

Classification with Gene Expression Data

BMI/CS 576

www.biostat.wisc.edu/bmi576/

Colin Dewey

cdewey@biostat.wisc.edu

Fall 2008

Classification of Gene Expression Profiles: The *Learning* Task

given:

- expression profiles for a set of genes or experiments/ individuals/time points (whatever columns represent)
- a class label for each expression profile

do: learn a model that is able to accurately predict the class labels for other expression profiles

Classification of Gene Expression Profiles: The *Classification* Task

given:

- expression profiles
- a classification model

do: predict the associated class for each expression profile

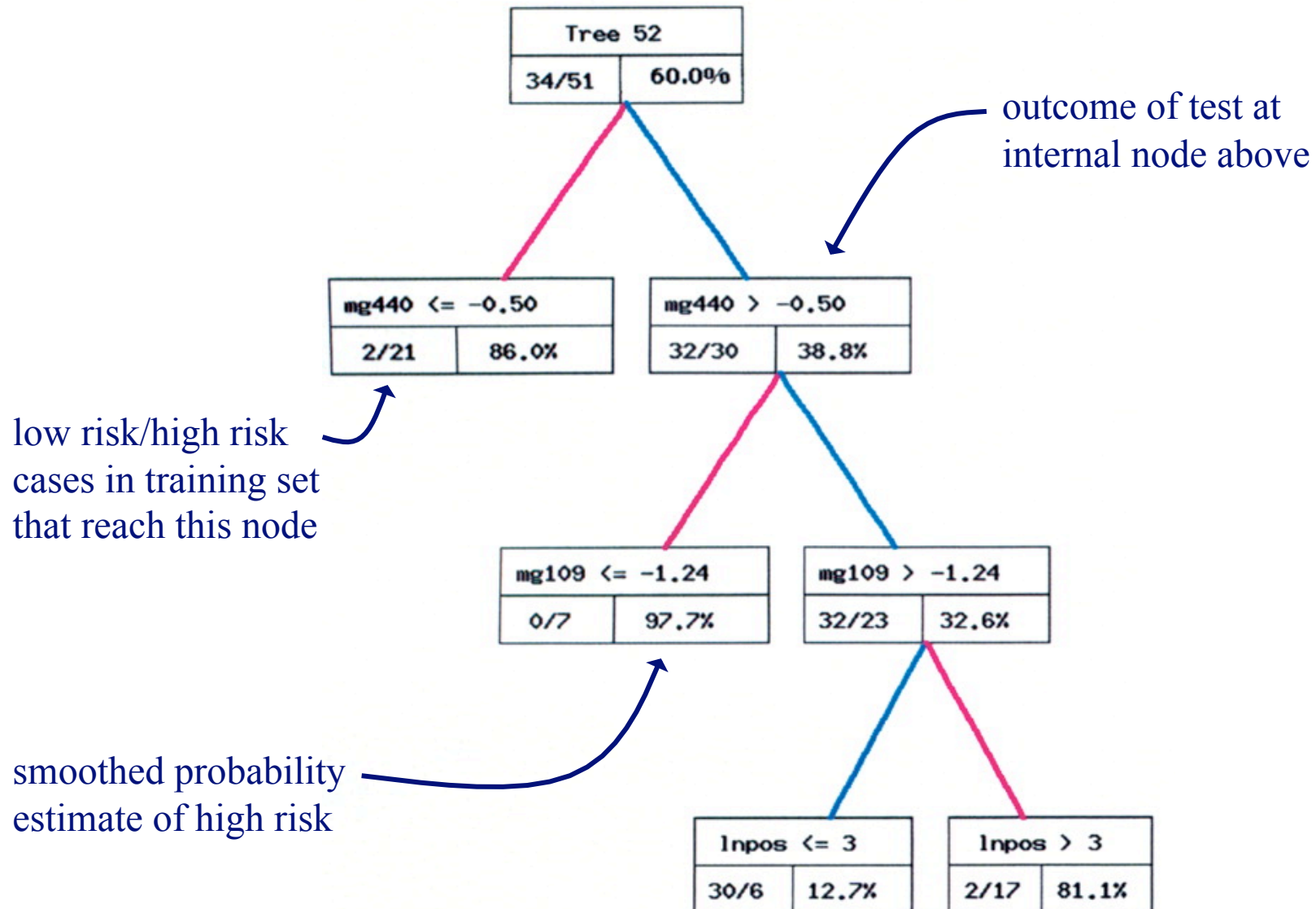
Breast Cancer Outcomes Prediction

- Nevins et al., *Lancet 2003*, *Human Molecular Genetics 2003*
- microarray and clinical data from 86 lymph-node positive breast cancer patients
 - 12,625 genes measured using Affymetrix arrays
- goal is to distinguish between high risk (recurrence w/in 5 years) and low risk (recurrence-free for 5 years)

Calculating “Metagenes”

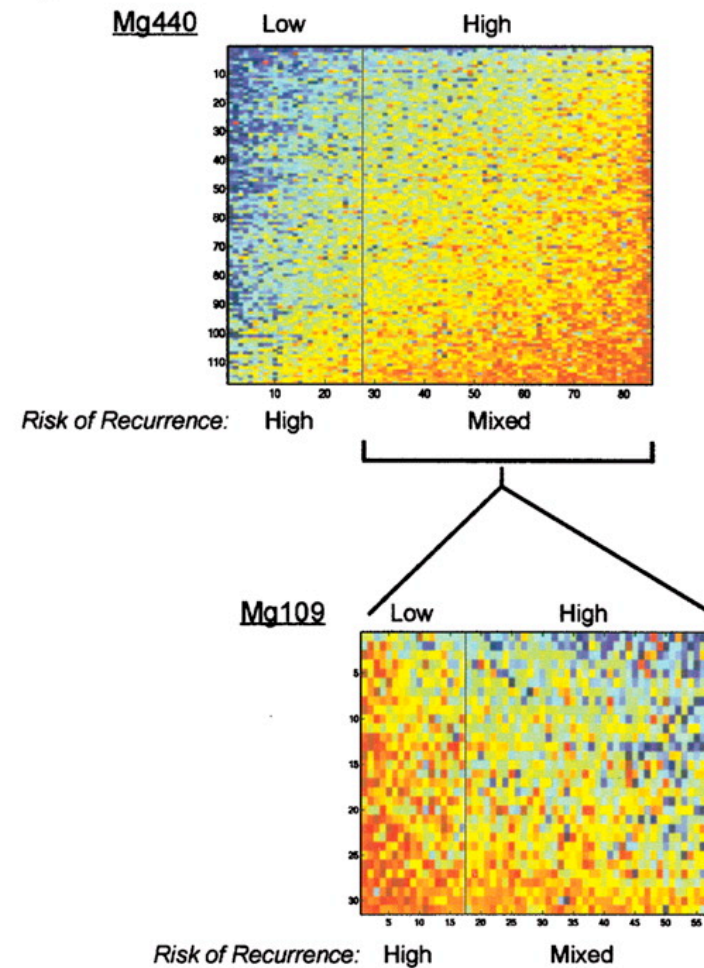
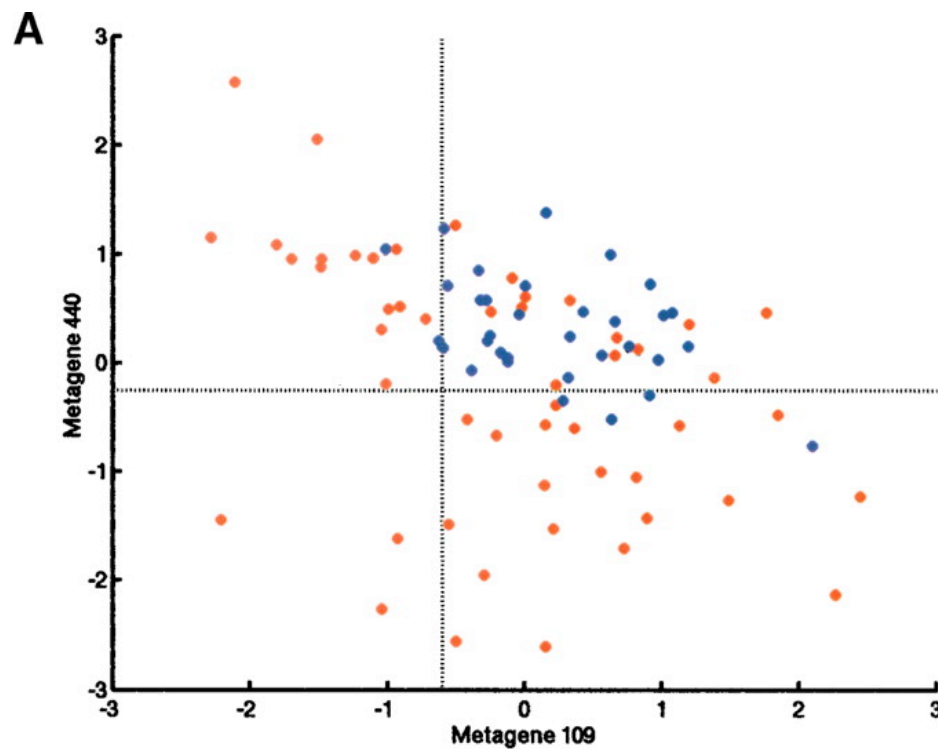
- the features used in their model are not mRNA measurements from individual genes
- instead they compute “metagenes”, which consist of linear combinations of gene measurements
- procedure
 - ran k -means clustering (with $k=500$) on original microarray data set
 - computed first *principal component* of each cluster
 - each of these principal components becomes a metagene

A Decision Tree Classifier



Decision Tree Classifiers

- tree-based classifiers partition the data using axis-parallel splits



Inducing Tree-Based Classifiers

- there are many decision-tree learning methods
- two most common are
 - C4.5 (Quinlan)
 - CART (Breiman, Friedman, Olshen, Stone)
- Nevins et al. use their own method
- all DT learning methods have the same basic algorithm structure: recursively grow a tree top-down

Generic DT Induction Pseudocode

MakeSubtree(set of instances I)

if stopping criteria met

make a leaf node N

determine class label/probabilities for N

else

make an internal node N

select best splitting criterion for N

for each outcome k of the split

I_k = subset of instances that have outcome k

k th child of N = MakeSubtree(I_k)

return subtree rooted at N

Breast Cancer Outcomes Prediction

- predictive accuracy estimated by cross-validation
85-90% correct predictions (low-risk, high-risk)

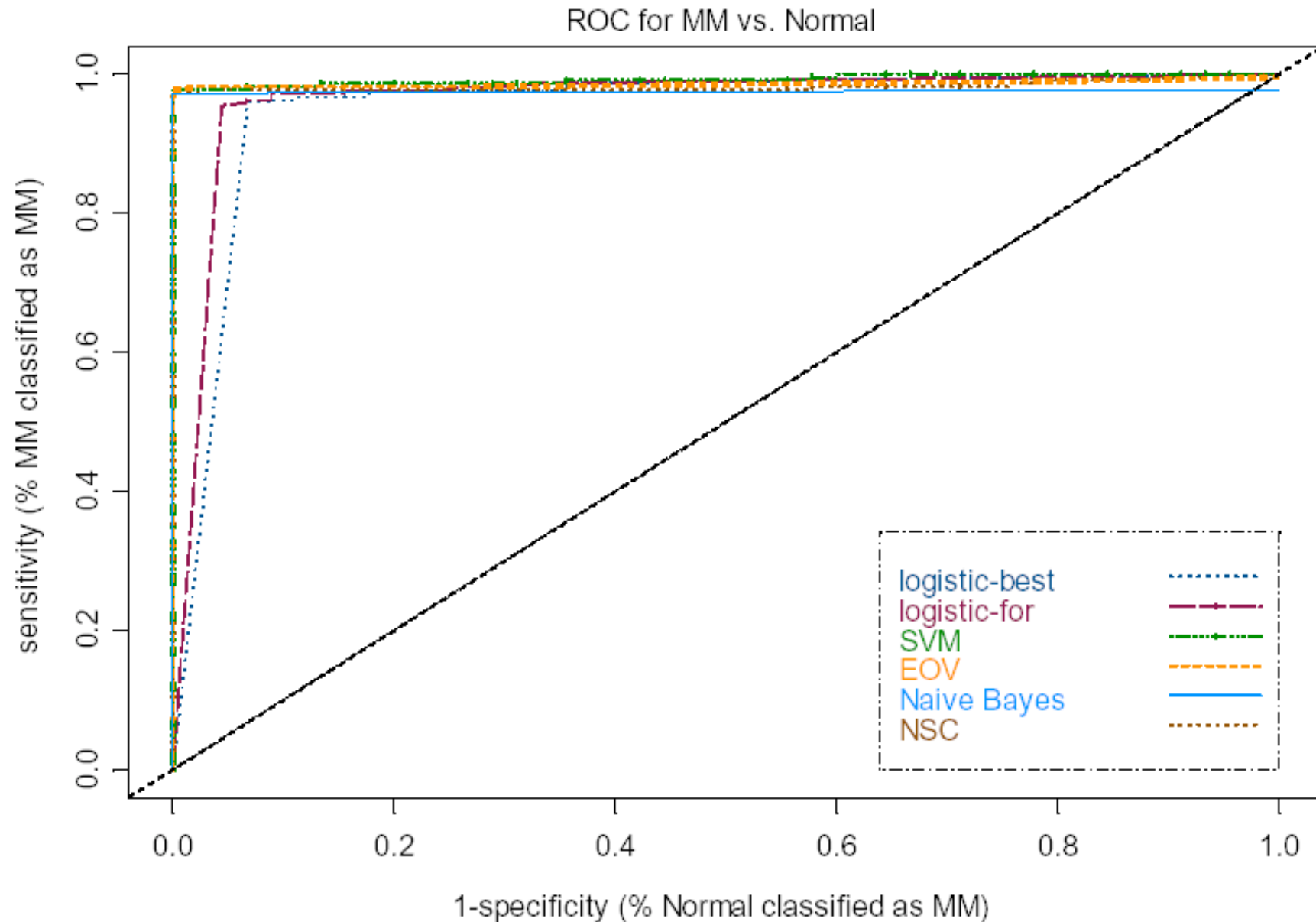
Multiple Myeloma vs. MGUS Classification

- Hardin et al., *Statistical Applications in Genetics and Mol. Bio.* '04
- standard lab classification of MM and MGUS is quite accurate... but biologically uninformative
- MGUS = Monoclonal gammopathy of undetermined significance
- MM = multiple myeloma
- Can this classification be done using expression profiles?
- learned models might
 - enable molecular diagnosis
 - lend insight into disease progression

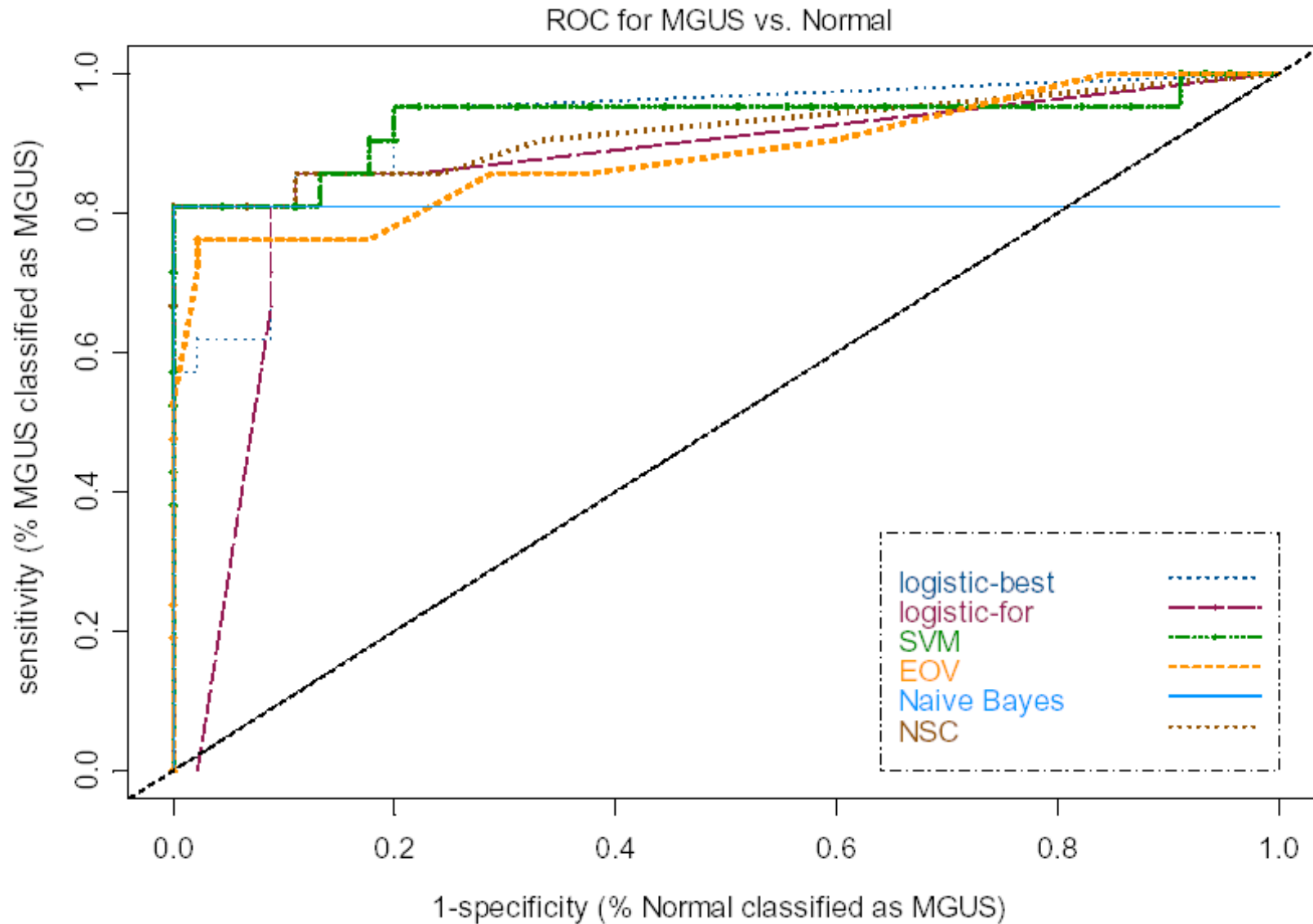
Hardin et al. Empirical Evaluation

- applied six supervised machine learning algorithms to this data set
 - logistic regression
 - C5.0 decision tree induction method
 - ensembles of voters
 - naïve Bayes
 - nearest shrunken centroid
 - support vector machines

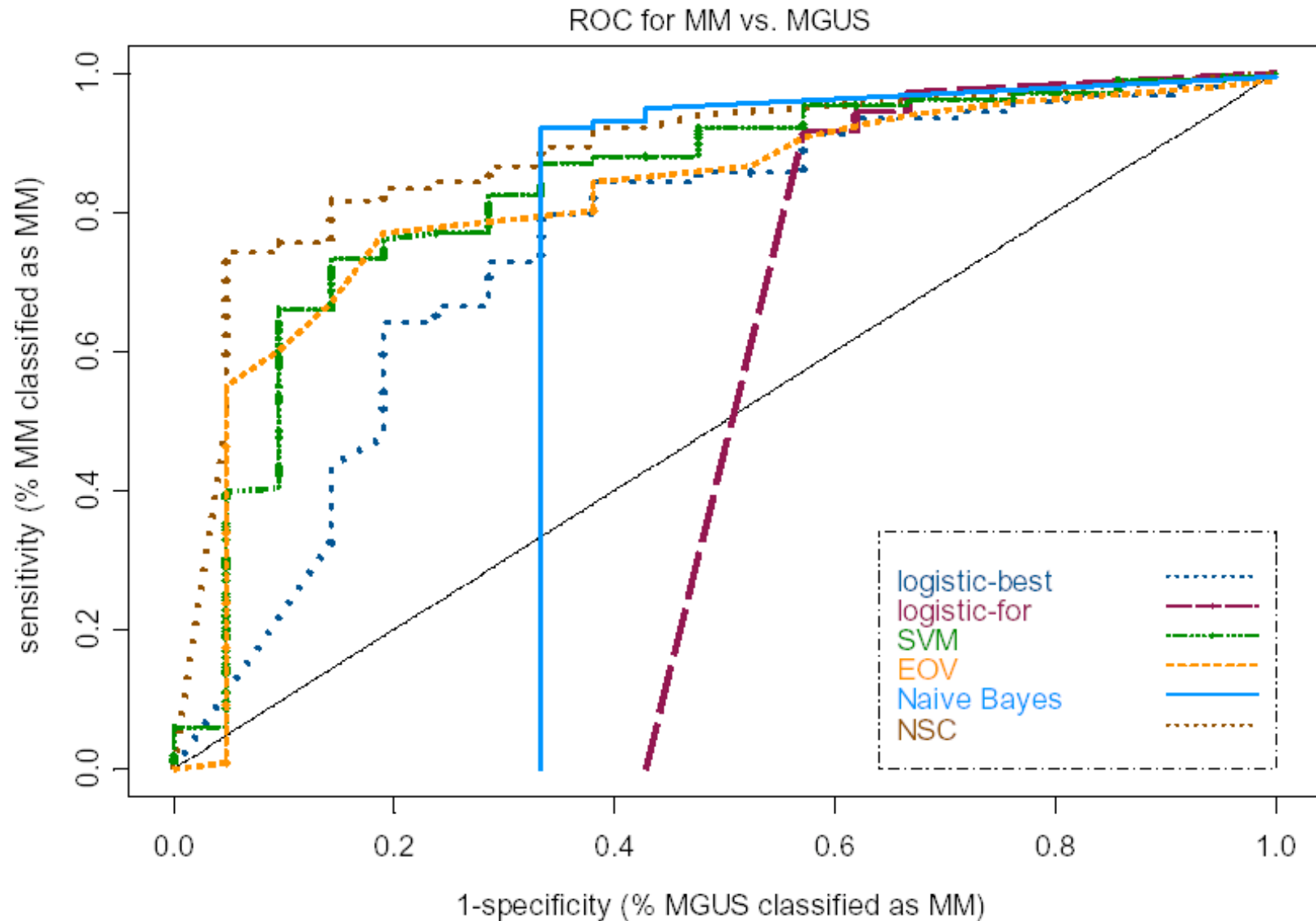
Empirical Evaluation:MM vs. Normal



Empirical Evaluation: MGUS vs. Normal



Empirical Evaluation: MM vs. MGUS



Comments Gene Expression Analysis

- we discussed two computational tasks
 - classification*: do this when you do know the categories of interest
 - clustering*: do this when you don't know the categories of interest
- *class discovery* is an interesting task that falls between classification and clustering
 - identify classes of profiles that don't seem to fit into any of the modeled categories
 - e.g. new subtypes of cancer, new types of toxic substances
- we've discussed methods in the context of microarray data, but they can be applied to a wide variety of high-throughput biological data