

Data Mining

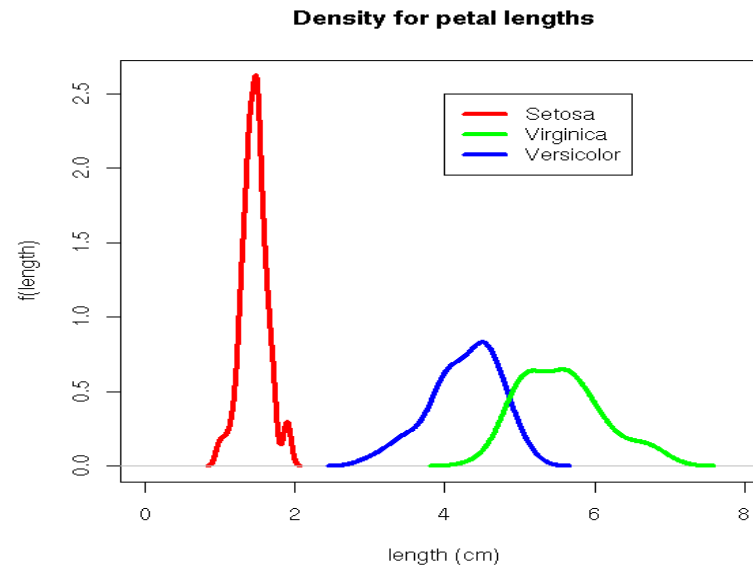
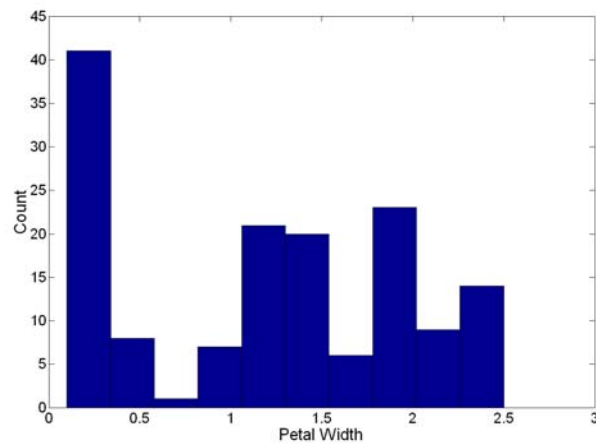
Classification: Alternative Techniques

Midterm review

Edited by J. Taylor for STATS202, Stanford University, Winter 2009

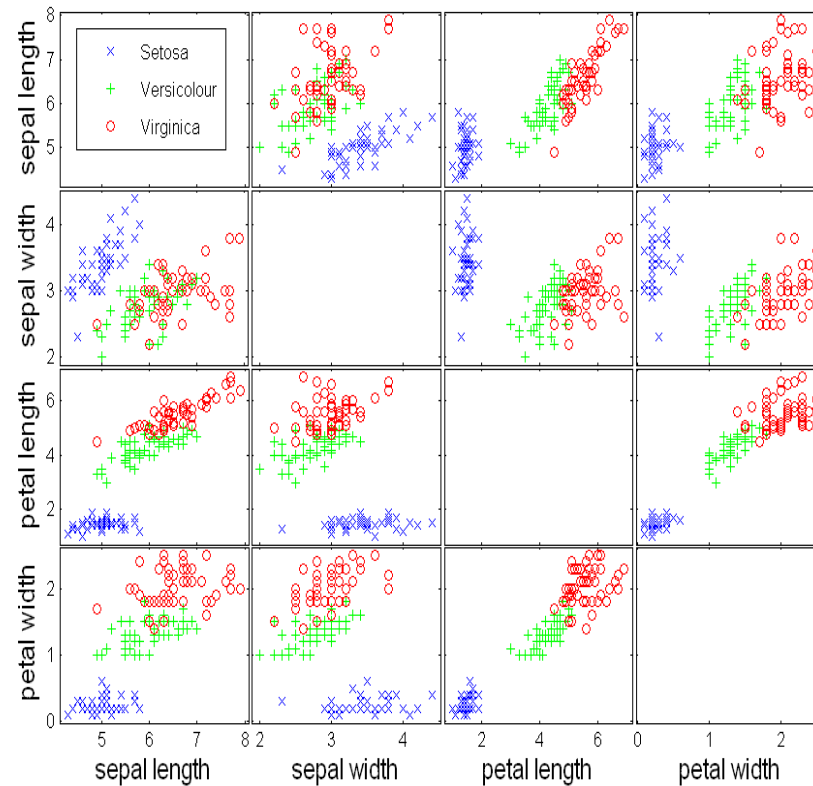
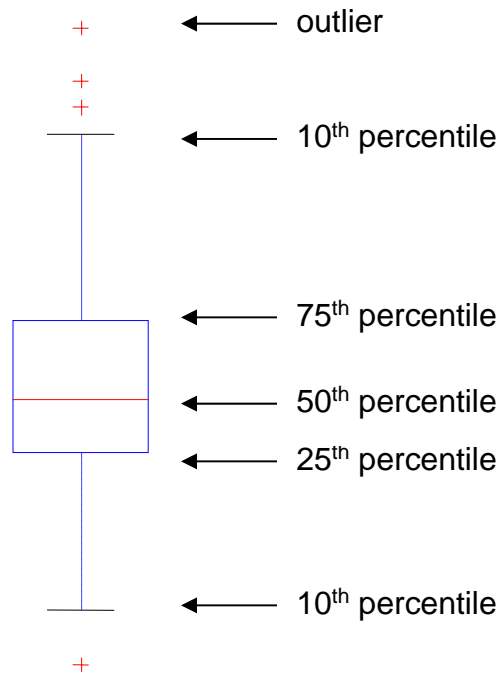
Chapter 3: Exploring Data

- Visualization tools



Chapter 3: Exploring Data

- Visualization tools

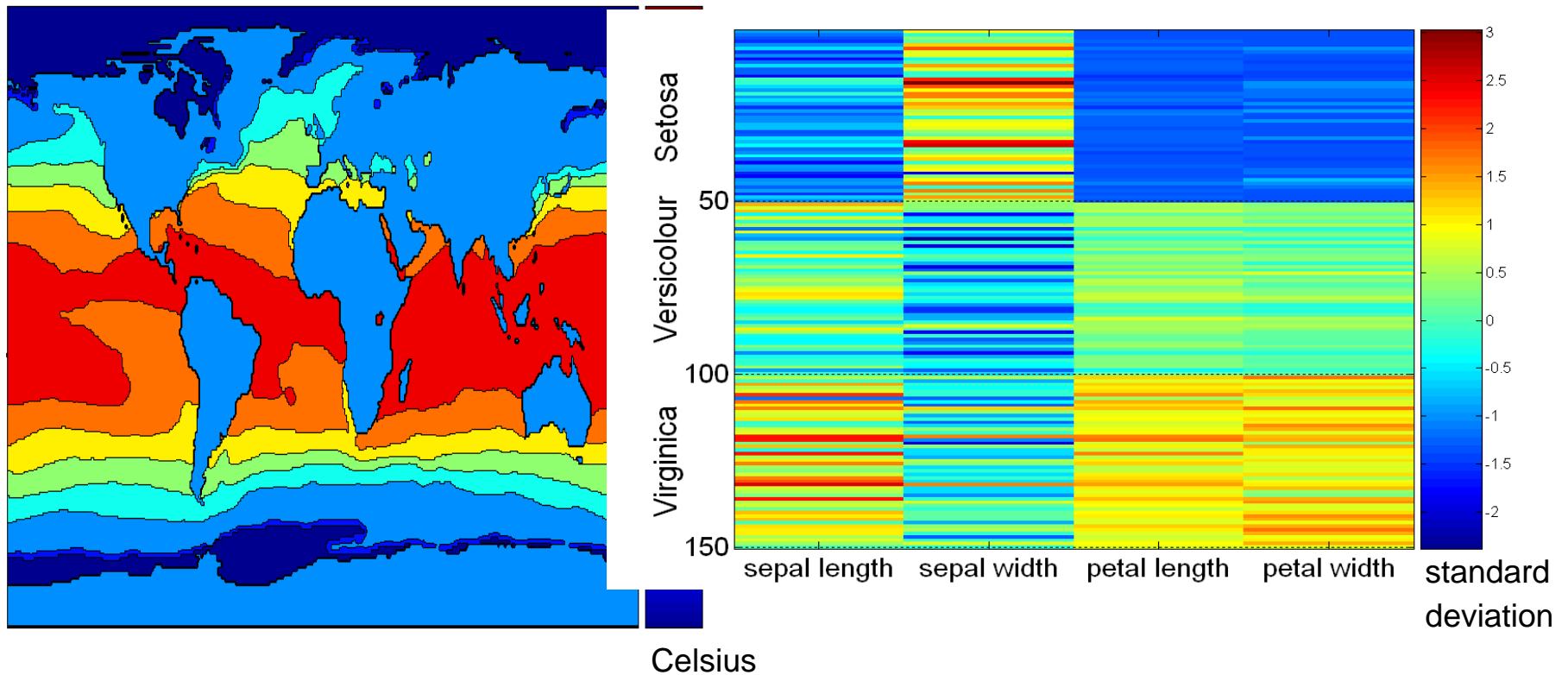


Chapter 3: Exploring Data

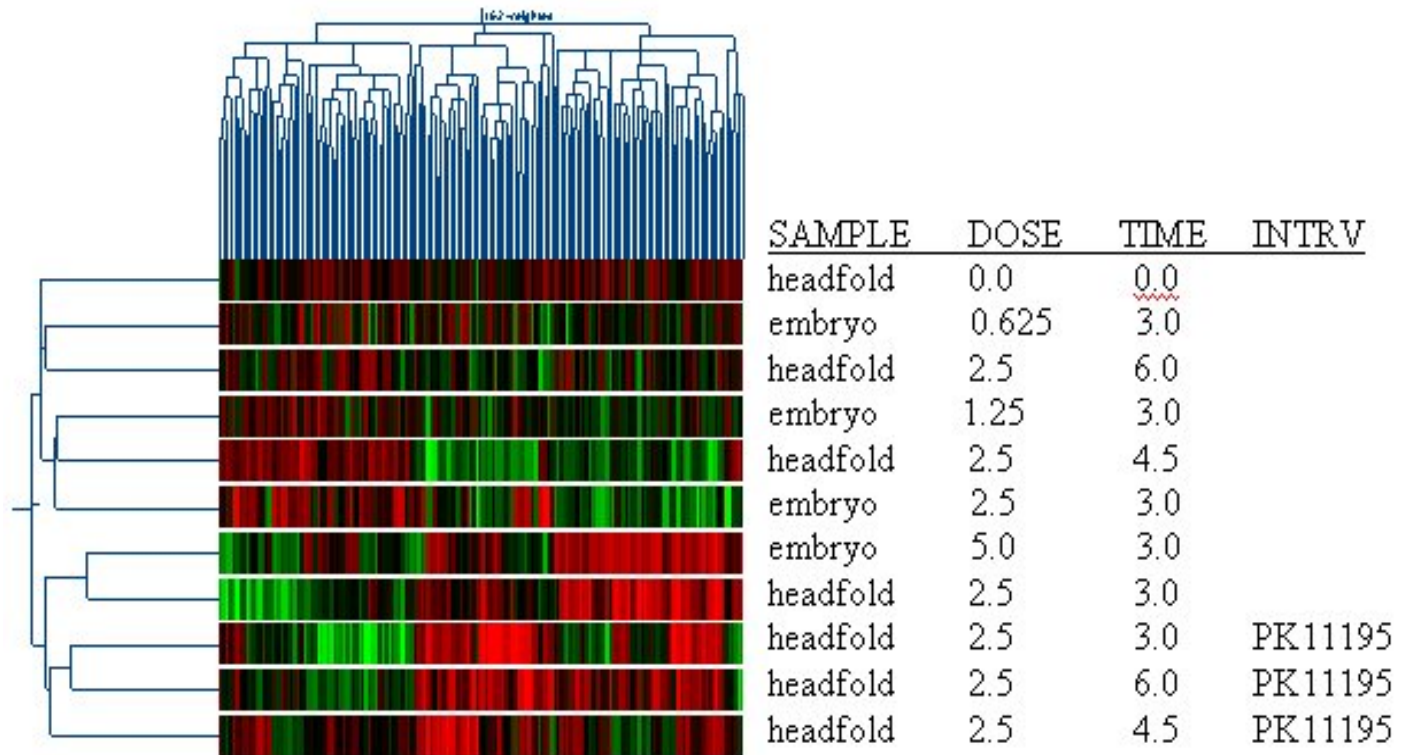
- Summary statistics:
 - mean
 - median, percentile
 - mode
 - standard deviation
 - skewness
 - range
 - covariance matrix

Chapter 3: Exploring Data

- Visualization tools

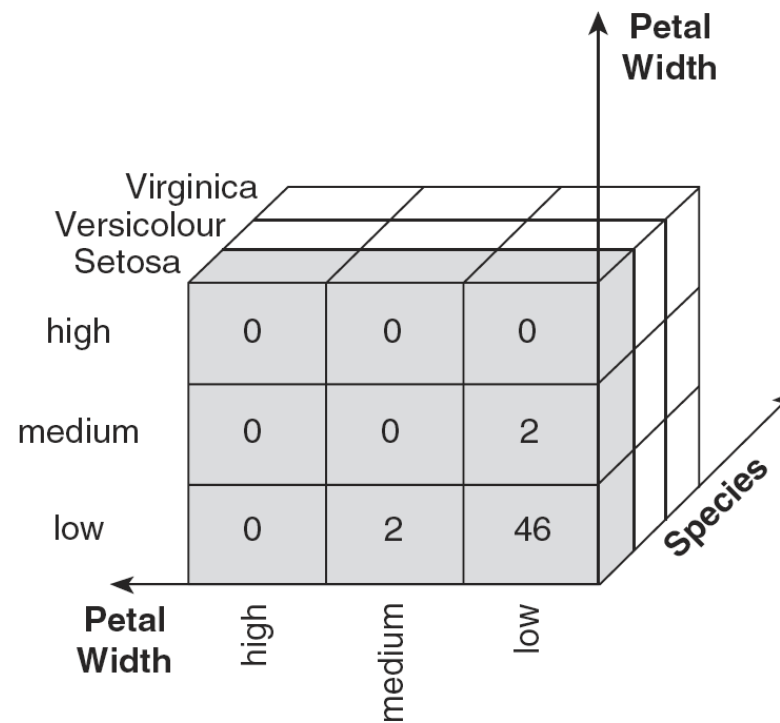


Chapter 3: Exploring Data

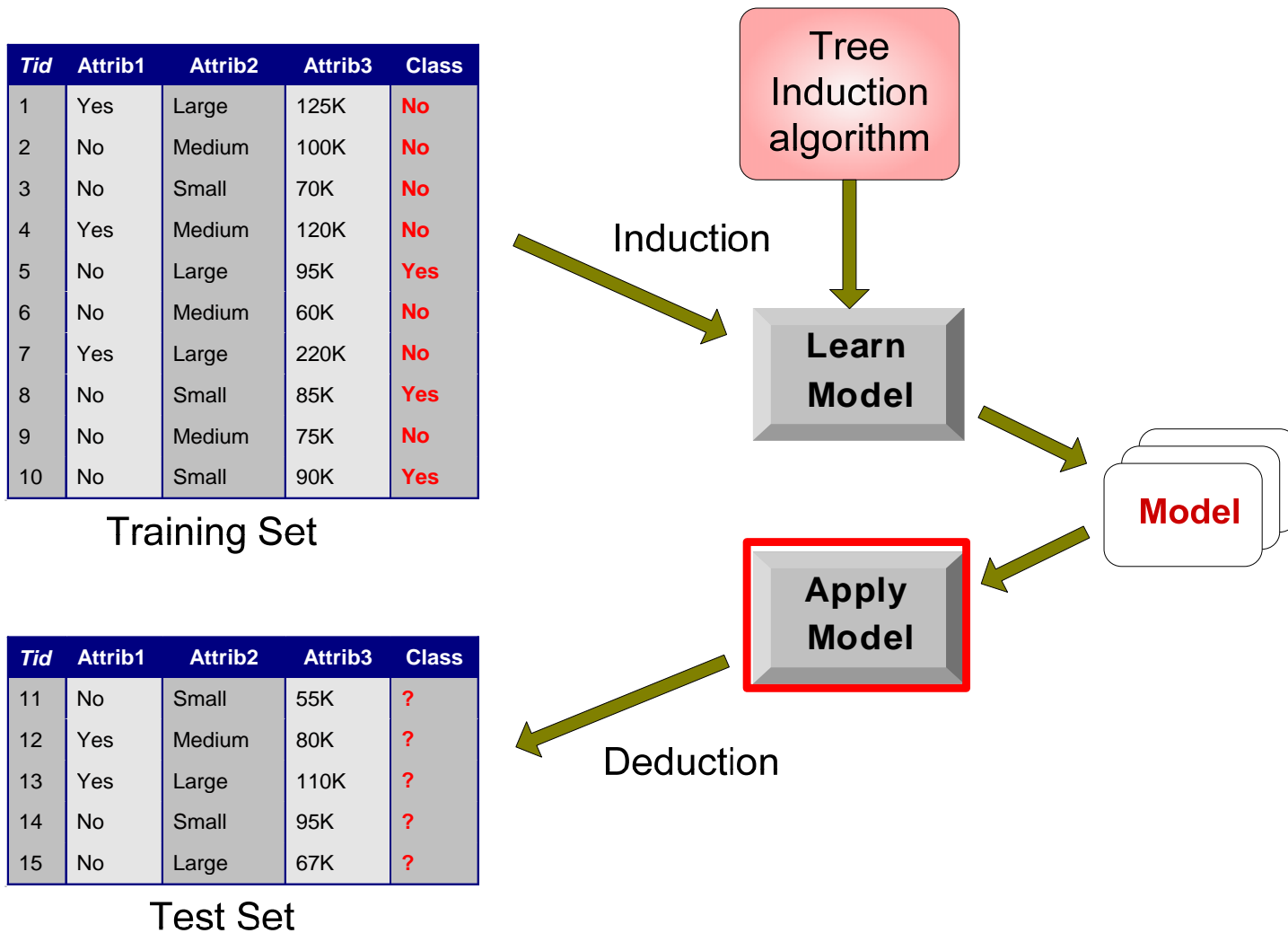


Chapter 3: Exploring Data

- OLAP / data cube

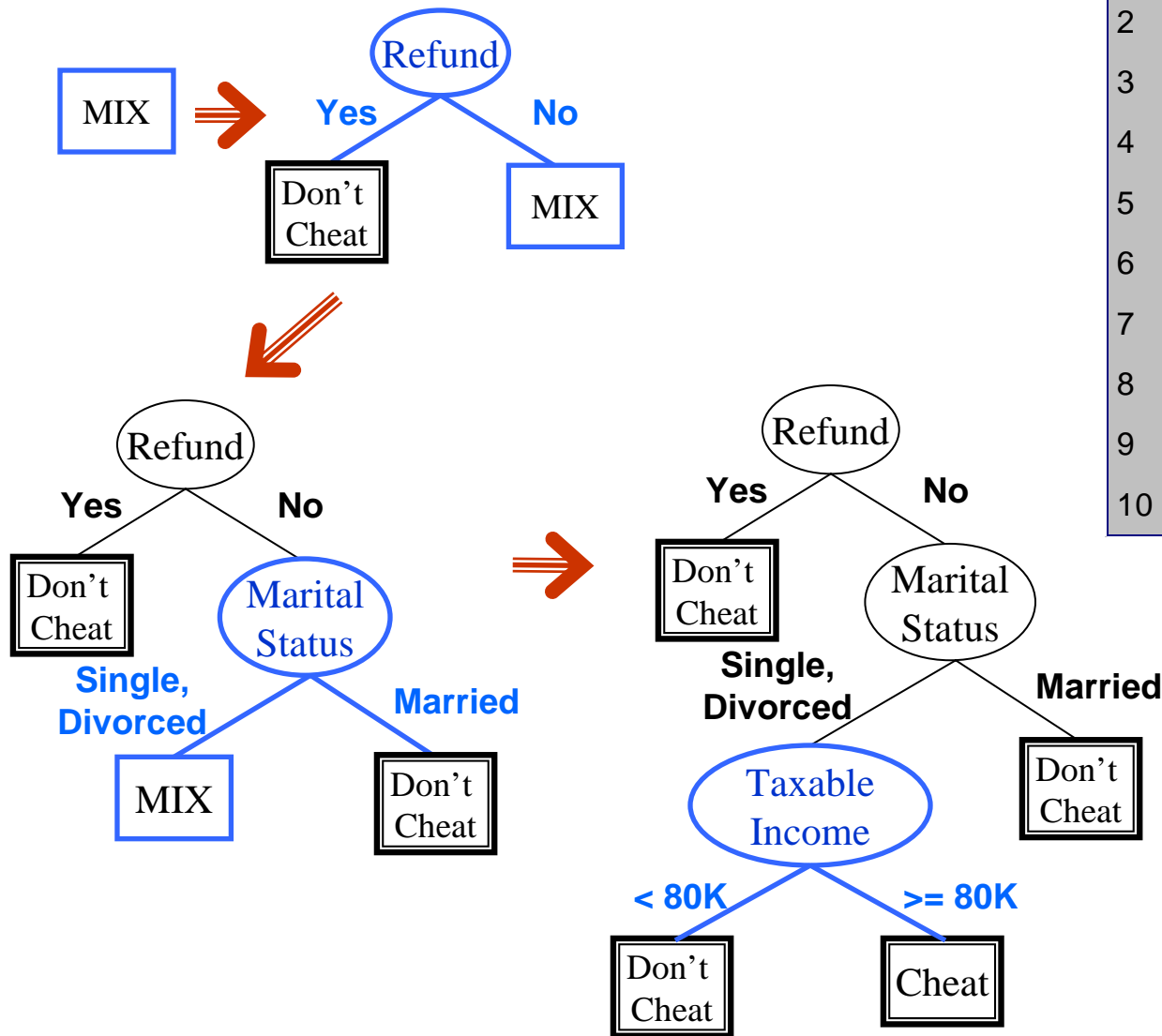


Chapter 4: Classification



Chapter 4: Hunt's algorithm

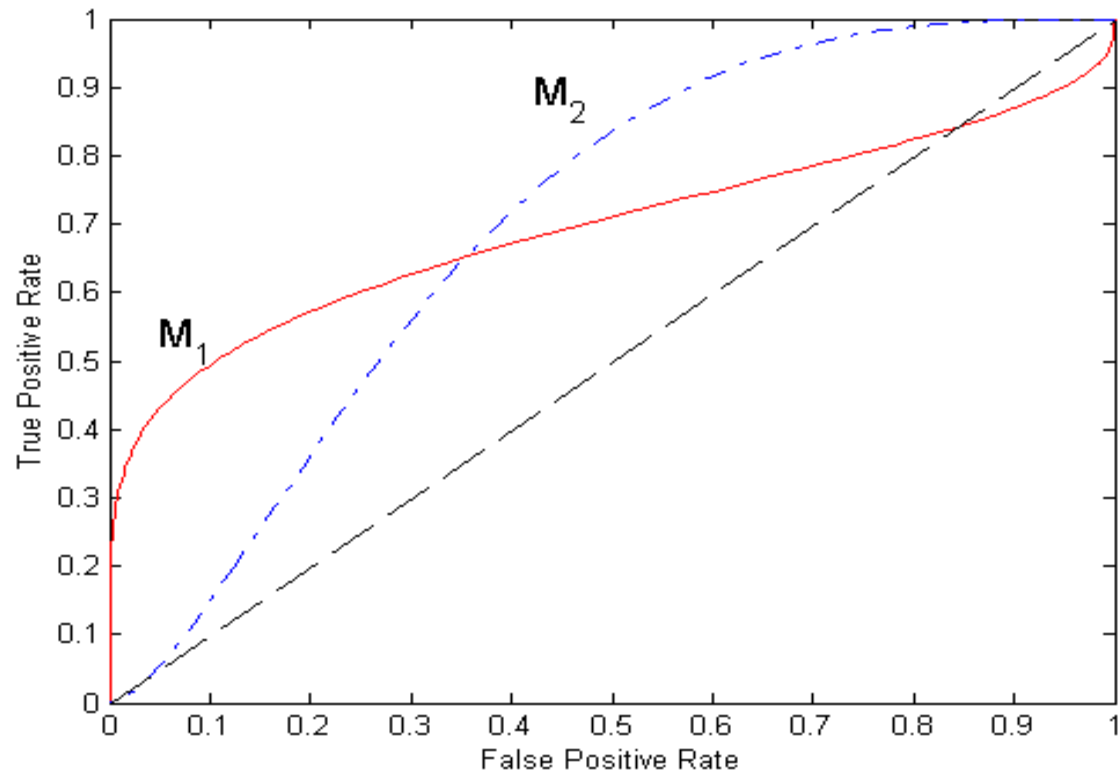
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



Chapter 4: Classification

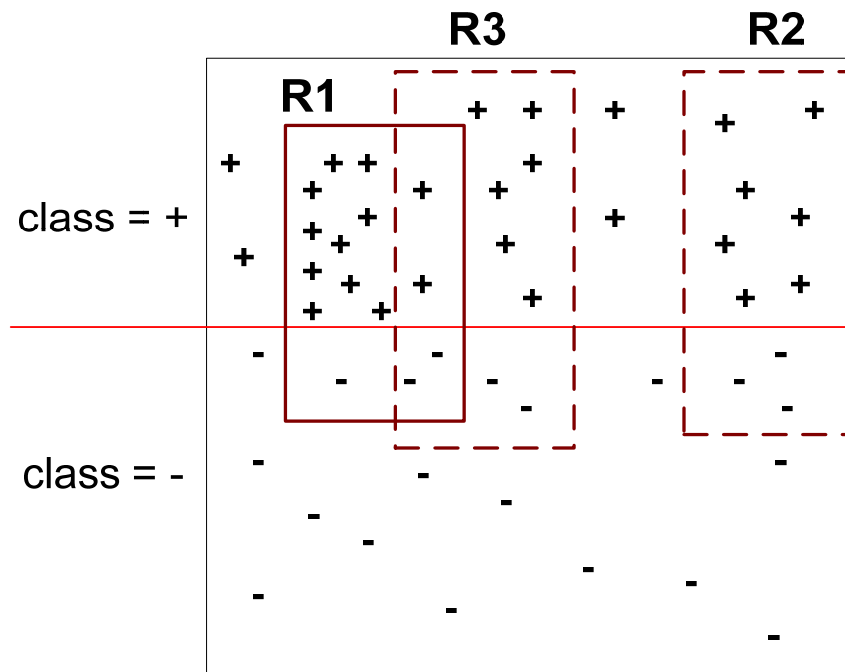
- Criteria for building trees:
 - information
 - misclassification error
 - Gini index
- Overfitting / estimating generalization error
 - optimistic rule
 - pessimistic rule
 - holdout rule
- Other criteria
 - TP/FP, TN/FN
 - confusion matrix
 - ROC curve
 - comparing two classifiers

Chapter 4: Classification



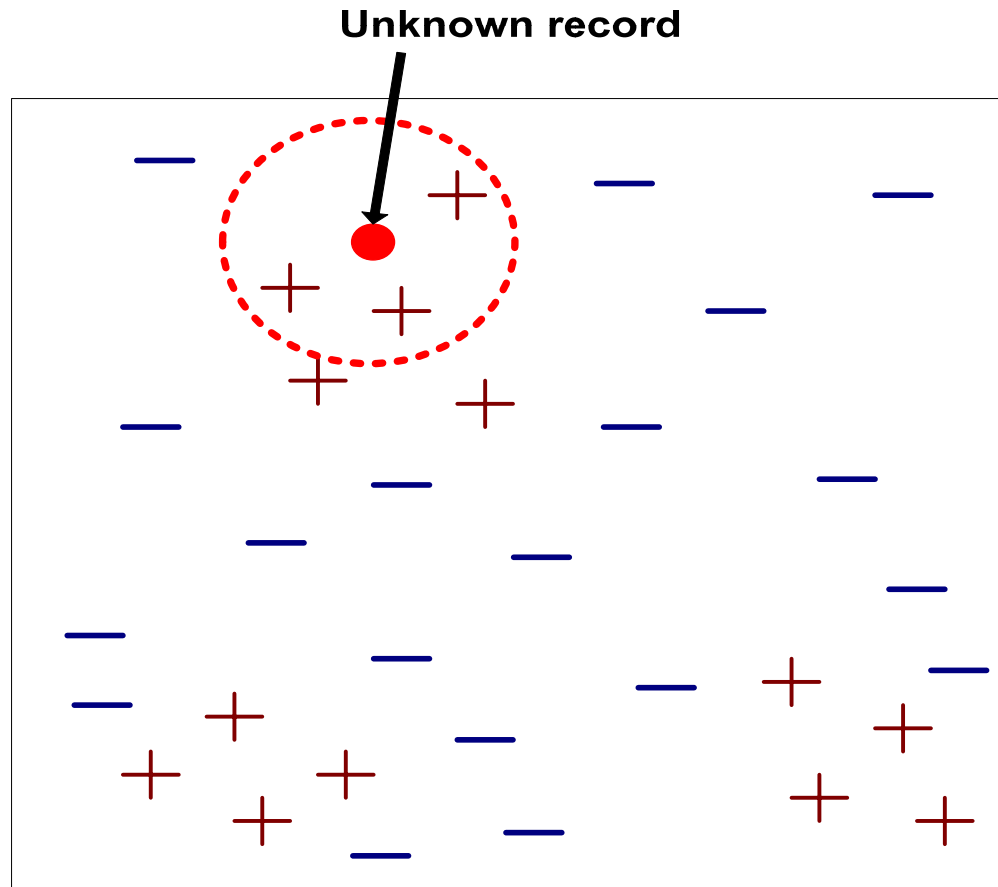
Chapter 5: Classification II

- Rule based classifiers:
 - Rule building
 - Rule accuracy / coverage
 - Rule quality



Chapter 5: Classification II

- Instance (prototype) based / nearest neighbour classifiers



Chapter 5: Classification II

- Naïve Bayes classifiers

- Bayes' rule:

$$P(C | A_1 A_2 \dots A_n) = \frac{P(A_1 A_2 \dots A_n | C) P(C)}{P(A_1 A_2 \dots A_n)}$$

- $P(X|Class=No) = P(Refund=No|Class=No)$
 $\times P(Married|Class=No)$
 $\times P(Income=120K|Class=No)$
 $= 4/7 \times 4/7 \times 0.0072 = 0.0024$
- $P(X|Class=Yes) = P(Refund=No|Class=Yes)$
 $\times P(Married|Class=Yes)$
 $\times P(Income=120K|Class=Yes)$
 $= 1 \times 0 \times 1.2 \times 10^{-9} = 0$

Since $P(X|No)P(No) > P(X|Yes)P(Yes)$

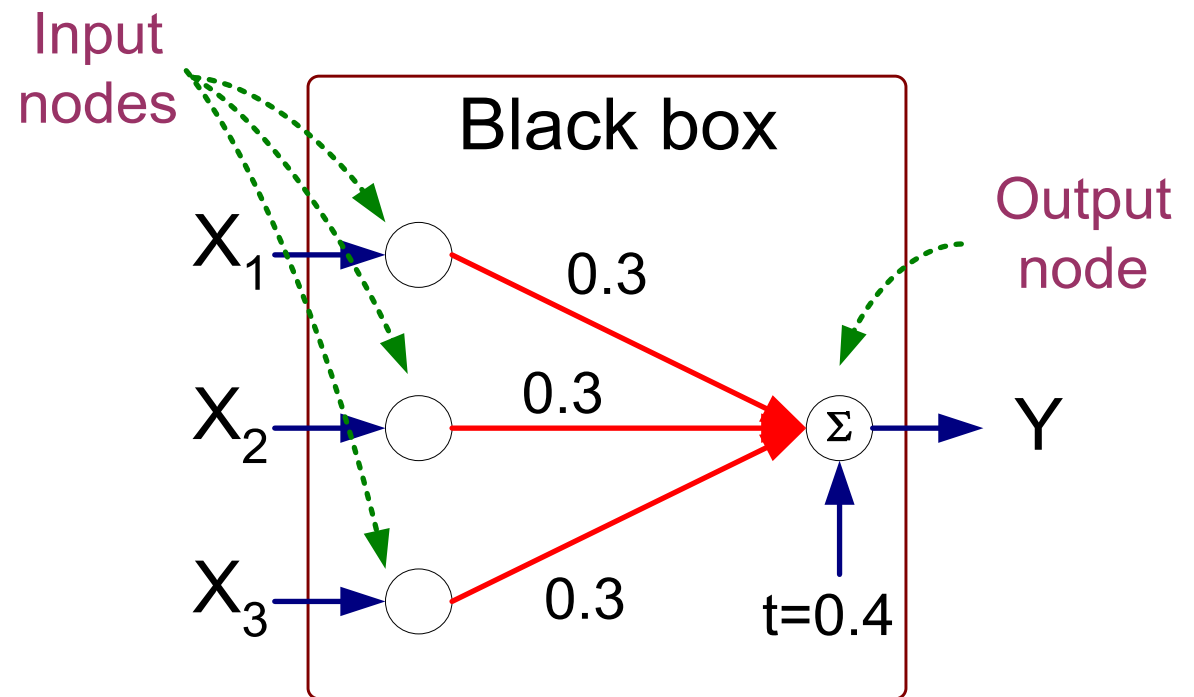
Therefore $P(No|X) > P(Yes|X)$

\Rightarrow Class = No

Chapter 5: Classification II

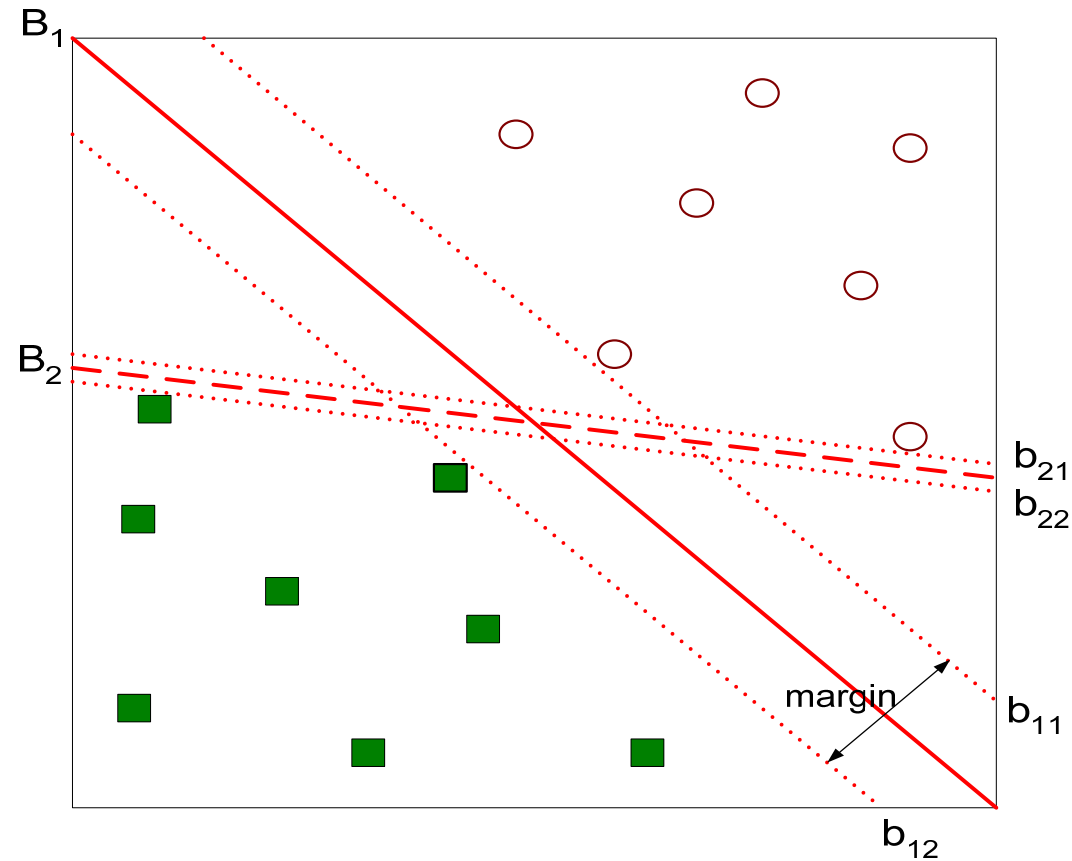
Neural nets

X_1	X_2	X_3	Y
1	0	0	0
1	0	1	1
1	1	0	1
1	1	1	1
0	0	1	0
0	1	0	0
0	1	1	1
0	0	0	0



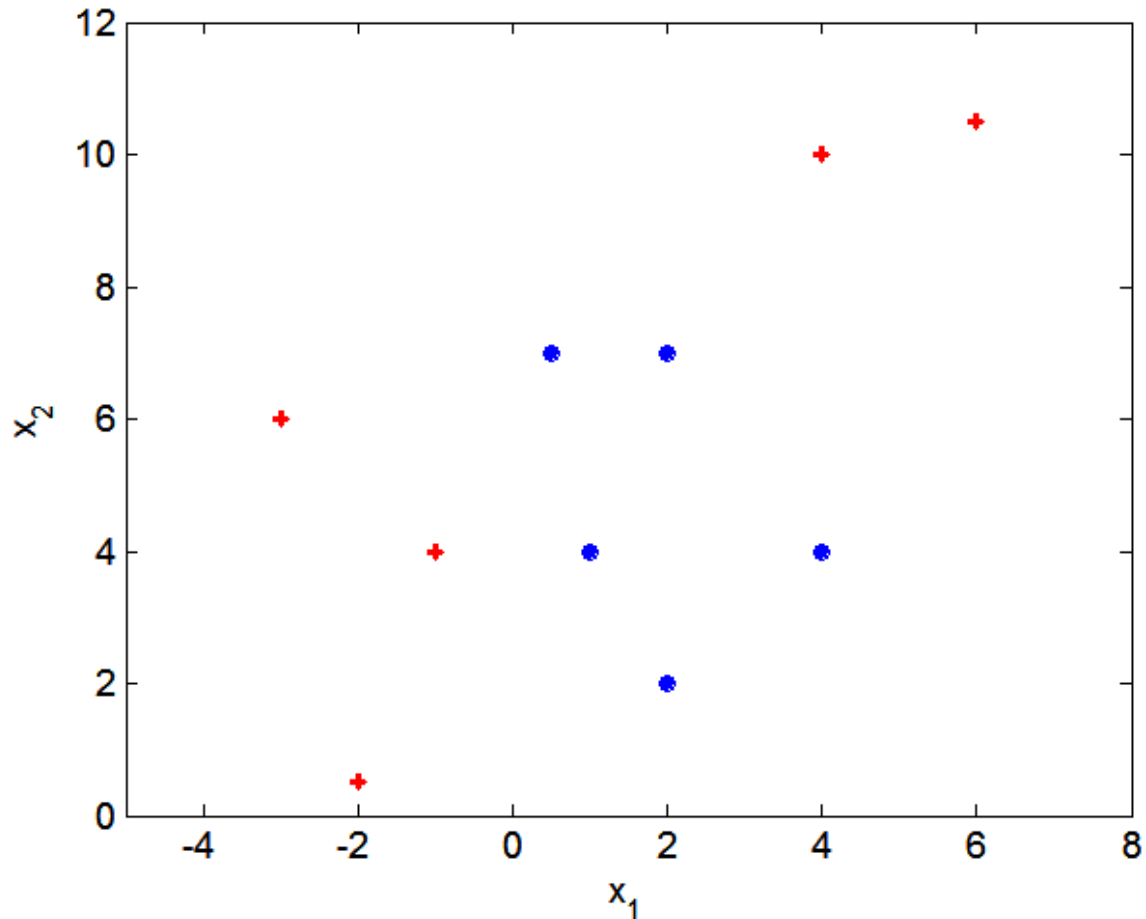
Chapter 5: Classification II

Support vector machines



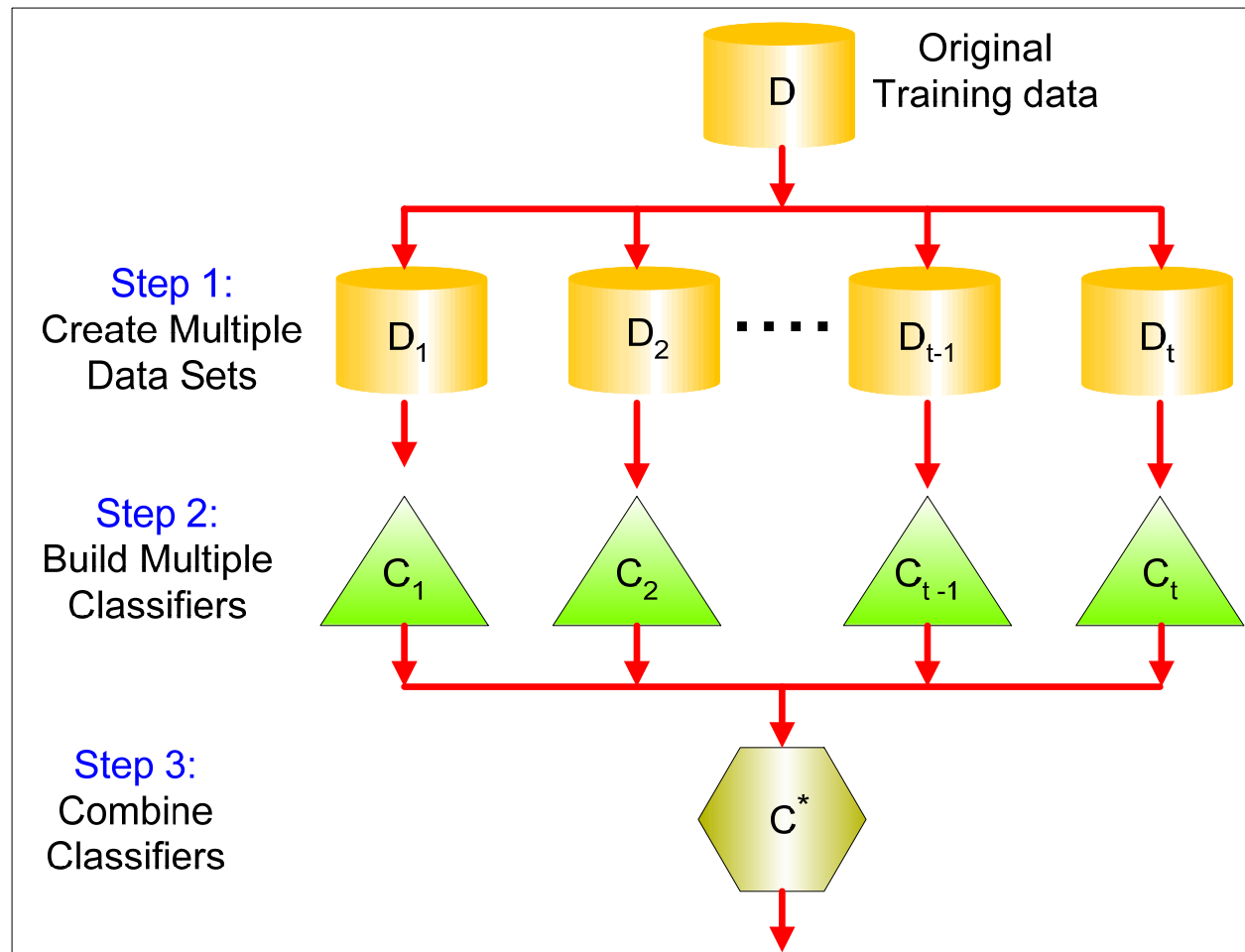
Chapter 5: Classification II

Support vector machines: nonlinear boundaries



Chapter 5: Classification II

Ensemble methods (bagging / random forest)



Chapter 5: Classification II

Ensemble methods: boosting

