
- Stacking

Prior Work:
John Hayward's Thesis

Targets Investigated:
- Tumor Size
- T-staging, N-staging
- Vascular Involvement
- Tumor Histology and Malignancy
- Survival Rates
- ECOG Score at 6-month, 9-month and 12-month intervals

Machine Learning Algorithms:
- Majority Class
- OneR
- J48
- Locally Weighted Learning
- Artificial Neural Networks
- Bayesian Networks
- Naïve Bayes
- Logistic Regression

Evaluation Techniques:
- 10 Fold Cross Validation

Feature Selection:
- CfsSubsetEval

Meta-Learning Algorithms:
- Bagging
- Boosting
- Stacking

Prior Work:
John Hayward's Thesis

Targets Investigated:
- Tumor Size
- T-staging, N-staging
- Vascular Involvement
- Tumor Histology and Malignancy
- Survival Rates
- ECOG Score at 6-month, 9-month and 12-month intervals

Machine Learning Algorithms:
- Majority Class
- OneR
- J48
- Locally Weighted Learning
- Artificial Neural Networks
- Bayesian Networks
- Naïve Bayes
- Logistic Regression

Evaluation Techniques:
- 10 Fold Cross Validation

Feature Selection:
- CfsSubsetEval

Meta-Learning Algorithms:
- Bagging
- Boosting
- Stacking

Conference Papers Published

J. Hayward, S.A. Alvarez, C. Ruiz, J. Tseng, M. Sullivan, G. Whalen. "Survival of Patients with Pancreatic Cancer Predicted using Machine Learning Techniques". Society of Surgical Oncology's 60th Annual Cancer Symposium. Washington DC, USA. March 2007.

S. Floyd, S.A. Alvarez, C. Ruiz, J. Hayward, M. Sullivan, M. Tseng, G. Whalen. "Improved Survival Prediction for Pancreatic Cancer using Machine Learning and Regression". The Society for Surgery of the Alimentary Tract (SSAT), Digestive Diseases Week. Washington DC, USA. May 2007

Thesis Overview

Targets Investigated:

- Survival Rates
  - All Attributes:
    - <6, 6-12, >12 Month Survival
    - <9, >9 Month Survival
    - <6, >6 Month Survival
  - Pre-Operative Attributes:
    - <6, 6-12, >12 Month Survival
    - <9, >9 Month Survival
    - <6, >6 Month Survival

Machine Learning Algorithms:

- Majority Class
- J48
- Artificial Neural Networks
- Support Vector Machines
- Bayesian Networks
- Naïve Bayes
- Logistic Regression

Evaluation Techniques:

- 10 Fold Cross Validation
- Attribute Selected Classifier
- ROC Curves

Feature Selection:

- Gain Ratio
- Principal Components
- ReliefF
- Support Vector Machines

Meta-Learning Algorithms:

- Bagging
- Boosting
- Stacking
- Our Model Selector

Thesis Overview

Targets Investigated:

- Survival Rates
  - All Attributes:
    - <6, 6-12, >12 Month Survival
    - <9, >9 Month Survival
    - <6, >6 Month Survival
  - Pre-Operative Attributes:
    - <6, 6-12, >12 Month Survival
    - <9, >9 Month Survival
    - <6, >6 Month Survival

Machine Learning Algorithms:

- Majority Class
- J48
- Artificial Neural Networks
- Support Vector Machines
- Bayesian Networks
- Naïve Bayes
- Logistic Regression

Evaluation Techniques:

- 10 Fold Cross Validation
- Attribute Selected Classifier
- ROC Curves

Feature Selection:

- Gain Ratio
- Principal Components
- ReliefF
- Support Vector Machines

Meta-Learning Algorithms:

- Bagging
- Boosting
- Stacking
- Our Model Selector

Thesis Overview

Targets Investigated:

- Survival Rates
  - All Attributes:
    - <6, 6-12, >12 Month Survival
    - <9, >9 Month Survival
    - <6, >6 Month Survival
  - Pre-Operative Attributes:
    - <6, 6-12, >12 Month Survival
    - <9, >9 Month Survival
    - <6, >6 Month Survival

Machine Learning Algorithms:

- Majority Class
- J48
- Artificial Neural Networks
- Support Vector Machines
- Bayesian Networks
- Naïve Bayes
- Logistic Regression

Evaluation Techniques:

- 10 Fold Cross Validation
- Attribute Selected Classifier
- ROC Curves

Feature Selection:

- Gain Ratio
- Principal Components
- ReliefF
- Support Vector Machines

Meta-Learning Algorithms:

- Bagging
- Boosting
- Stacking
- Our Model Selector

Thesis Overview

Targets Investigated:

- Survival Rates
  - All Attributes:
    - <6, 6-12, >12 Month Survival
    - <9, >9 Month Survival
    - <6, >6 Month Survival
  - Pre-Operative Attributes:
    - <6, 6-12, >12 Month Survival
    - <9, >9 Month Survival
    - <6, >6 Month Survival

Machine Learning Algorithms:

- Majority Class
- J48
- Artificial Neural Networks
- Support Vector Machines
- Bayesian Networks
- Naïve Bayes
- Logistic Regression

Evaluation Techniques:

- 10 Fold Cross Validation
- Attribute Selected Classifier
- ROC Curves

Feature Selection:

- Gain Ratio
- Principal Components
- ReliefF
- Support Vector Machines

Meta-Learning Algorithms:

- Bagging
- Boosting
- Stacking
- Our Model Selector

Thesis Overview

Targets Investigated:

- Survival Rates
  - All Attributes:
    - <6, 6-12, >12 Month Survival
    - <9, >9 Month Survival
    - <6, >6 Month Survival
  - Pre-Operative Attributes:
    - <6, 6-12, >12 Month Survival
    - <9, >9 Month Survival
    - <6, >6 Month Survival

Machine Learning Algorithms:

- Majority Class
- J48
- Artificial Neural Networks
- Support Vector Machines
- Bayesian Networks
- Naïve Bayes
- Logistic Regression

Evaluation Techniques:

- 10 Fold Cross Validation
- Attribute Selected Classifier
- ROC Curves

Feature Selection:

- Gain Ratio
- Principal Components
- ReliefF
- Support Vector Machines

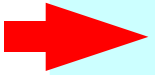
Meta-Learning Algorithms:

- Bagging
- Boosting
- Stacking
- Our Model Selector

Thesis Overview

Targets Investigated:

- Survival Rates
  - All Attributes:
    - <6, 6-12, >12 Month Survival
    - <9, >9 Month Survival
    - <6, >6 Month Survival
  - Pre-Operative Attributes:
    - <6, 6-12, >12 Month Survival
    - <9, >9 Month Survival
    - <6, >6 Month Survival



Machine Learning Algorithms:

- Majority Class
- J48
- Artificial Neural Networks
- Support Vector Machines
- Bayesian Networks
- Naïve Bayes
- Logistic Regression

Evaluation Techniques:

- 10 Fold Cross Validation
- Attribute Selected Classifier
- ROC Curves

Feature Selection:

- Gain Ratio
- Principal Components
- ReliefF
- Support Vector Machines

Meta-Learning Algorithms:

- Bagging
- Boosting
- Stacking
- Our Model Selector

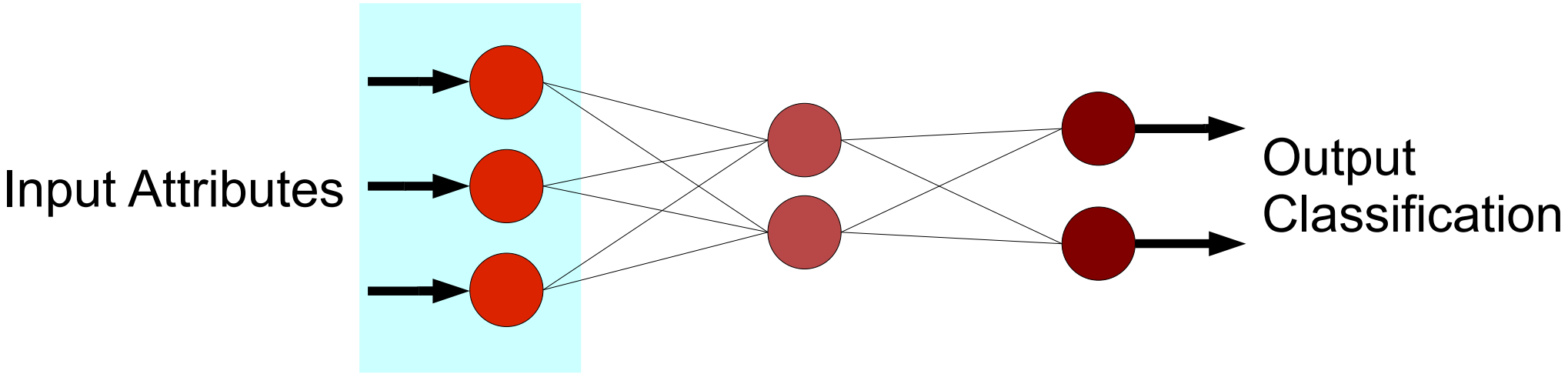
Benchmark Algorithms

- Logistic Regression
  - Trusted by medical community
  - Find probability of a response given set of explanatory variables
- Majority Class
  - Always returns the response variable that is most common in the training set

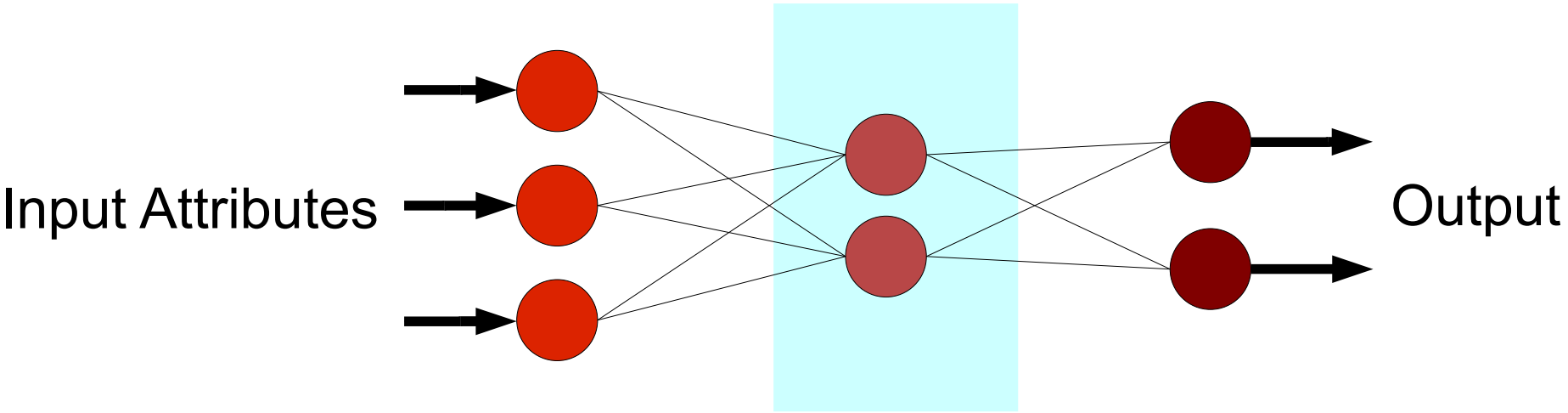
Machine Learning Algorithms

- Machine learning algorithms construct models
- Each data point in a dataset is an instance
  - Models built using training set of instances
  - Models tested using test set of instances
- Accuracy of a model is percentage of test instances correctly predicted by model
- Input attributes are the explanatory variables
- Class or target class is the response variable

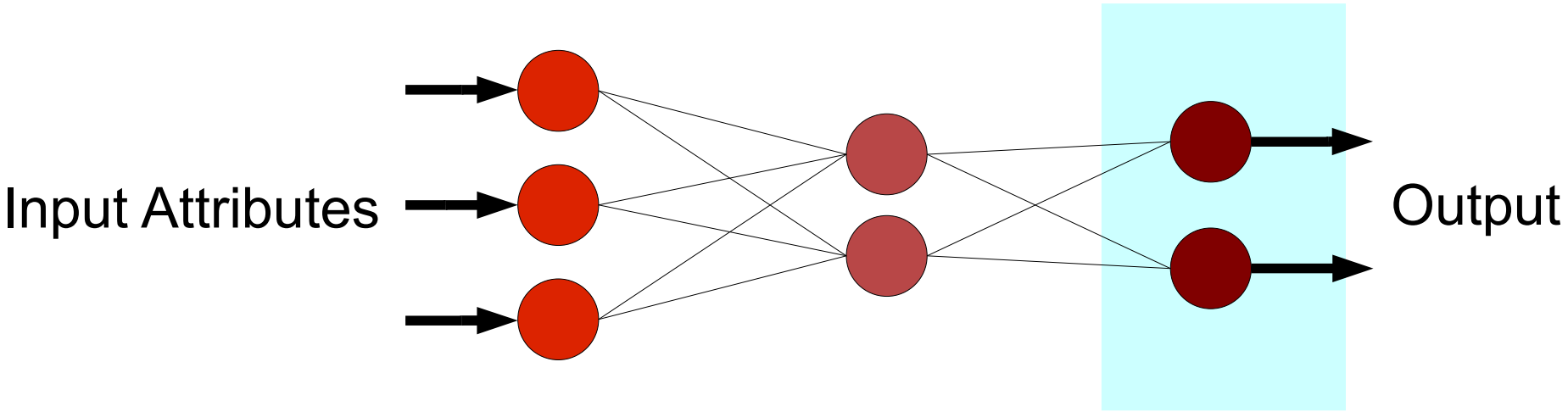
Machine Learning Algorithms: Artificial Neural Networks



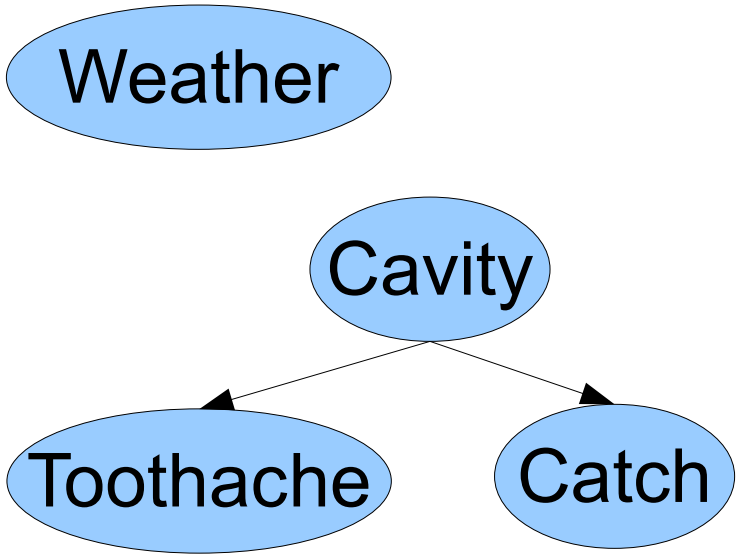
Machine Learning Algorithms: Artificial Neural Networks



Machine Learning Algorithms: Artificial Neural Networks



Machine Learning Algorithms: Bayesian Networks



- Nodes model dependencies
- Weather independent of other variables
- Toothache & Catch conditionally independent given Cavity

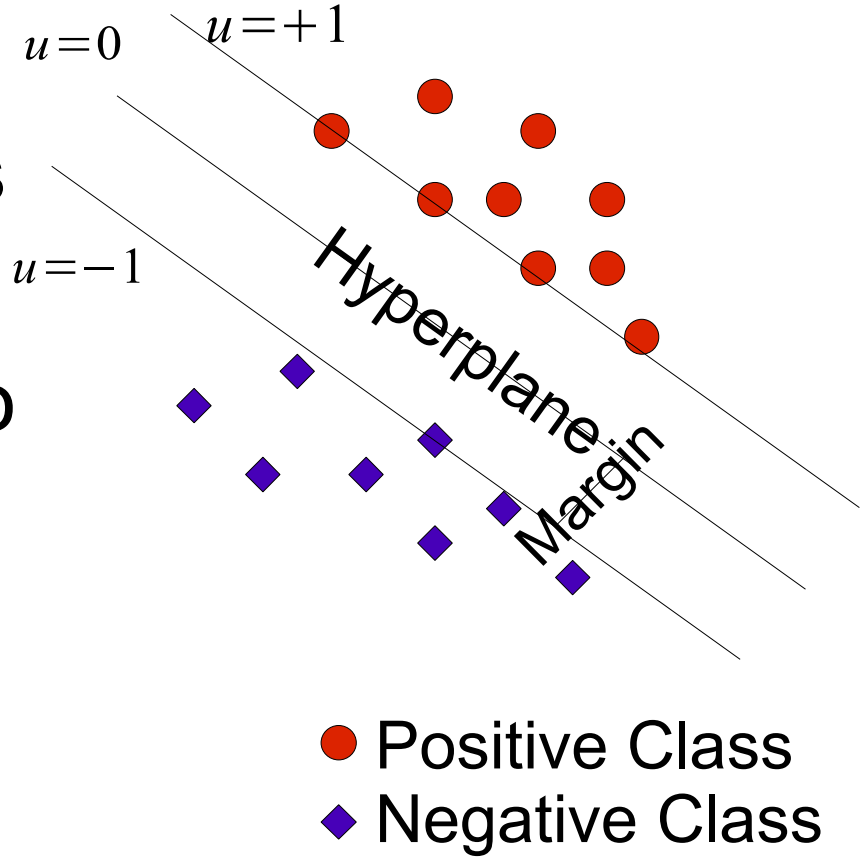
Based on Bayes Theorem:

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

Example from: Stuart J. Russell and Peter Norvig. Artificial Intelligence A Modern Approach. Upper Saddle River NJ: Prentice Hall, 200

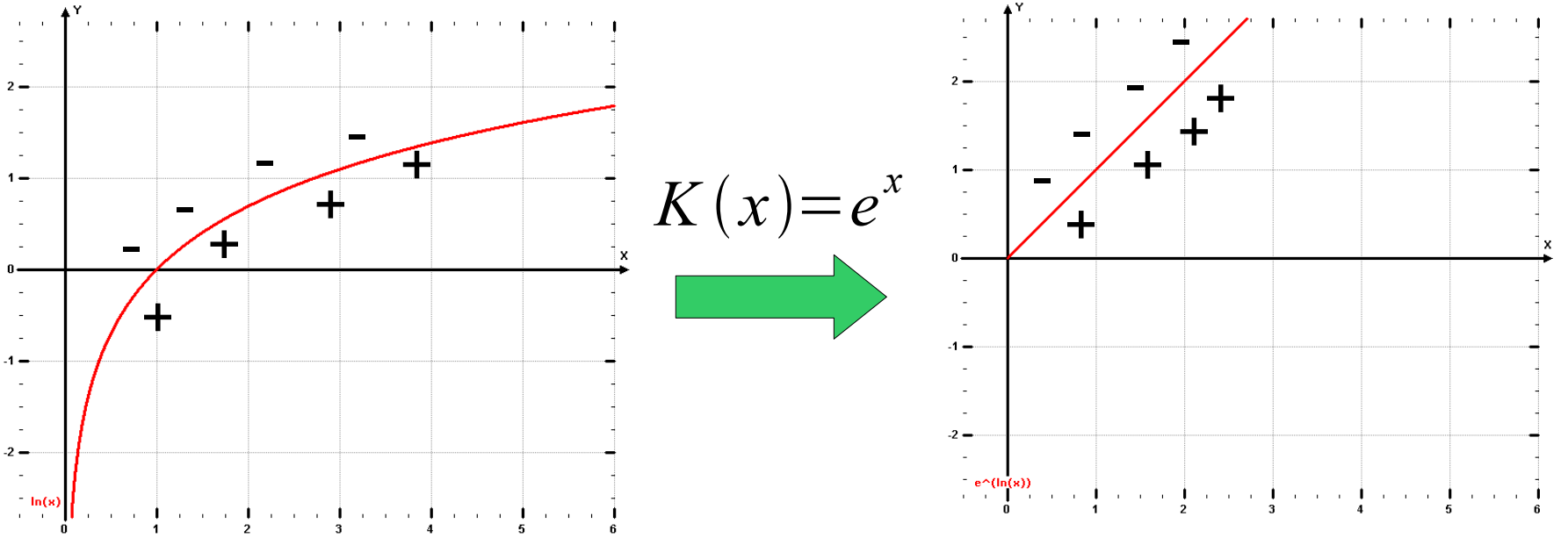
Machine Learning Algorithms: Support Vector Machines

- Attempt to fit a hyperplane that separates two groups of instances
- Use kernel function to transform instance space to make linearly separable



Machine Learning Algorithms: Support Vector Machines - Kernel Functions

- Map input space into new feature space
  - Consider Input Space $f(x)=\ln(x)$
  - Use Kernel Function: $K(x)=e^{x}$
  - To Transform Input Space Into: $K(f(x))=e^{\ln(x)}=x$
- Many Types of Generic Kernel Functions



Machine Learning Algorithm Results

Model Construction Method	Accuracy	Better then Majority Class
Majority Class	50.0	
Logistic	58.8	
SMO_Kernel:1.0	62.5	Yes
ANN_1HU	58.5	
ANN_2HU	58.0	
NaiveBayes	49.3	
Bayes Net 1 Parent	55.0	
Bayes Net 2 Parents	64.7	Yes

$P<0.05$

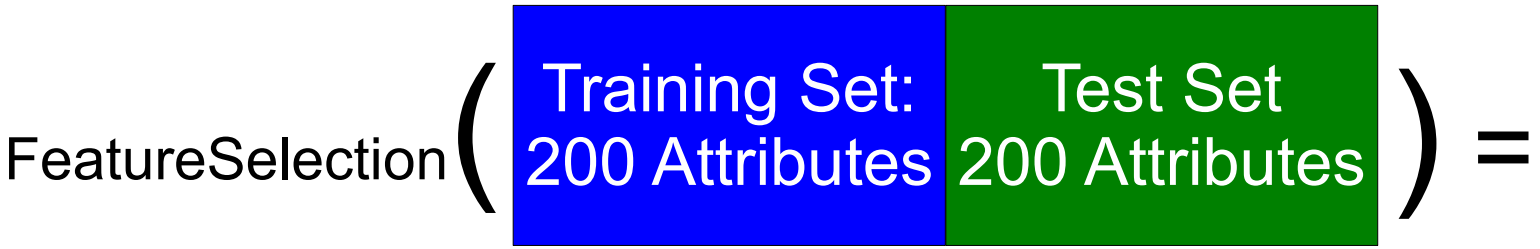
Feature Selection

- Reduces dimensionality of the input space
- Lets machine learning algorithms focus on modeling most relevant features
- Thesis investigates:
  - Gain Ratio
  - Principal Components
  - ReliefF
  - Support Vector Machines

Feature Selection: How Not To Select Best Features

- When use training data as test data, get much higher accuracy then when a model is tested on new data
  - Same applies to feature selection!

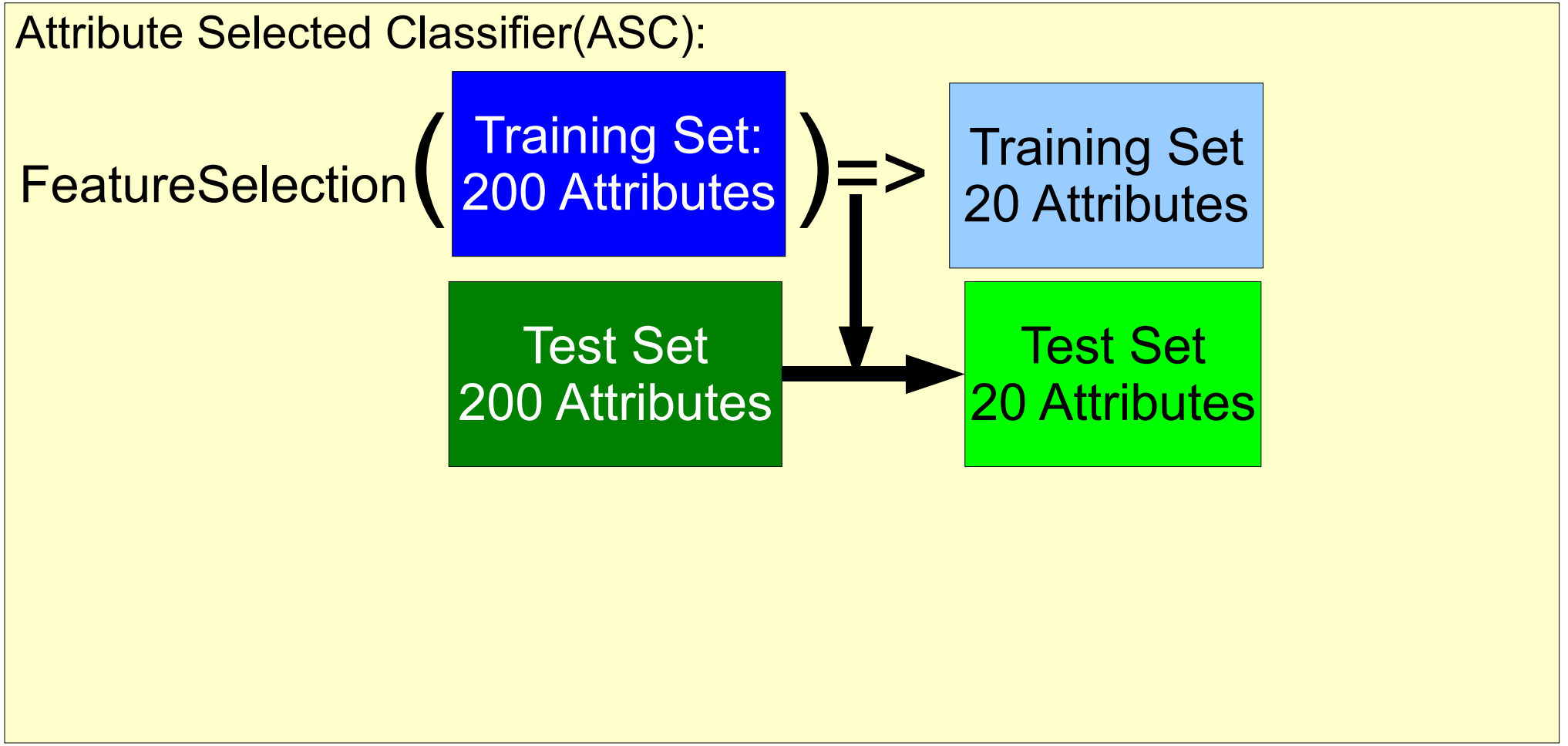
Hayward's Use of Feature Selection:



Feature Selection:
Attribute Selected Classifier

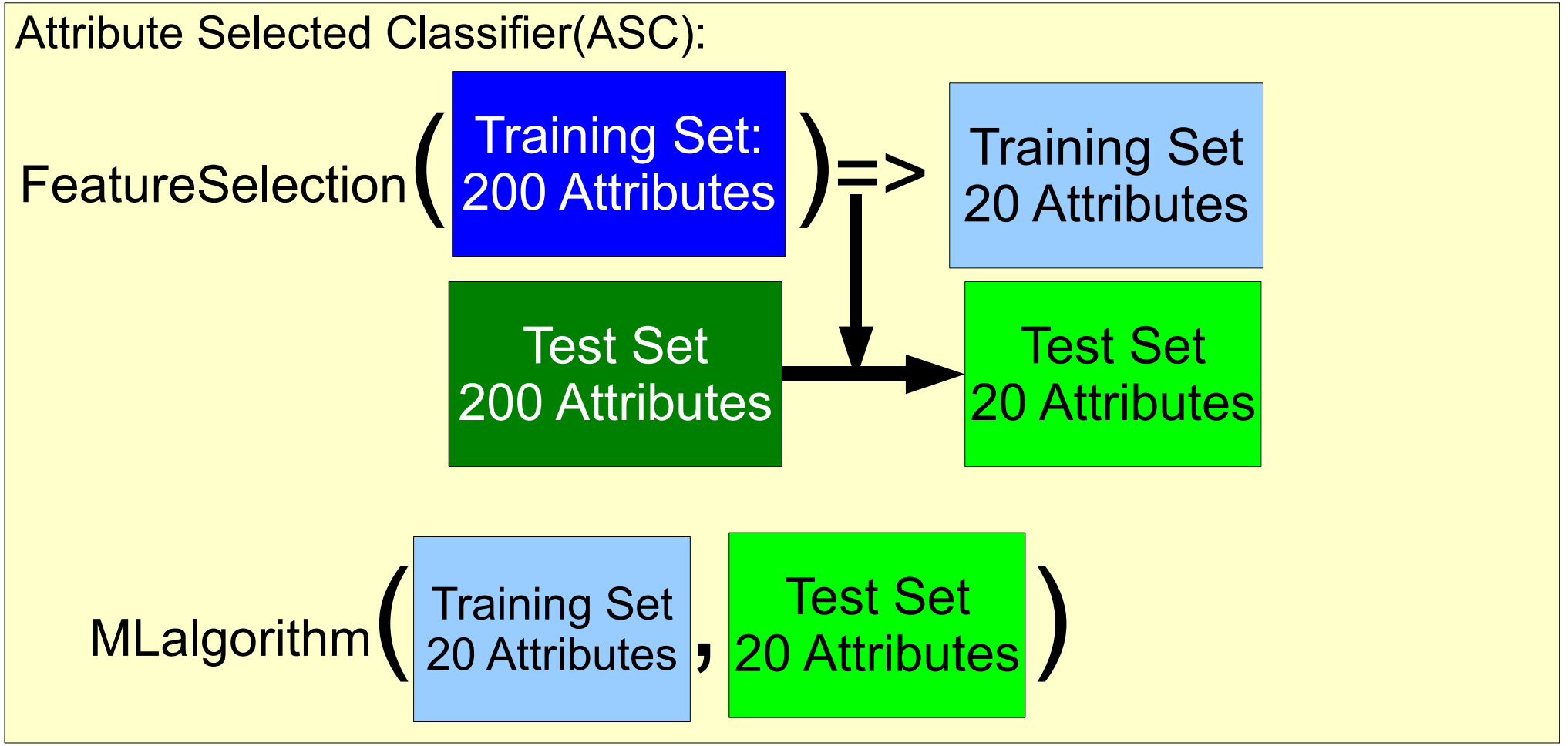
Feature Selection: Attribute Selected Classifier

ASC(MLalgorithm, Feature Selection, Training Set: 200 Attributes , Test Set 200 Attributes)

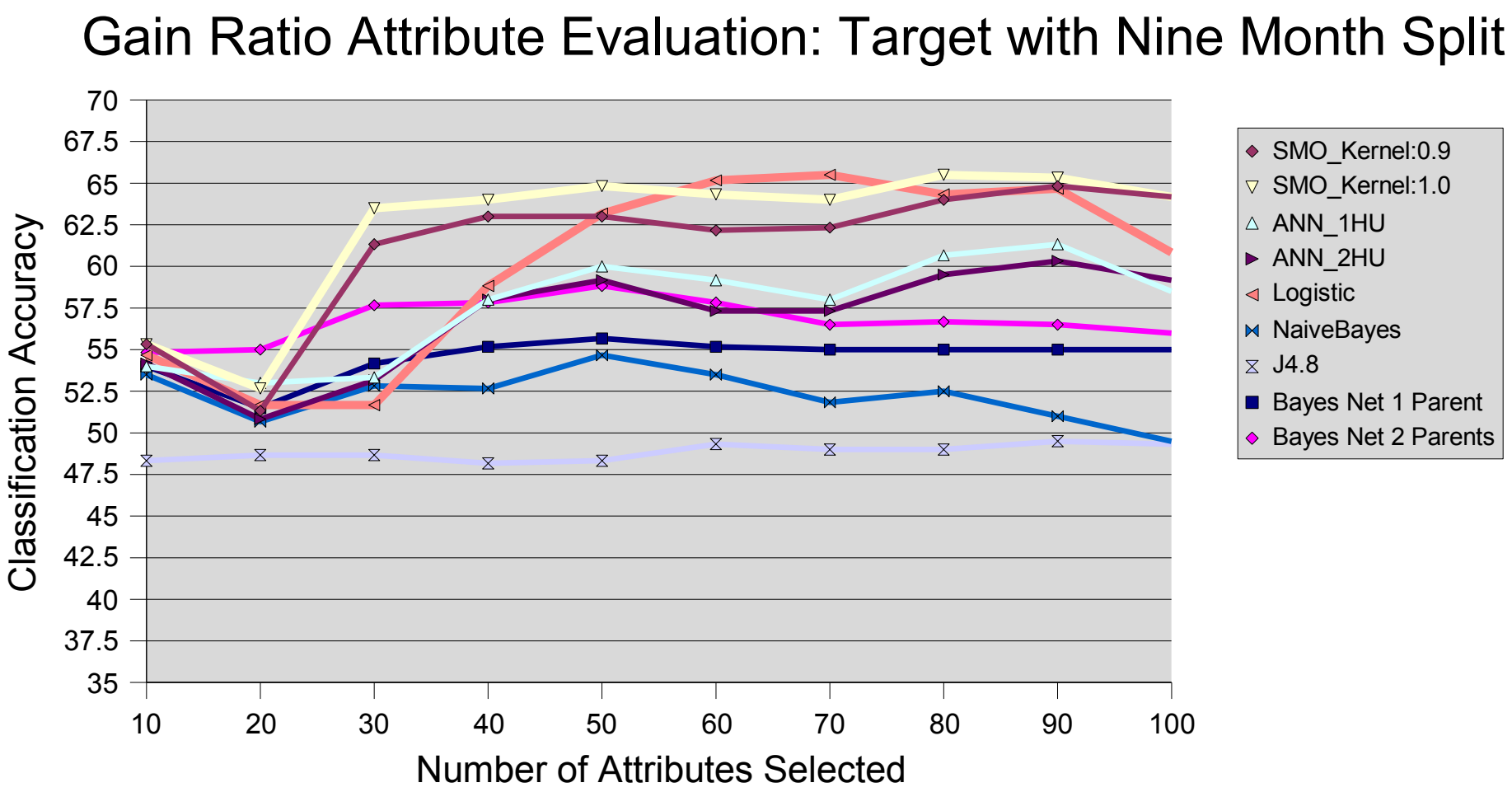


Feature Selection: Attribute Selected Classifier

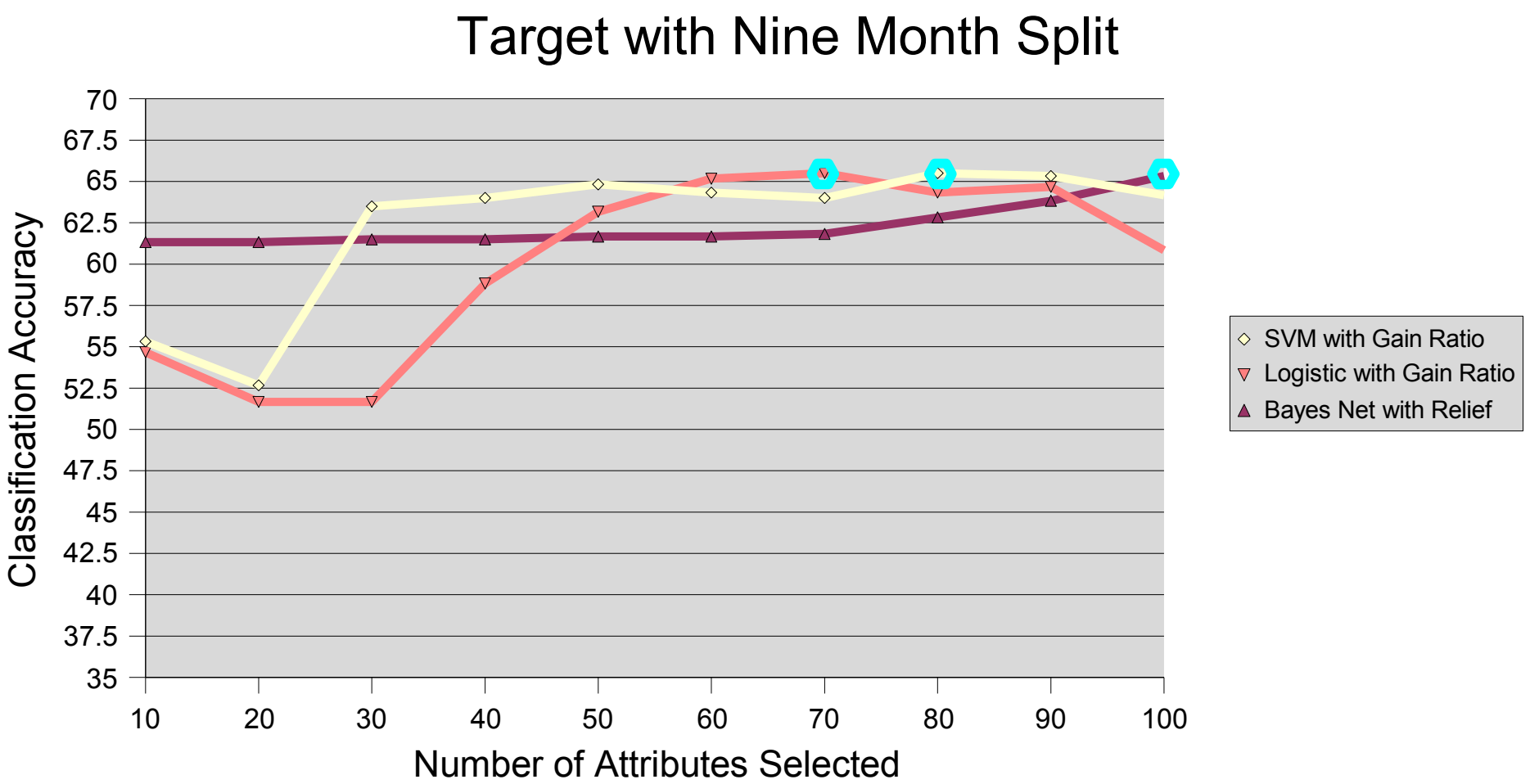
ASC(MLalgorithm, Feature Selection, Training Set: 200 Attributes , Test Set 200 Attributes)



Example Result Set



Best Classifiers



Best Classifiers

- 65.5%: Logistic Regression
  - Feature Selection: Gain Ratio to select 70 attributes
- 65.5%: Support Vector Machines, linear kernel
  - Feature Selection: Gain Ratio to select 80 attributes
- 65.3%: Bayesian Network, Two Parents
  - Feature Selection: ReliefF to select 100 attributes

Compare Classifiers

- 50.0% Majority Class
- 58.8% Logistic Regression, no feature selection
  - No Statistically Significant Difference
- 65.5%: Logistic Regression, feature selection
  - Statistically Significantly Better than Majority Class
- 65.5%: Support Vector Machines, linear kernel
  - Statistically Significantly Better than Majority Class
- 65.3%: Bayesian Network, Two Parents
  - Statistically Significantly Better than Majority Class

All Statistical Significance: $p<0.05$

Comparison of Feature Selection with Human Expert

Nine Month Split	Giles	Gain Ratio	ReliefF	SVM
ZeroR	50	50	50	50
Logistic	49	49	58	40
SMO_Kernel:0.9	52	53	58	39
ANN_1HU	62	54	60	38
NaiveBayes	52	60	55	51
J4.8	51	50	56	49
Bayes Net 1 Parent	62	57	59	48

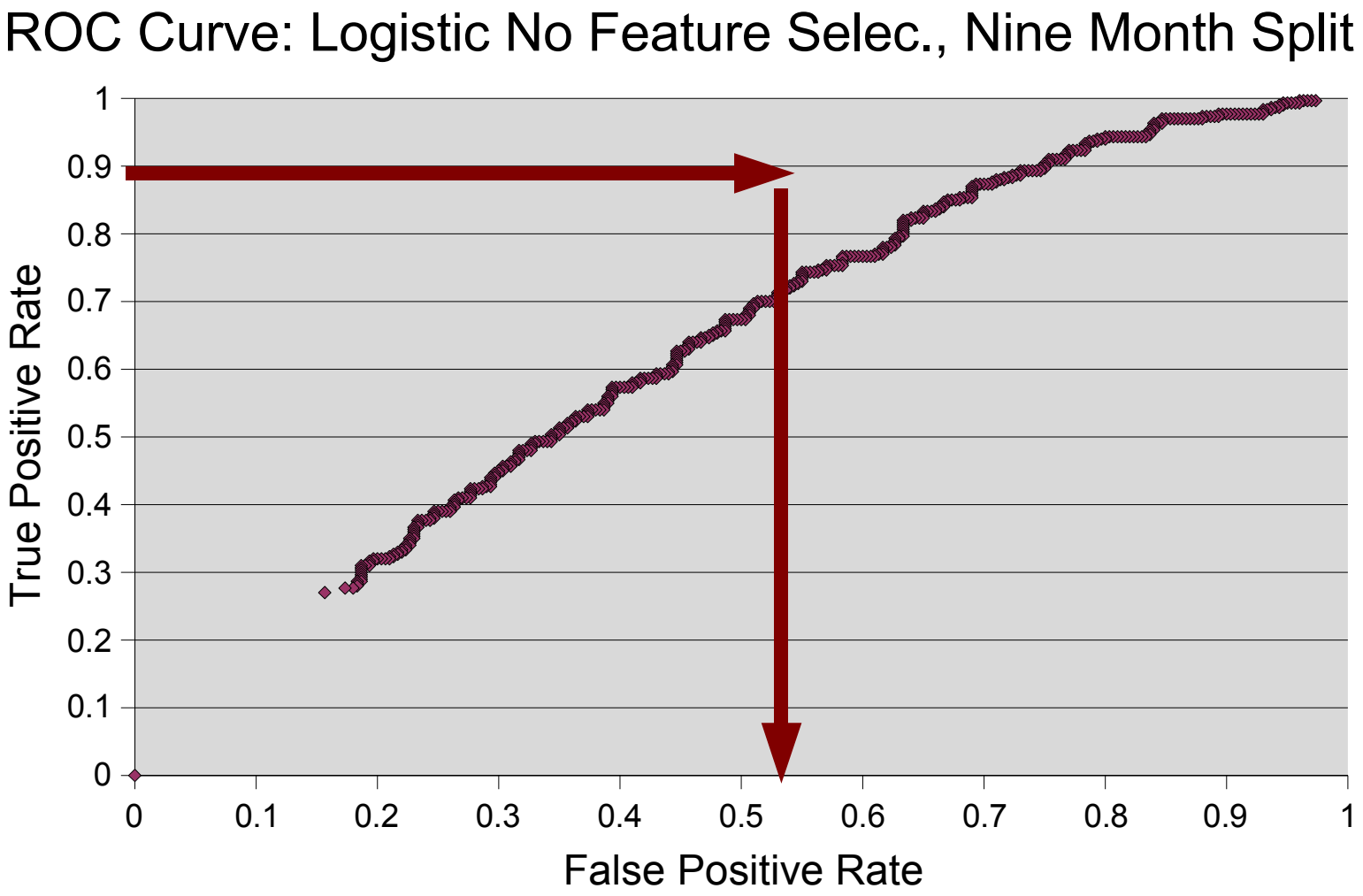
Six and Twelve Month Split	Giles	Gain Ratio	ReliefF	SVM
ZeroR	33	33	33	33
Logistic	35	49	47	22
SMO_Kernel:0.9	37	46	46	27
ANN_1HU	38	46	48	27
NaiveBayes	35	47	46	35
J4.8	38	44	39	33
Bayes Net 1 Parent	34	48	42	29

Comparison of Attributes Selected

Giles		Relief	
DemECOG	TxChemo	TxLap	SxFati
SxWtloss	TxPal	SxSatiety	DemWeight
SxChola	TxPalCeliac	ResPODischStatus	RadOncName
SxAbd	ResTransfusion	SxPru	ResBloodLoss
SxBack	ResAttemptUn	SxBack	SxWtloss
CxHF	ResPODays	TxResect	ResPOPulmComp
CxLiver	ResPOLeak	SurOncName	TxChemoGem
LabAlb	ResPOLiverTB	PTCStent	TxChemoFlu
CTSMA	ResPathM	PTCDx	EUSStagingT
CTHepatic	NoResNoHandle	ResPOInfection	TxRadia
CTPortal	NoResMagnitude	GIMDName	TxPal
EUSCyto	NoResCeliacInvolve	Histology	TxChemo
Histology	NoResSMAInvolve	DemECOG	LabCEA
PreOutlook	NoResCirrhosis	CxDiab	CxDiabOral
TxResect	NoResMetastatic	CxPriorCancerSurgery	PresumptiveDx

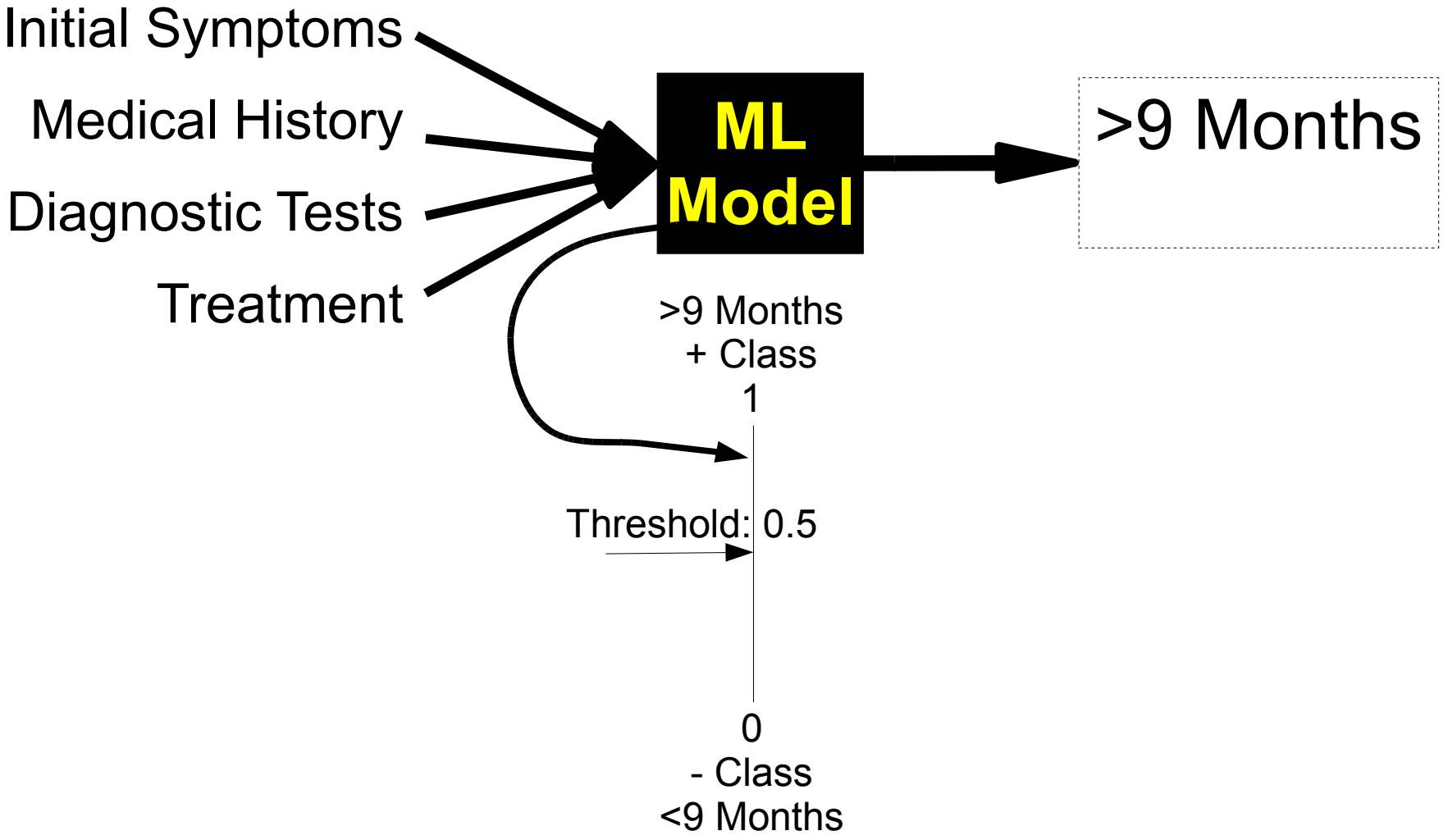
Evaluation Techniques: ROC Curves

- Show trade off between true and false positive rates

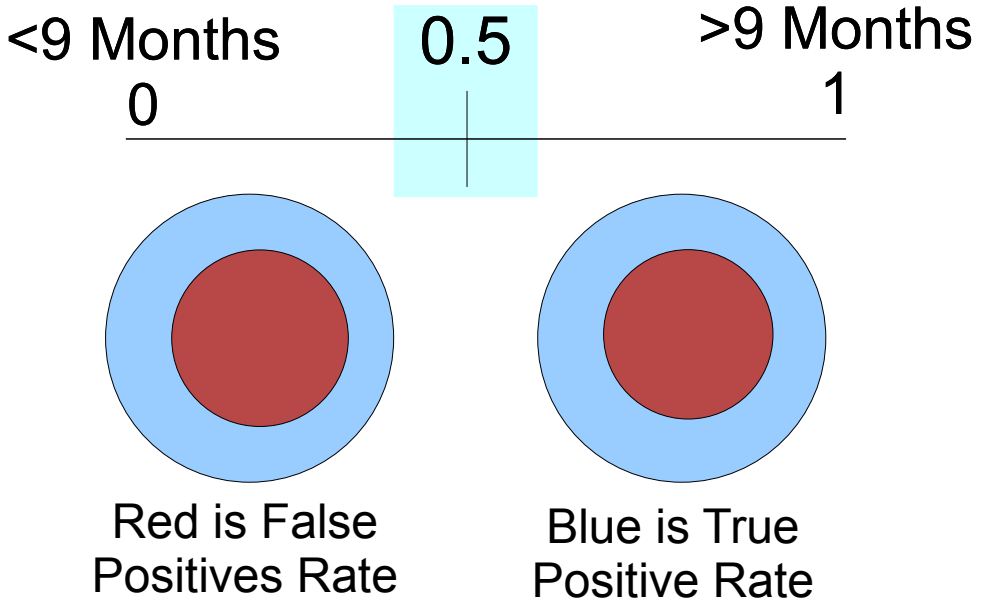


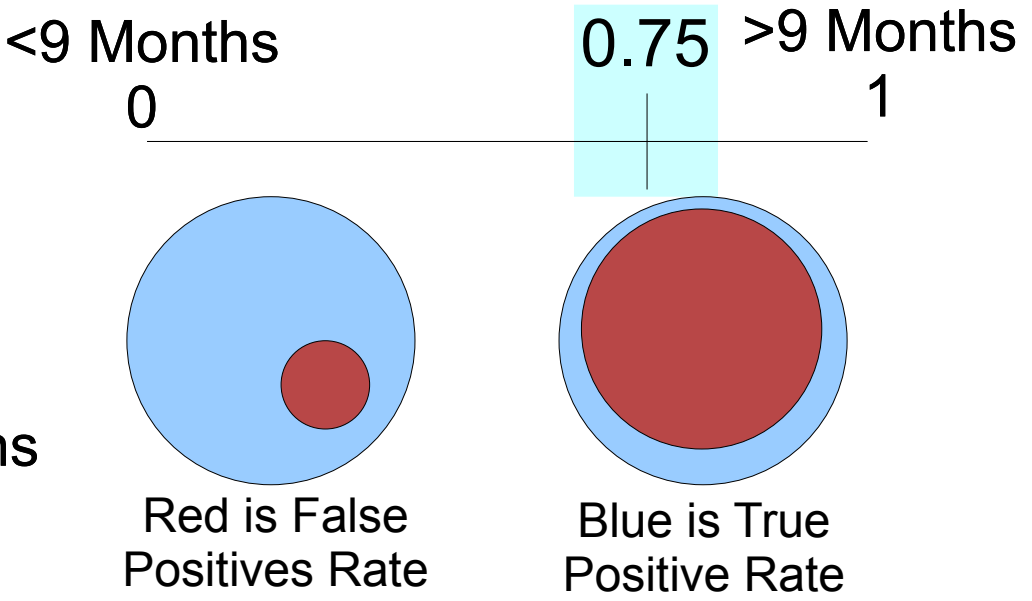
Evaluation Techniques: ROC Curves

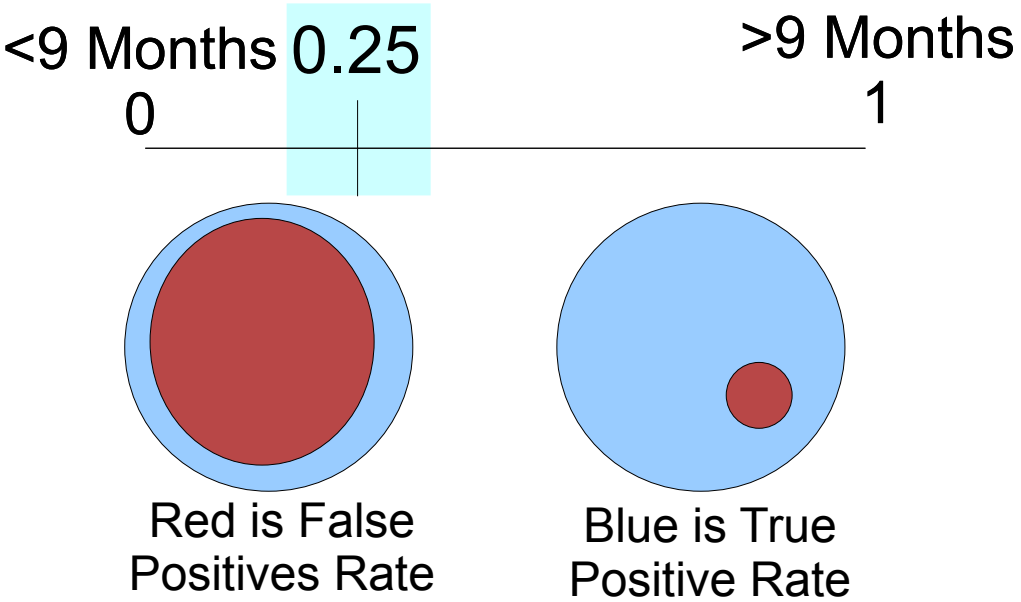
- Within most ML Model is a threshold



Evaluation Techniques: ROC Curves

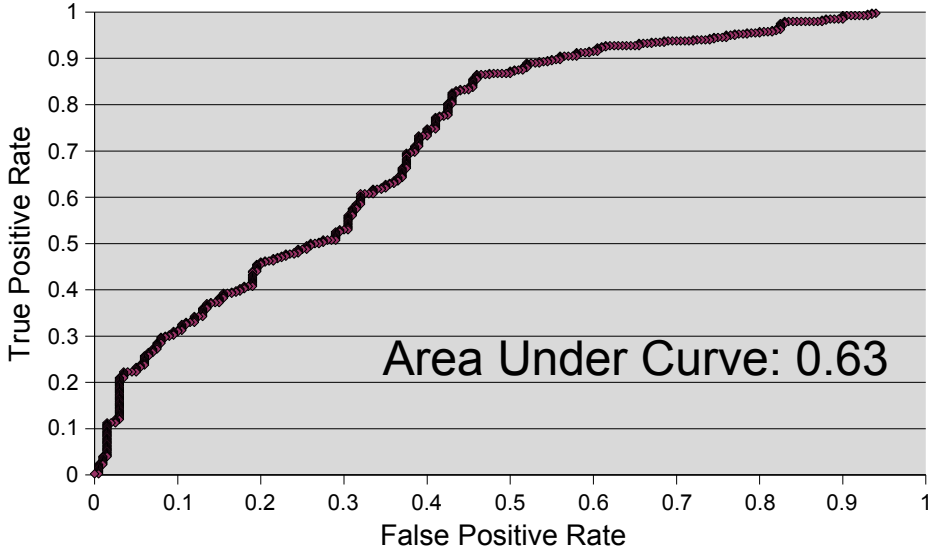




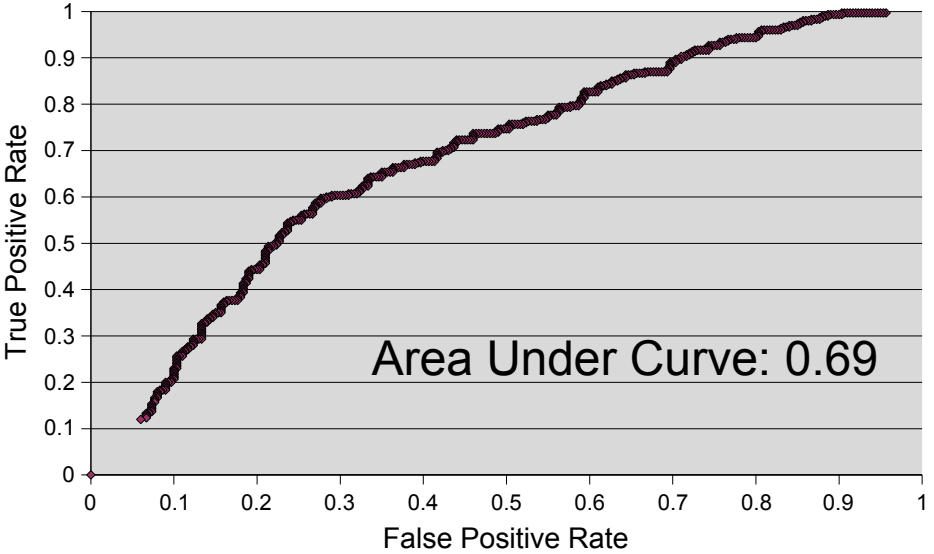


ROC Curves

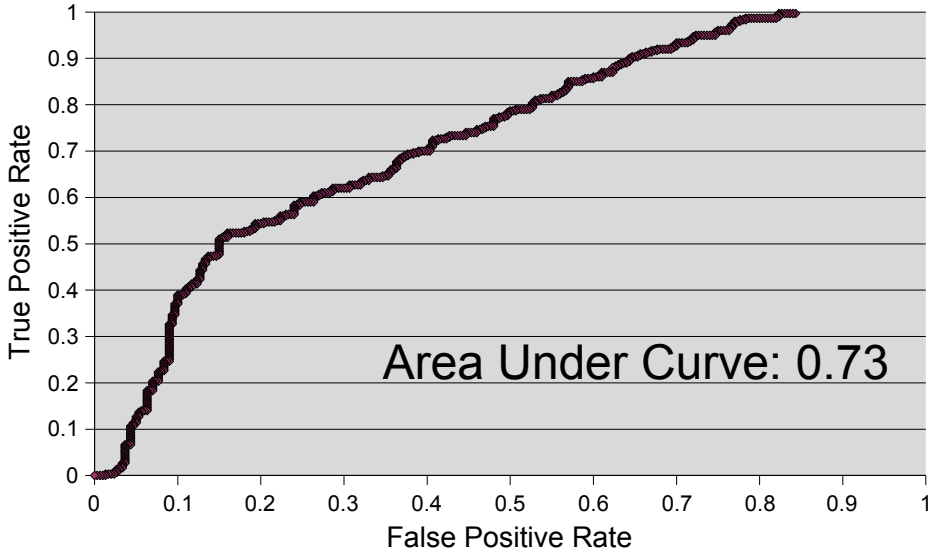
Example ROC Curve



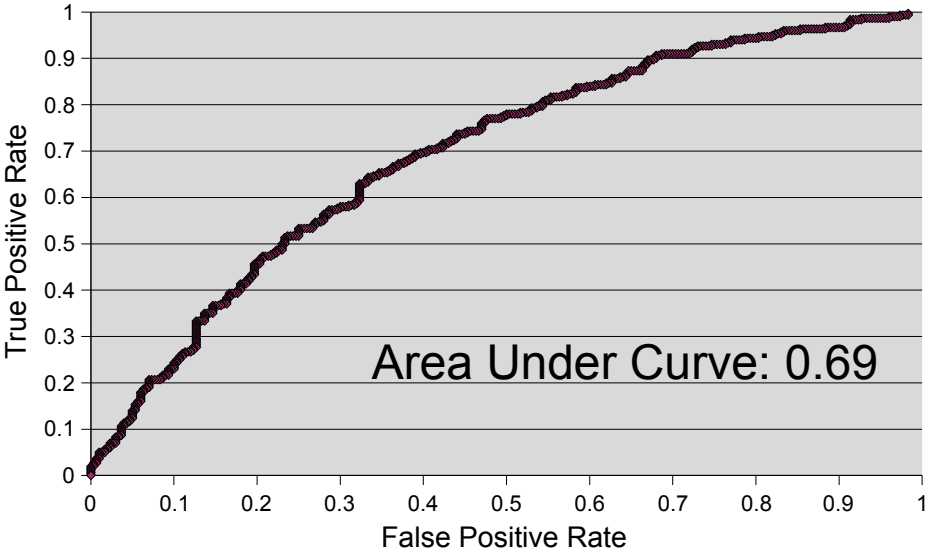
ROC Curve: Logistic with Feature Selec., Nine Month Split



ROC Curve: Support Vector Machines, Nine Month Split

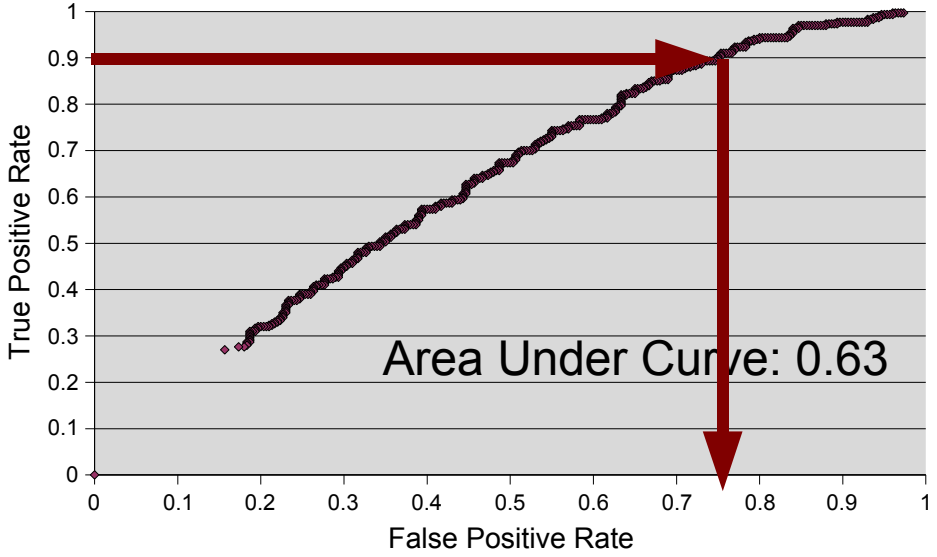


ROC Curve: Bayesian Network, Nine Month Split

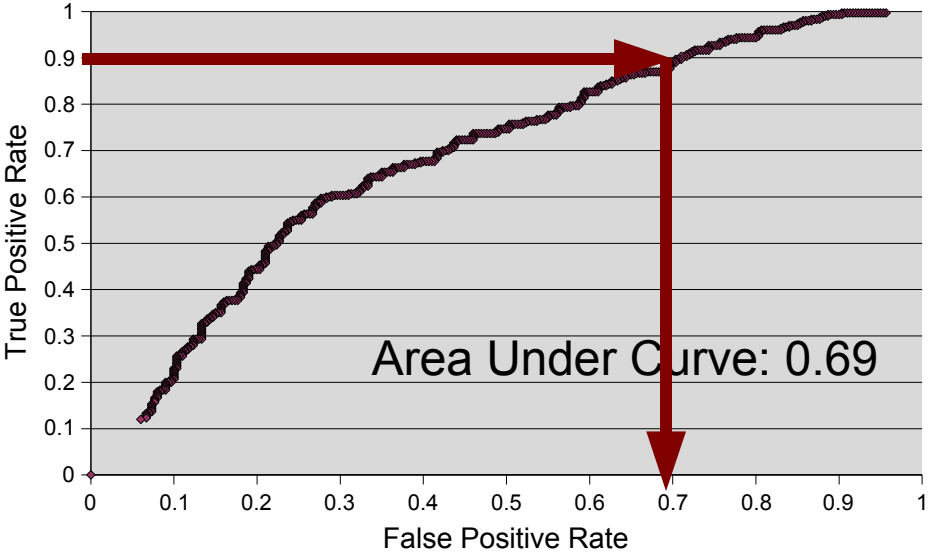


ROC Curves

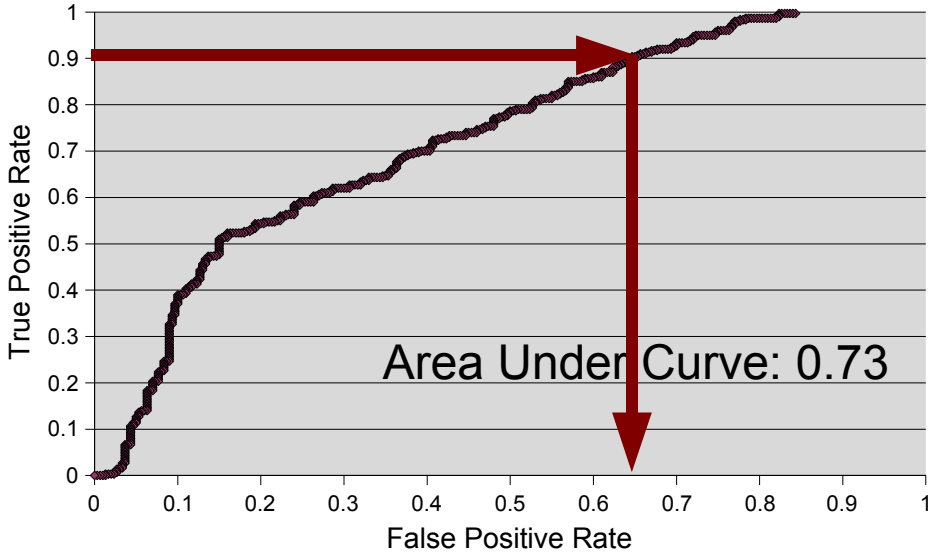
ROC Curve: Logistic No Feature Selec., Nine Month Split



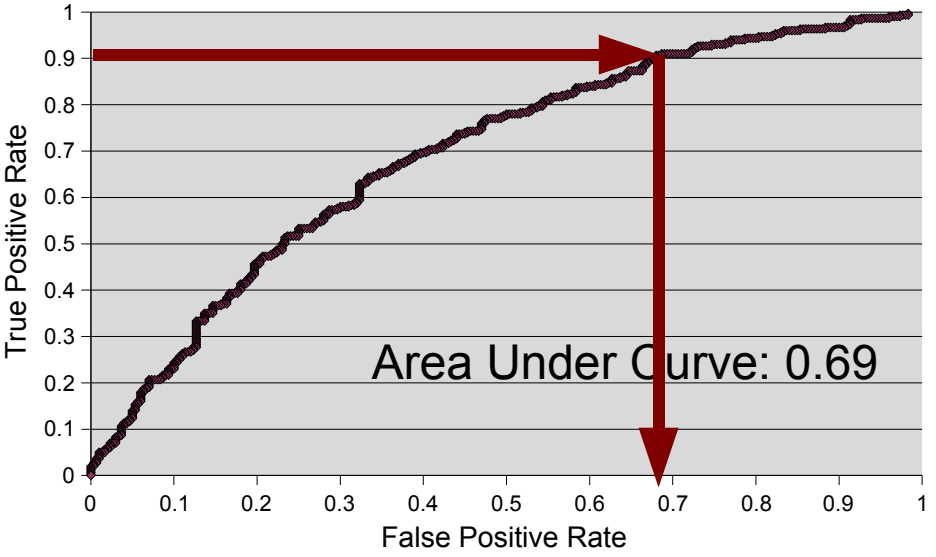
ROC Curve: Logistic with Feature Selec., Nine Month Split



ROC Curve: Support Vector Machines, Nine Month Split

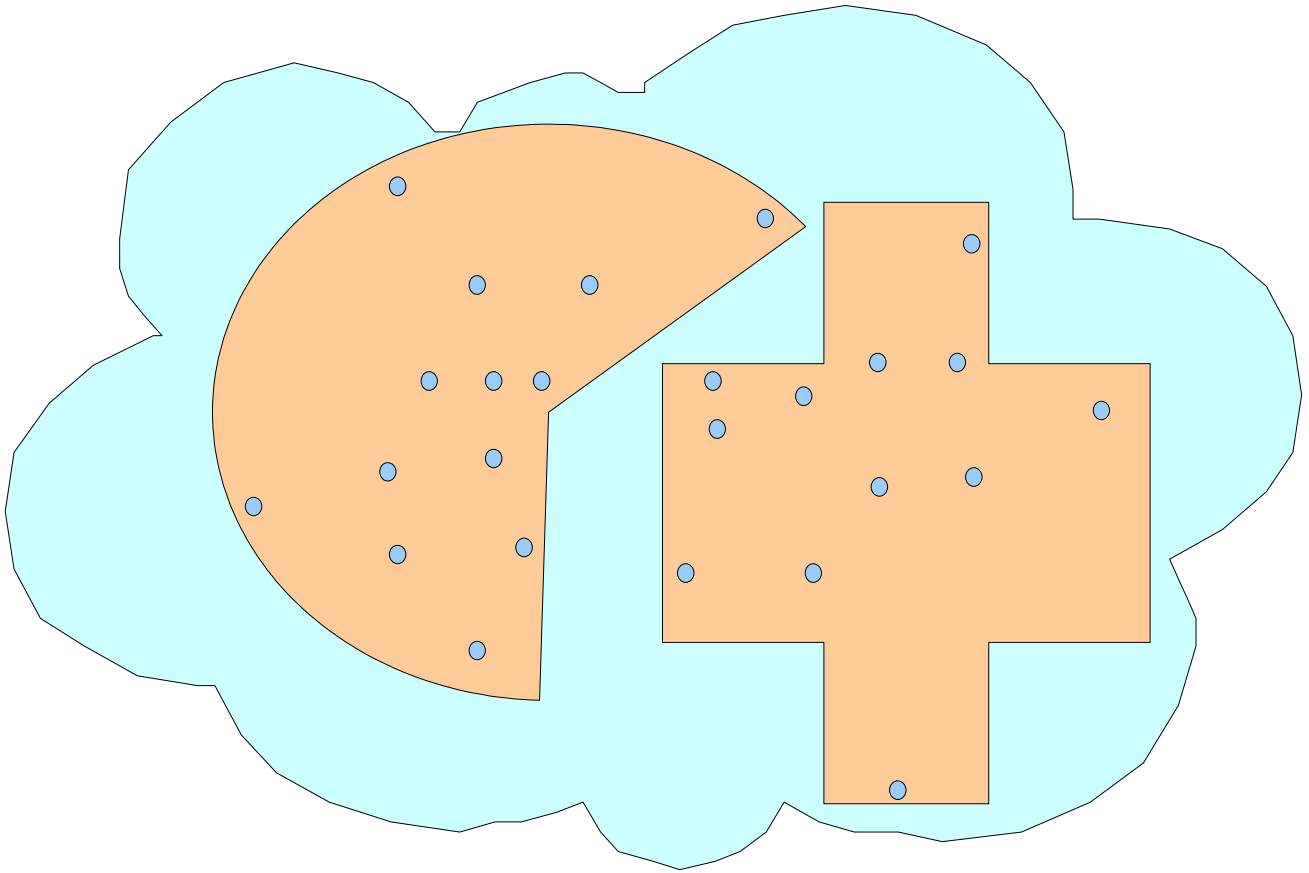


ROC Curve: Bayesian Network, Nine Month Split



Machine Learning Algorithms: Meta Learning

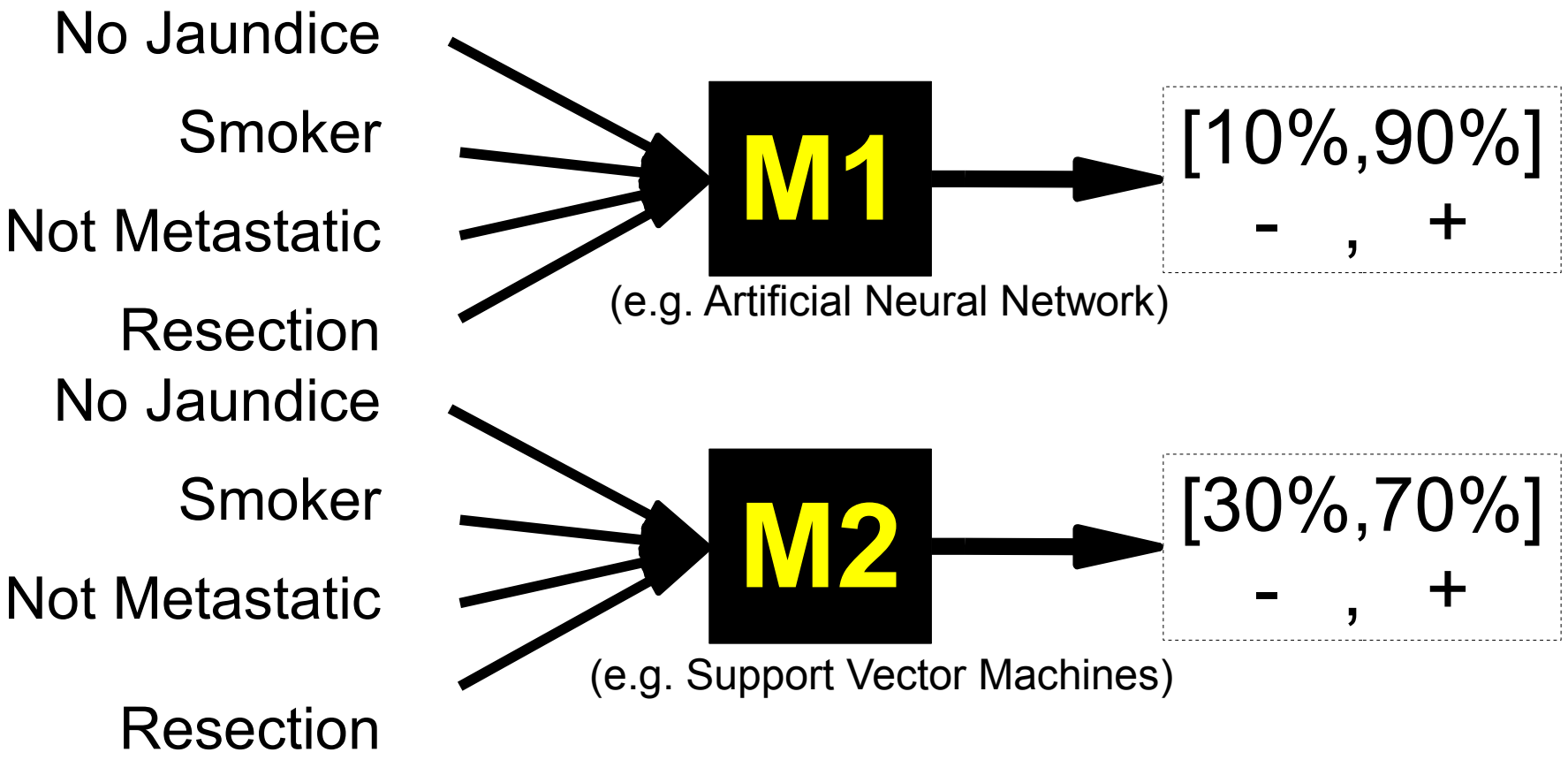
- Can we construct a model to predict which of the original models will best classify a test instance?



Machine Learning Algorithms: Our Model Selector - Example

Original Instance:

{No Jaundice, Smoker, Not Metastatic, Resection, +}

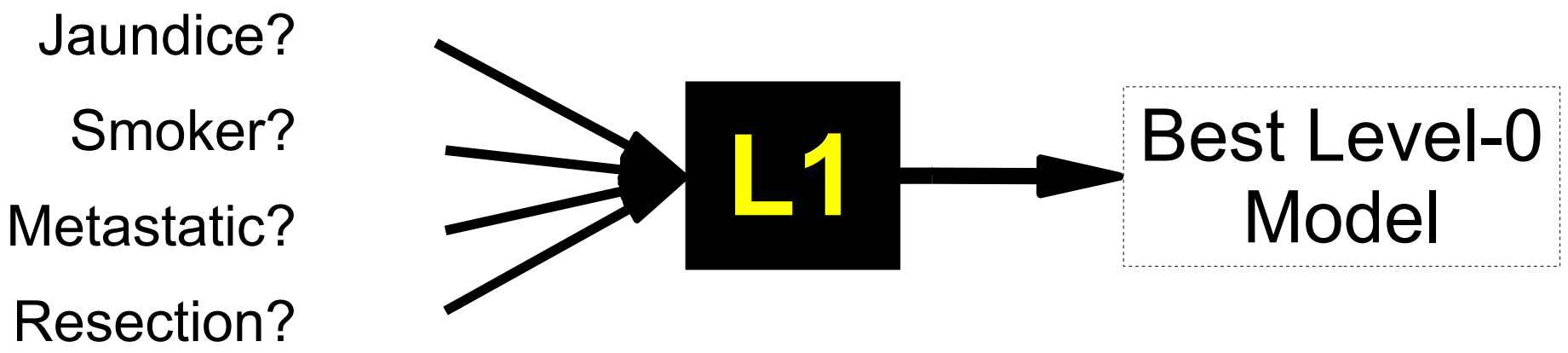


New Instance:

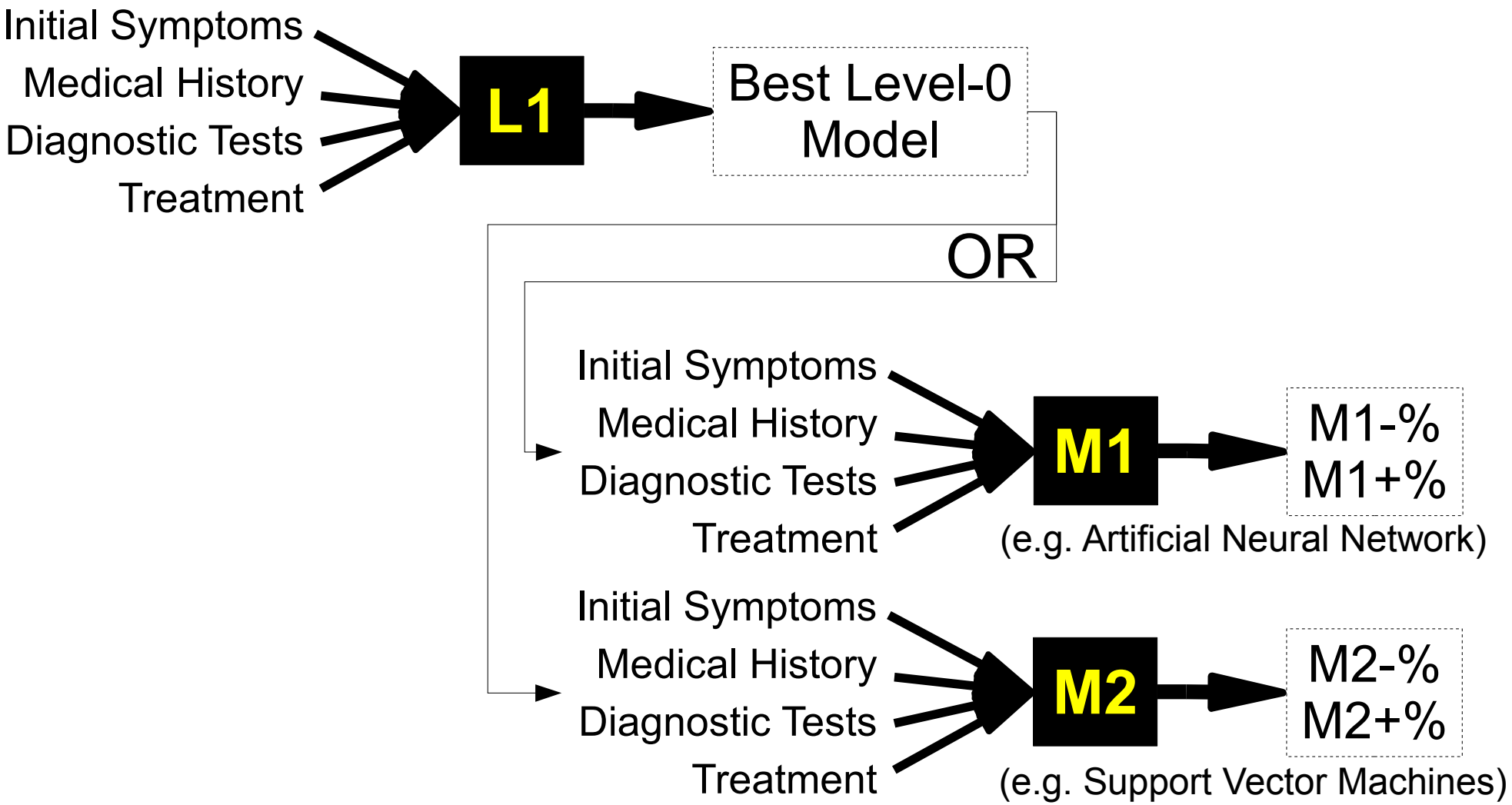
{No Jaundice, Smoker, Not Metastatic, Resection, M1}

Machine Learning Algorithms: Our Model Selector - Training

{Jaundice, Smoker, Metastatic, No Resection, M2}
{No Jaundice, Smoker, Not Metastatic, Resection, M1}
{Jaundice, Not Smoker, Metastatic, Resection, M2}



Machine Learning Algorithms: Our Model Selector – Test Instance



Our Model Selector: Compare Classifiers

Meta Model	Models Combined			
	Logistic, SVM, Bayes	SVM, Bayes	Logistic, Bayes	Logistic, SVM
ANN 1 Hidden Unit	64.6	65.2	65.7	66.5
ANN 2 Hidden Units	64.3	65.2	65.8	66.2
J48	64.5	65.8	65.2	65.8
Naïve Bayes	64.8	64.2	62.8	67.3
SVM	65.3	66.2	65.0	65.7

- 65.5%: Logistic Regression, feature selection
- 65.5%: Support Vector Machines, linear kernel
- 65.3%: Bayesian Network, Two Parents

Thesis Contribution

- Investigation into variety of Feature Selection techniques
- Use of Attribute Selected Classifier to evaluate Feature Selection
- Support Vector Machines over this domain
- Designed and Implemented new Meta-Learning Algorithm: Model Selector

General Conclusions

- Logistic Regression can be improved through feature selection
- Over datasets where no statistical difference between logistic regression and majority class, statistical difference using more advanced techniques
- Would be good to have more patients to reduce model variance

Feature Selection Conclusions

- Feature Selection helpful for selecting most important attributes in estimating survival
- Attribute selection, when run over both training and test set biased
  - Bias removed through use of Attribute Selected Classifier
- Gain Ratio best feature selector over this domain
- Accuracy of feature selection very close to, and often better then, selections of domain expert

Meta-Learning Conclusions

- Stacking, Bagging, and Boosting not useful over this domain
- Our Model Selector can increase classification accuracy by a few percent without increasing standard deviation

Future Work

- Investigate Quality of Life in addition to survival
- Compare findings to national databases
  - Much larger number of patients
  - Much less information on each patient
- See that UMass continues to add new patients to the database

Data Mining Techniques for Prognosis in Pancreatic Cancer

Masters Thesis Presentation

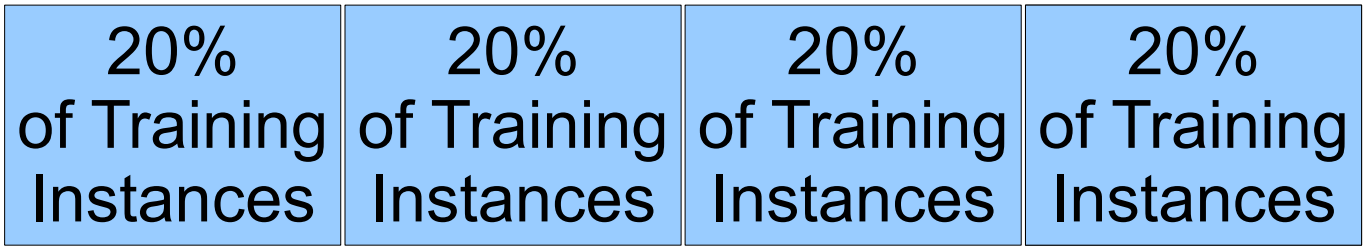
Stuart Floyd
AIRG
April 26, 2007

Advisors: Professor Carolina Ruiz,
Professor Sergio Alvarez (Boston College)

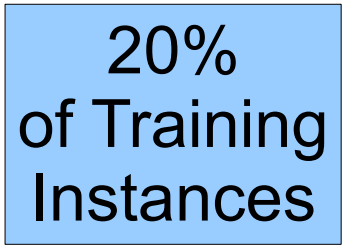
UMass Collaborators: Professor Jennifer Tseng,
Professor Giles Whalen

Cross Validation

Used For Training:

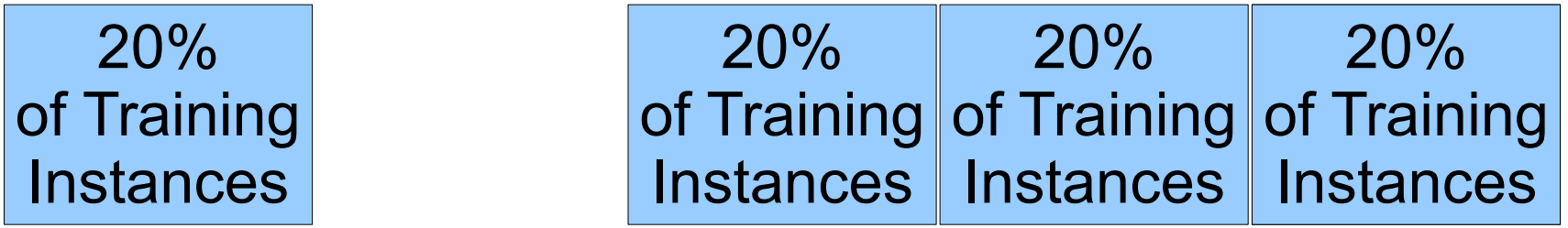


Used For Testing:

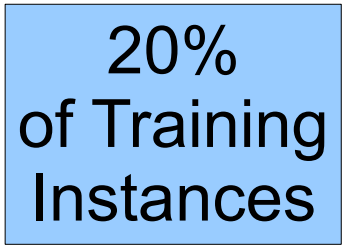


Cross Validation

Used For Training:



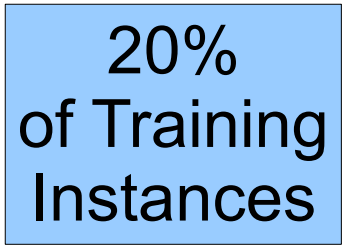
Used For Testing:



Cross Validation

Used For Training:

Used For Testing:



Cross Validation

Used For Training:

20% of Training Instances	20% of Training Instances	20% of Training Instances		20% of Training Instances

Used For Testing:

				20% of Training Instances

Cross Validation

Used For Training:

20% of Training Instances	20% of Training Instances	20% of Training Instances	20% of Training Instances

Used For Testing:

20% of Training Instances