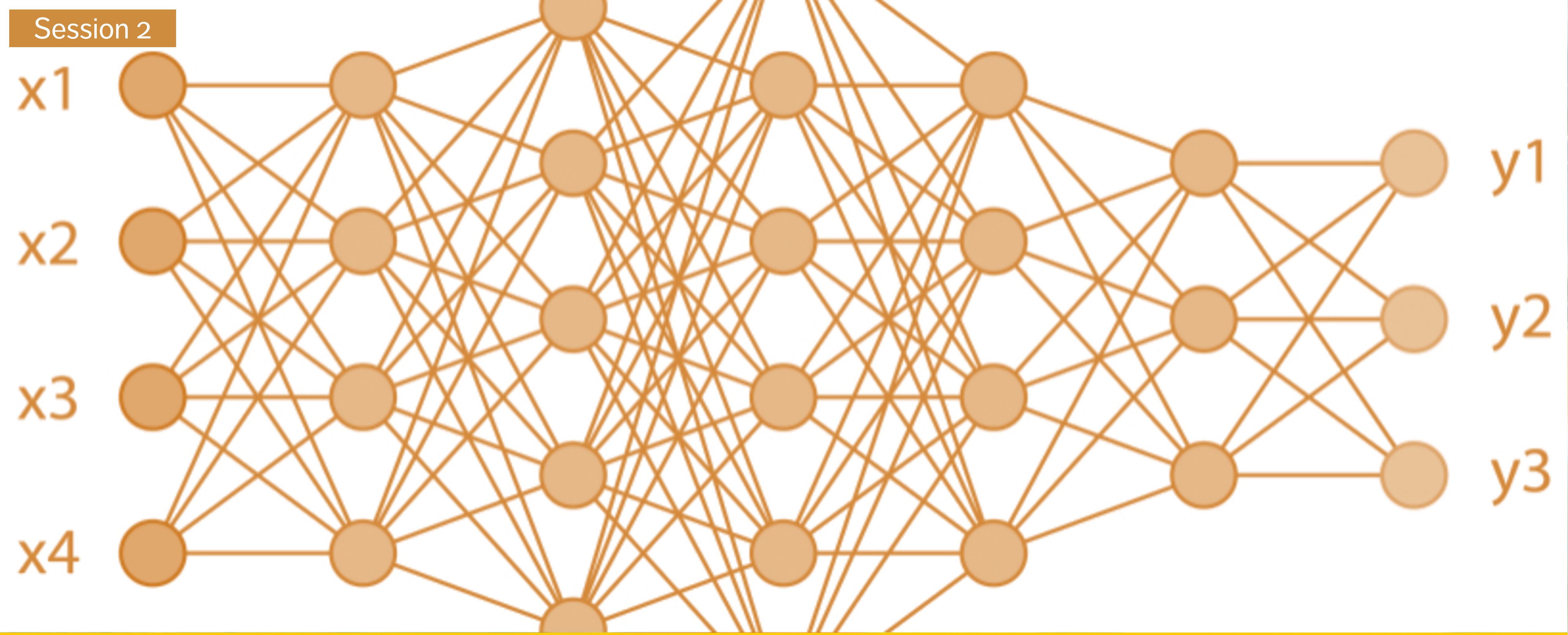
Classification

INTRODUCTION TO MACHINE LEARNING

The diversity of problems that can be addressed by classification algorithms is significant, and covers many domains. It is difficult to comprehensively discuss all the methods in a single book.

[C.C. Aggarwal]



4. Classification Overview

Overview

In **classification**, a sample set of data (the **training** set) is used to determine rules and patterns that divide the data into pre-determined groups, or **classes** (supervised learning; predictive analytics).

The training data usually consists of a **randomly** selected subset of the **labeled** (target) data.

Value estimation (regression) is similar to classification when the target variable is **numerical**.

Overview

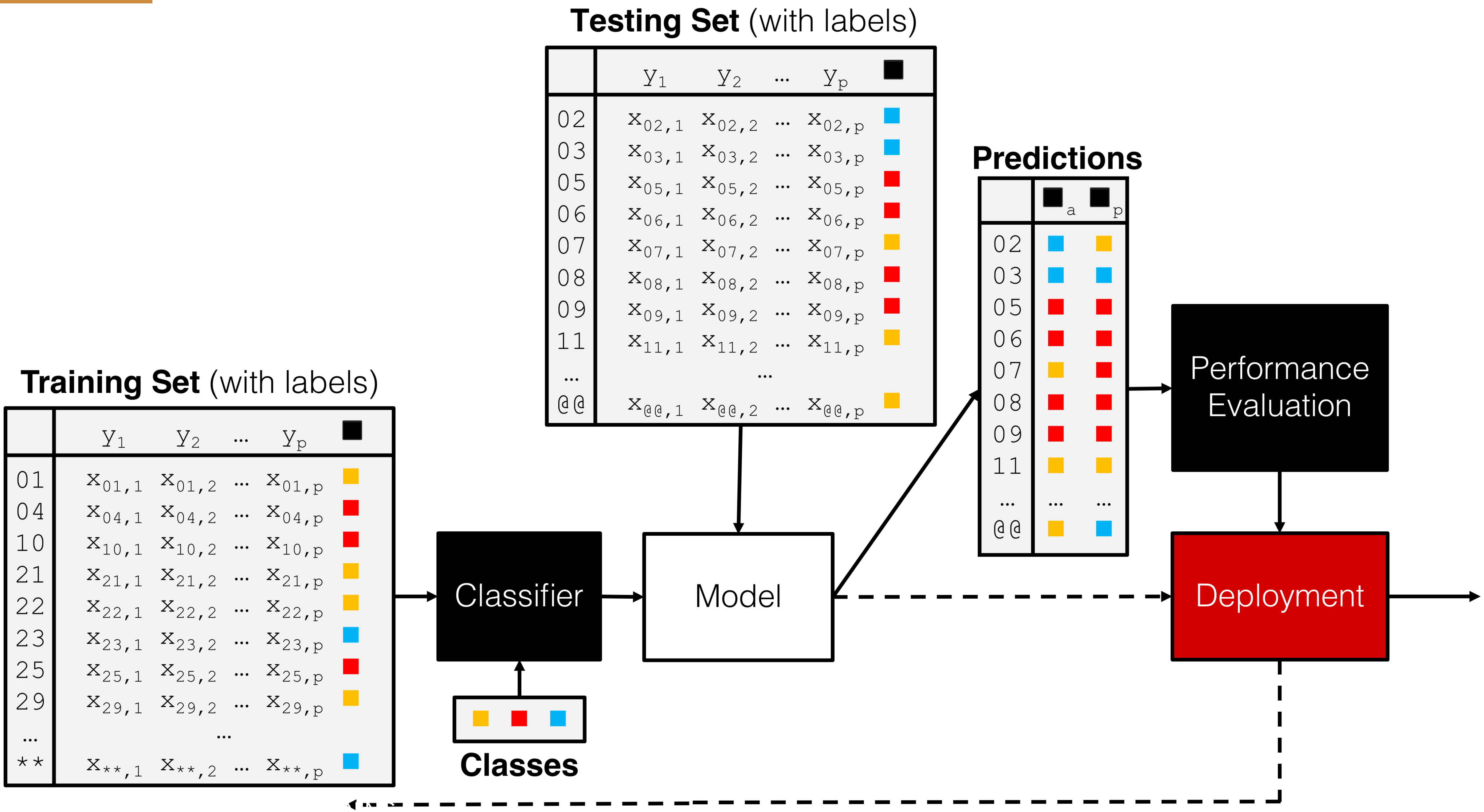
In the **testing** phase, the model is used to assign a class to observations for which the label is hidden, but ultimately known (the **testing** set).

The performance of a classification model is evaluated on the testing set, **never** on the training set. In the **absence** of testing data, classification may be **descriptive** but not predictive.

Technical issues:

- selecting the features to include in the model
- selecting the algorithm
- etc.

Session 2


Testing Set (with labels)
y1 y2 … yp
02 x02,1 x02,2 … x02,p
03 x03,1 x03,2 … x03,p
05 x05,1 x05,2 … x05,p
06 x06,1 x06,2 … x06,p
07 x07,1 x07,2 … x07,p
08 x08,1 x08,2 … x08,p
09 x09,1 x09,2 … x09,p
11 x11,1 x11,2 … x11,p
…
@@ x@@,1 x@@,2 … x@@,p
Predictions
a p
02
03
05
06
07
08
09
11
…
@@
Performance Evaluation
Training Set (with labels)
y1 y2 … yp
01 x01,1 x01,2 … x01,p
04 x04,1 x04,2 … x04,p
10 x10,1 x10,2 … x10,p
21 x21,1 x21,2 … x21,p
22 x22,1 x22,2 … x22,p
23 x23,1 x23,2 … x23,p
25 x25,1 x25,2 … x25,p
29 x29,1 x29,2 … x29,p
…
** x**,1 x**,2 … x**,p
Classifier
Model
Deployment
Classes

Applications

Medicine and Health Science

- predicting which patient is at risk of suffering a second, fatal heart attack within 30 days based on health factors (blood pressure, age, sinus problems, etc.)

Social Policies

- predicting the likelihood of requiring assisting housing in old age based on demographic information/survey answers

Marketing and Business

- predicting which customers are likely to switch to another cell phone company based on demographics and usage

Applications

Other uses include:

- Predicting that an object belongs to a particular class.
- Organizing and grouping instances into categories.
- Enhancing the detection of relevant objects
  - **avoidance:** “this object is an incoming vehicle”
  - **pursuit:** “this borrower is unlikely to default on her mortgage”
  - **degree:** “this dog is 90% likely to live until it’s 7 years old”
- Predicting the inflation rate for the coming two years based on a number of economic indicators.

Examples

Scenario:

A motor insurance company has a fraud investigation dept. that studies up to 30% of all claims made, yet money is still getting lost on fraudulent claims.

Questions: can we predict

- whether a claim is likely to be fraudulent?
- whether a customer is likely to commit fraud in the near future?
- whether an application for a policy is likely to result in a fraudulent claim?
- the amount by which a claim will be reduced if it is fraudulent?

Examples

Scenario:

Customers who make a large number of calls to a mobile phone company's customer service number have been identified as churn risks. The company is interested in reducing said churn.

Questions: can we predict

- the overall lifetime value of a customer?
- which customers are more likely to churn in the near future?
- what retention offer a particular customer will best respond to?

Case Study

Minnesota Tax Audits

Hsu *et al.*
[Data Mining Based Tax Audit Selection: A Case Study of a Pilot Project at the Minnesota Department of Revenue]
Real World Data Mining Applications, 2015

Objective

The *U.S. Internal Revenue Service* (IRS) estimated that there were large gaps between **revenue owed** and **revenue collected** for 2001 and for 2006.

Using DoR data (*Minnesota Department of Revenue*), the authors increased **efficiency** in the audit selection process and sought to **reduce the gap** between revenue owed and revenue collected.

Case Study

Minnesota Tax Audits

Hsu *et al.*
[Data Mining Based Tax Audit Selection: A Case Study of a Pilot Project at the Minnesota Department of Revenue]
Real World Data Mining Applications, 2015

Methodology

1. **data selection and separation:** experts selected several hundred cases to audit and divided them into training, testing and validating sets
2. **classification modeling** using MultiBoosting, Naïve Bayes, C4.5 decision trees, multilayer perceptrons, support vector machines, etc.
3. **evaluation of all models** on the testing set – models performed poorly until the size of the business being audited was recognized to have an effect, leading to two separate tasks (large/small businesses).
4. **model selection/validation** compared the estimated accuracy between different classification model predictions and the actual field audits (MultiBoosting with Naïve Bayes was selected as the final model; suggesting improvements to increase audit efficiency).

Case Study

Minnesota Tax Audits

Hsu *et al.*
[Data Mining Based Tax Audit Selection: A Case Study of a Pilot Project at the Minnesota Department of Revenue]
Real World Data Mining Applications, 2015

Data

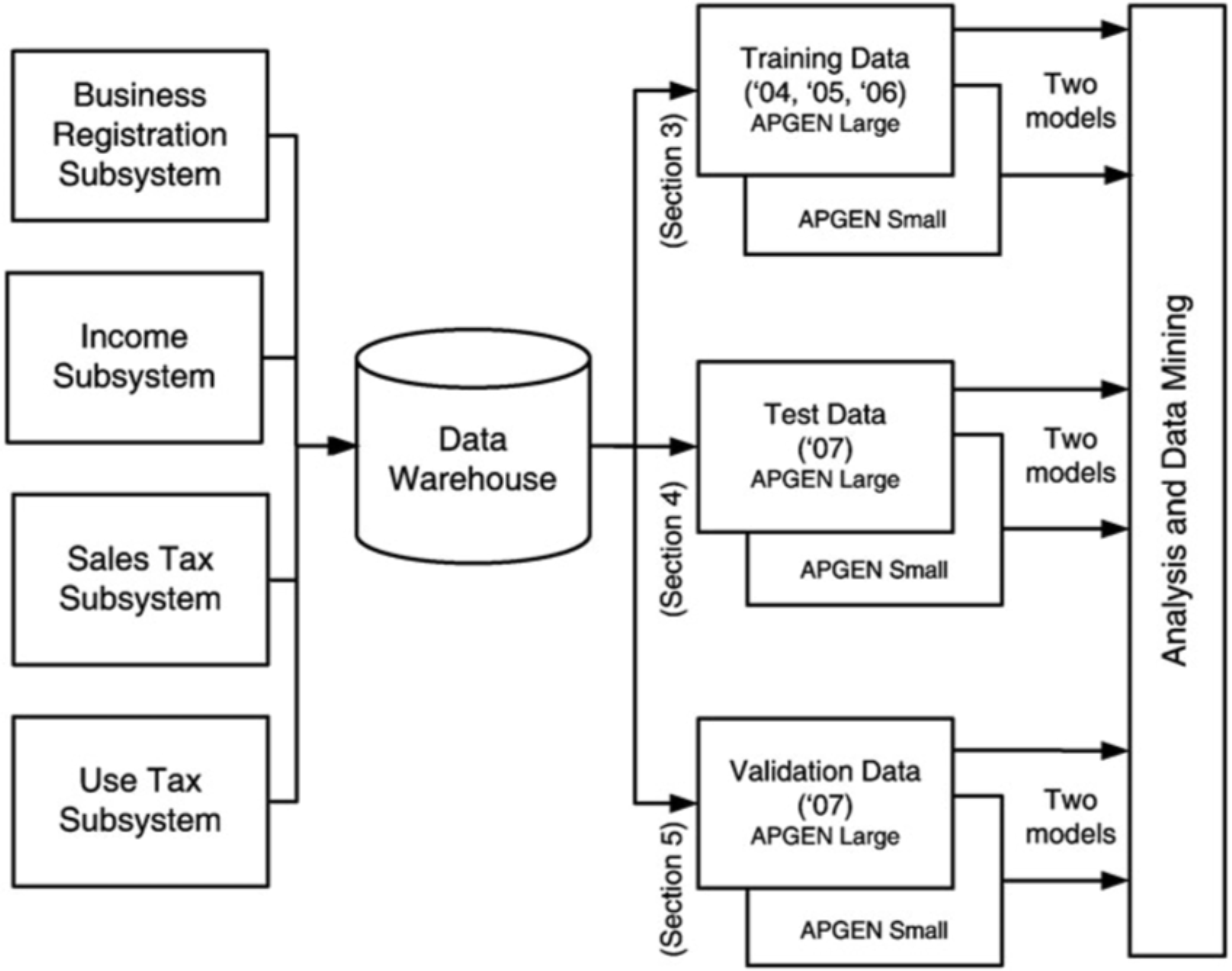
Selected tax audit cases from 2004 to 2007, collected by the audit experts, which were split into training, testing and validation sets:

- the **training data** set consisted of *Audit Plan General* (APGEN) *Use Tax* audits and their results for the years 2004-2006
- the **testing data** consisted of APGEN Use Tax audits conducted in 2007 and was used to test or evaluate models (for Large and Smaller businesses) built on the training dataset
- while **validation** was assessed by actually conducting field audits on predictions made by models built on 2007 Use Tax return data processed in 2008.

Case Study

Minnesota Tax Audits

Hsu *et al.*
Data Mining Based Tax Audit Selection: A Case Study of a Pilot Project at the Minnesota Department of Revenue
Real World Data Mining Applications, 2015



Case Study

Minnesota Tax Audits

Hsu *et al.*
[Data Mining Based Tax Audit Selection: A Case Study of a Pilot Project at the Minnesota Department of Revenue]
Real World Data Mining Applications, 2015

Strengths and Limitations of the Algorithms

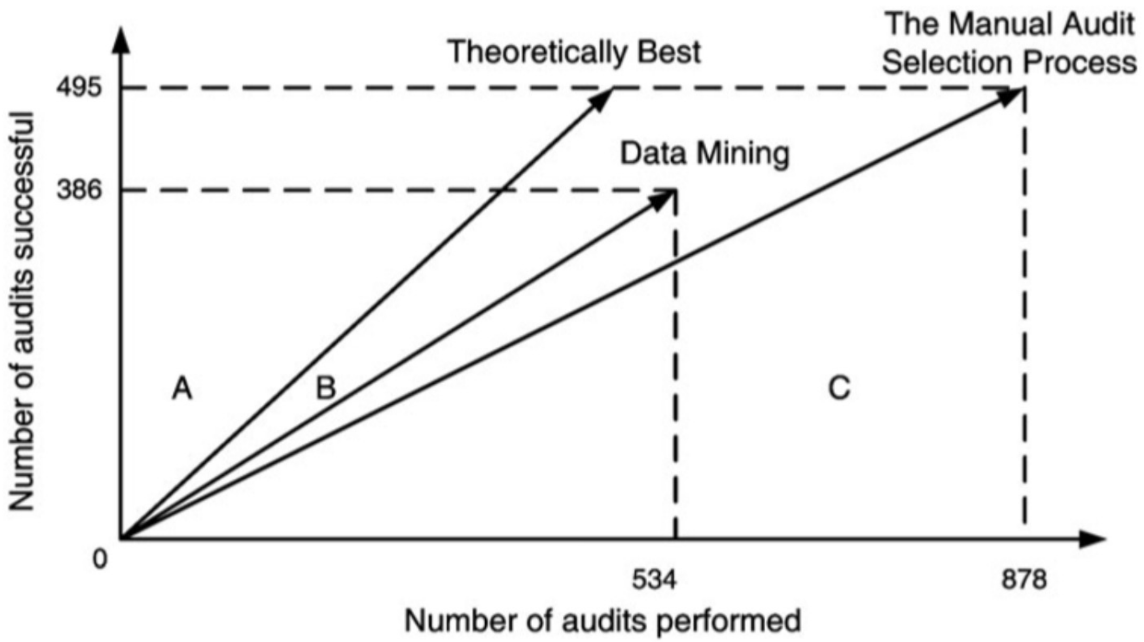
- Naïve Bayes classification assumes independence of the features, which rarely occurs in real-world situations. This approach is also known to potentially introduce bias to classification schemes. In spite of this, classification models built using it have a successfully track record.
- MultiBoosting is an **ensemble technique** that uses committee (i.e. groups of classification models) and "group wisdom" to make predictions; unlike other ensemble techniques, it is different from other ensemble techniques in the sense that it forms a committee of sub-committees, which has a tendency to reduce both bias and variance of predictions.

Case Study

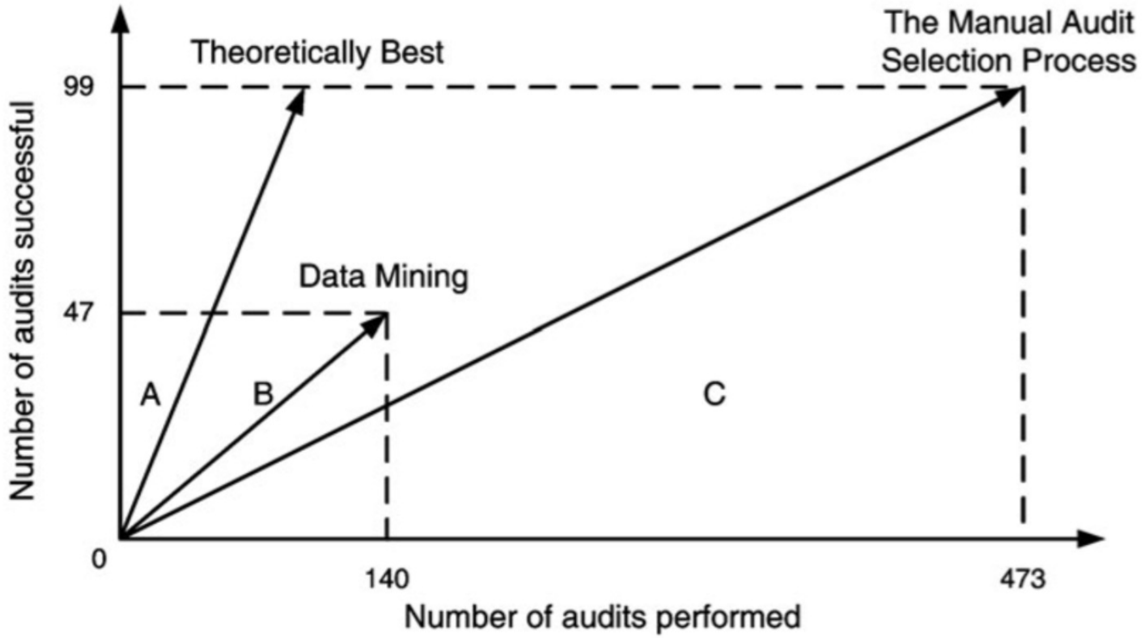
Minnesota Tax Audits

Hsu *et al.*
[Data Mining Based Tax Audit Selection: A Case Study of a Pilot Project at the Minnesota Department of Revenue]
Real World Data Mining Applications, 2015

APGEN Large



APGEN Small



Case Study

Minnesota Tax Audits

Hsu *et al.*
[Data Mining Based Tax Audit Selection: A Case Study of a Pilot Project at the Minnesota Department of Revenue]
Real World Data Mining Applications, 2015

APGEN Large

	Predicted as good	Predicted as bad
Actually good	386 (Use tax collected) R = \$5,577,431 (83.6 %) C = \$177,560 (44 %)	109 (Use tax lost) R = \$925,293 (13.9 %) C = \$50,140 (12.4 %)
Actually bad	148 (costs wasted) R = \$72,744 (1.1 %) C = \$68,080 (16.9 %)	235 (costs saved) R = \$98,105 (1.4 %) C = \$108,100 (26.7 %)

APGEN Small

	Predicted as good	Predicted as bad
Actually good	47 (Use tax collected) R = \$263,706 (42.5 %) C = \$21,620 (9.9 %)	52 (Use tax lost) R = \$264,101 (42.5 %) C = \$23,920 (11 %)
Actually bad	93 (costs wasted) R = \$24,441 (3.9 %) C = \$42,780 (19.7 %)	281 (costs saved) R = \$68,818 (11.1 %) C = \$129,260 (59.4 %)

Case Study

Minnesota Tax Audits

Hsu *et al.*
[Data Mining Based Tax Audit Selection: A Case Study of a Pilot Project at the Minnesota Department of Revenue]
Real World Data Mining Applications, 2015

Take-Aways

- Many models were churned out before the team made a final selection.
- Past performance of a model family in a previous project can guide the selection, but remember the *No Free Lunch (NFL) Theorem*: nothing works best all the time!
- The feature selection process could very well require a number of visits to domain experts before the feature set yields promising results.
- Data analysis teams should seek out individuals with a good understand of both data and context.
- Domain-specific knowledge has to be integrated in the model in order to beat random classifiers, on average.
- Even slight improvements over the current approach can find a useful place in an organization – data science is not solely about Big Data and disruption!

General Classification Comments

Classification is linked to **probability estimation**

- approaches based on regression models could prove fruitful

Rare occurrences (often more interesting or important):

- historical data at Fukushima's nuclear reactor prior to the meltdown could not have been used to learn about meltdowns, for instance
- predicting no meltdown will yield correct predictions roughly 99.99% of the time, but will miss the point of the exercise

No Free-Lunch Theorem: no classifier works best for all data.

With big datasets, algorithms must also consider efficiency.

Suggested Reading

Classification Overview

Data Understanding, Data Analysis, Data Science
Volume 3: Spotlight on Machine Learning

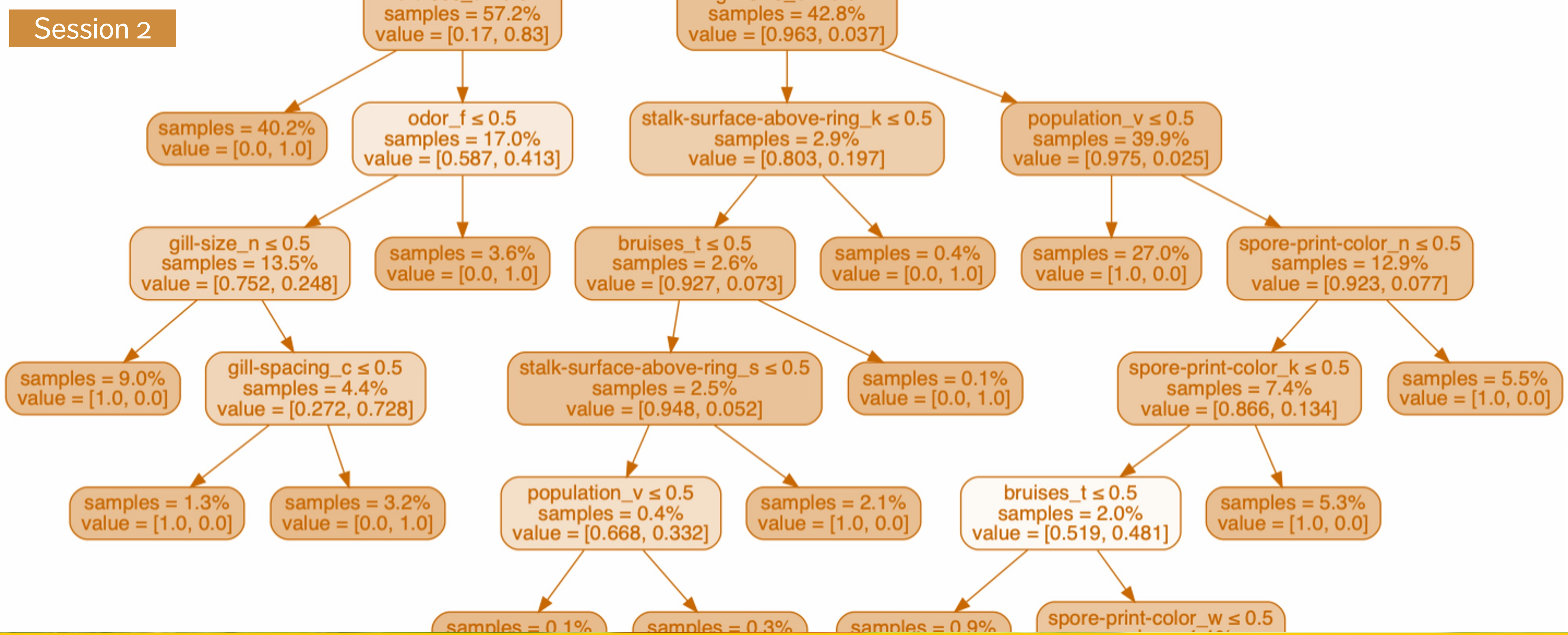
19. Introduction to Machine Learning
    19.4 Classification and Regression
    - Overview
    - Case Study: Minnesota Tax Audits

21. Focus on Classification and Supervised Learning
    21.1 Overview
    - Formalism

Exercises

Classification Overview

1. How would you use standard statistical modeling techniques to answer the questions presented in the two scenarios in the slides?
2. Identify scenarios and questions that could use classification and/or value estimation in your every day work activities.



5. Decision Trees and Other Algorithms

Classification Algorithms

Logistic Regression

- classical model
- affected by variance inflation and variable selection process

Neural Networks

- hard to interpret
- requires all variables to be of the same type
- easier to train since backpropagation (chain rule)

Bayesian Methods

Decision Trees

- may overfit the data if not pruned correctly (manually?)

Classification Algorithms

Naïve Bayes Classifiers

- quite successful for text mining applications (spam filter)
- assumptions not often met in practice

Support Vector Machines

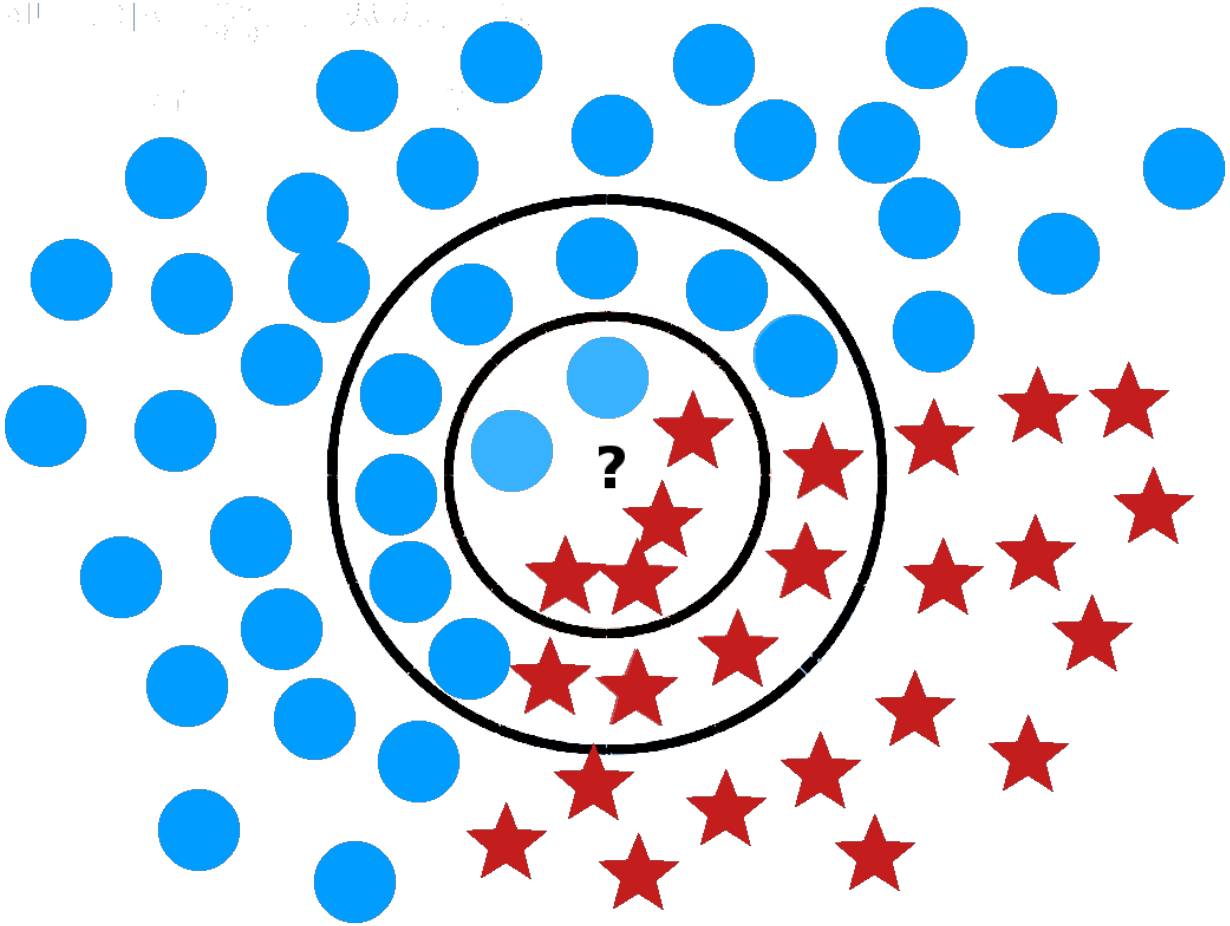
- may be difficult to interpret (non-linear boundaries)
- can help mitigate big data difficulties

Boosting Methods

Nearest Neighbours Classifiers

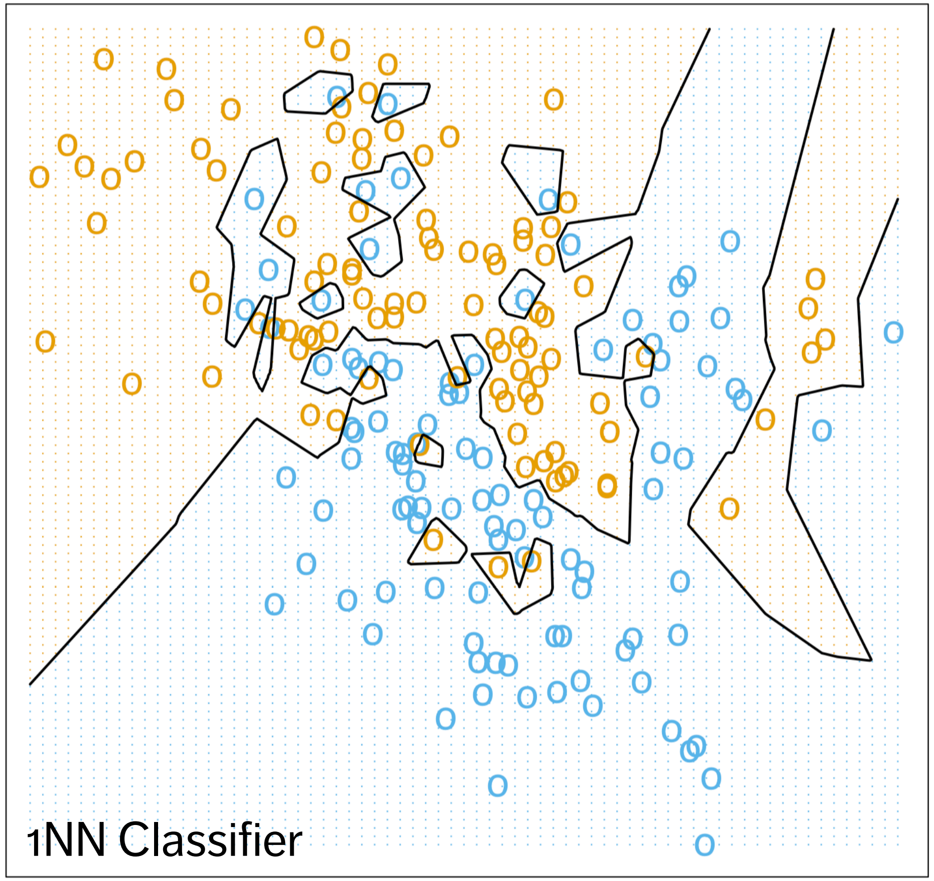
- require very little assumptions about the data
- not very stable (adding points may substantially modify the boundary)

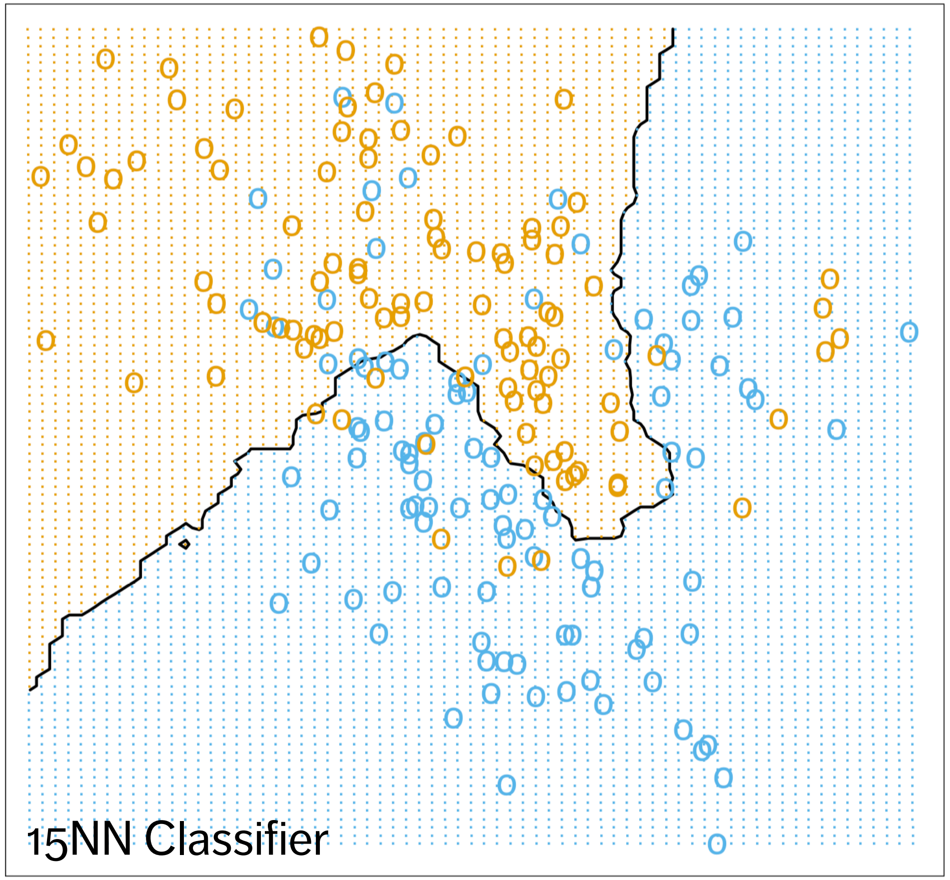
k –Nearest Neighbours Classifier



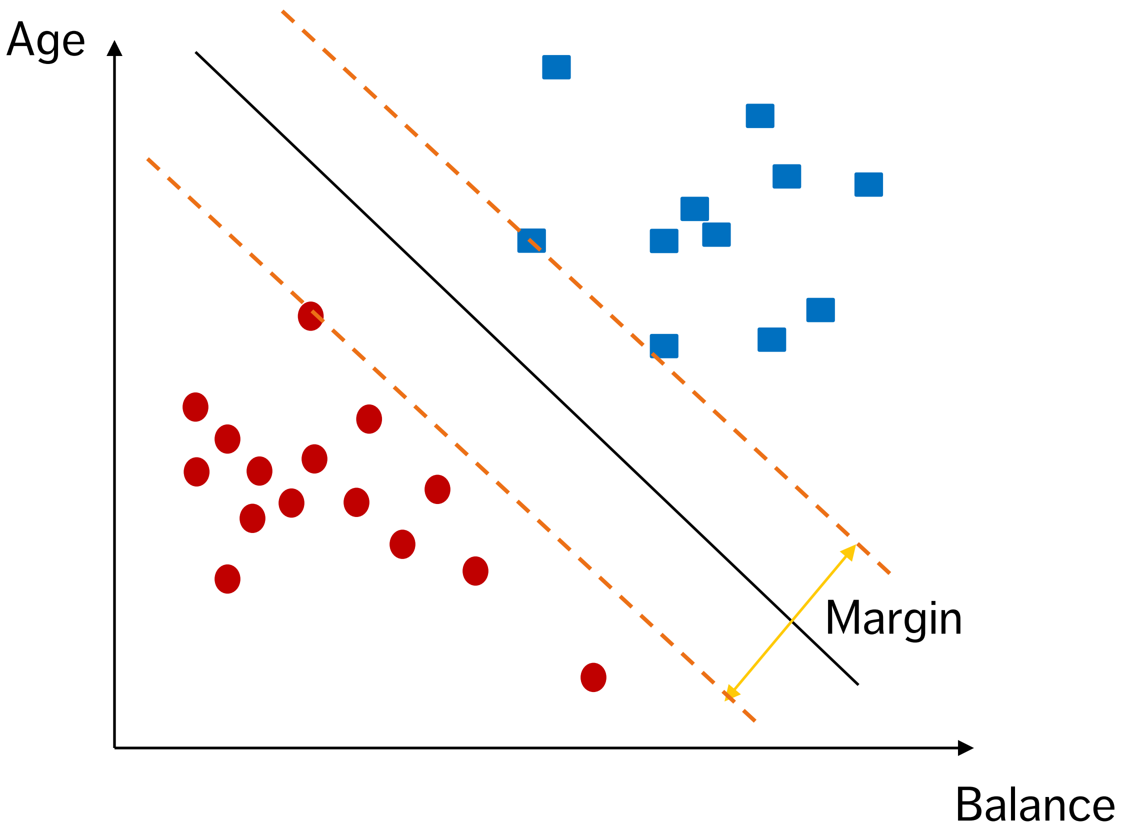
[Tibshirani, Hastie & Friedman, *The Elements of Statistical Learning*]

k –Nearest Neighbours Classifier



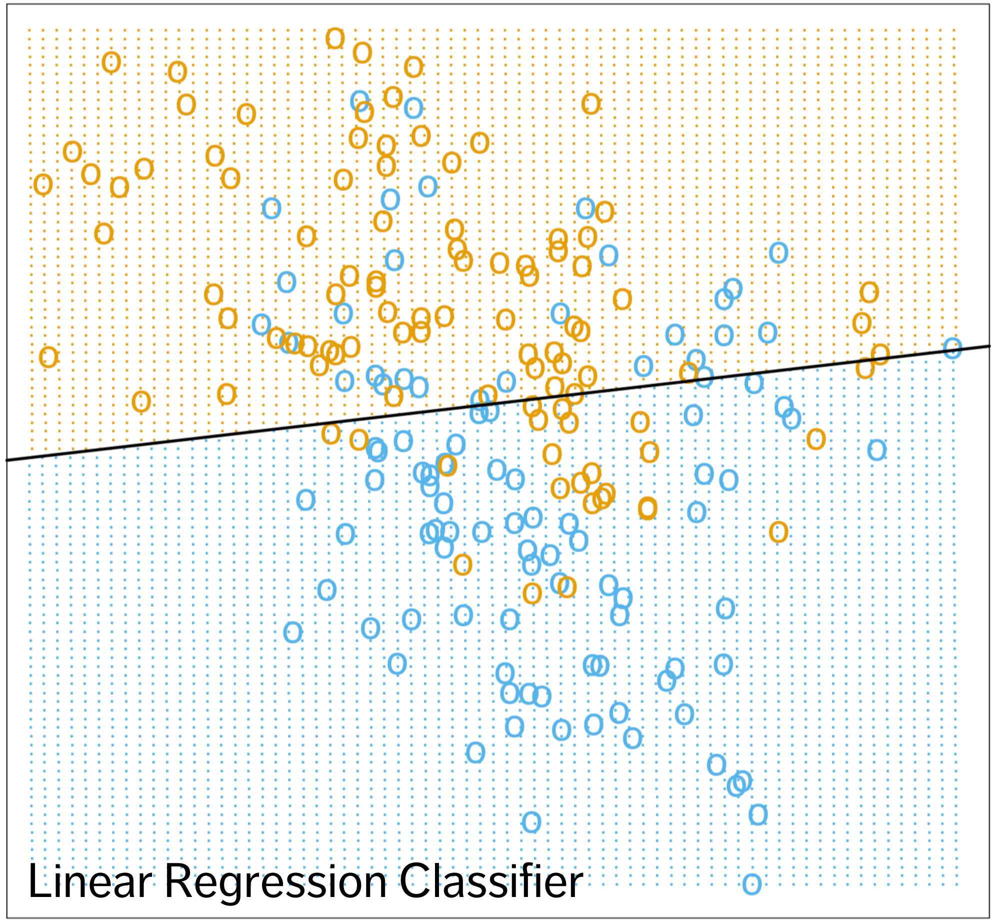


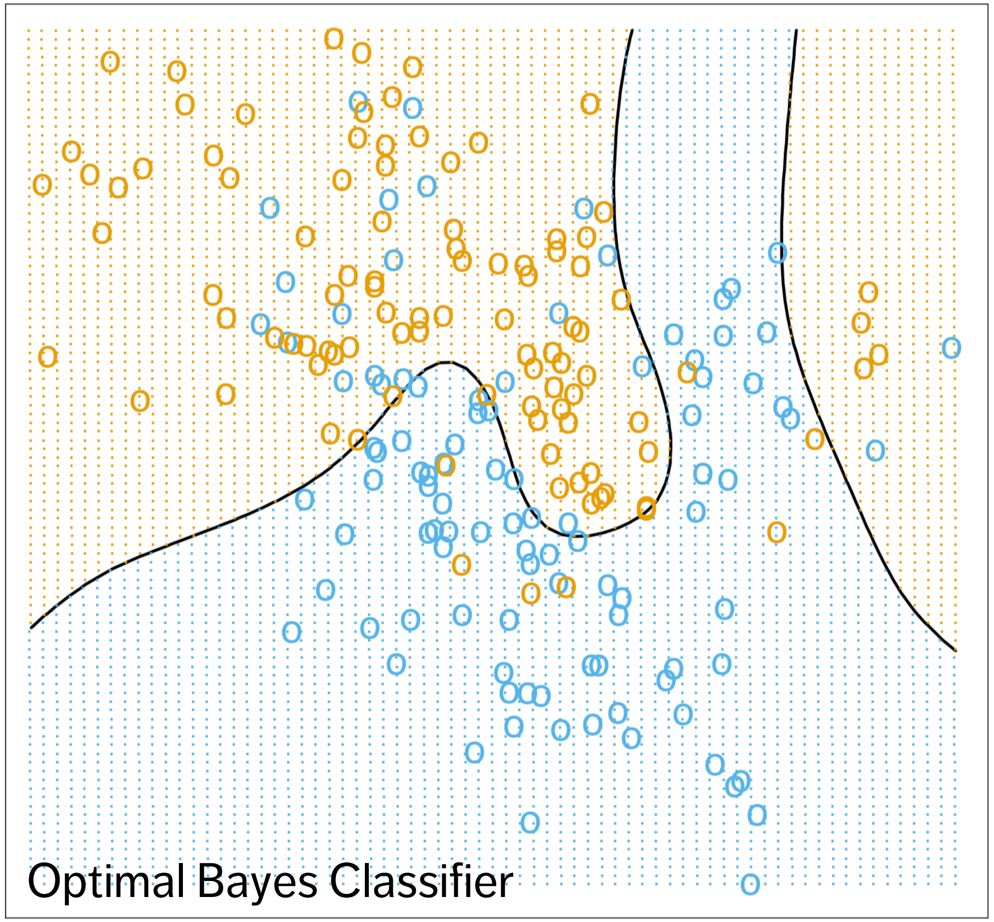
Support Vector Machines



[Tibshirani, Hastie & Friedman, *The Elements of Statistical Learning*]

Other Classifiers





[adapted from A. Ng, K. Soo, *Numsense!*, Surviving a Disaster]

Decision Trees

Decision trees are perhaps the most **intuitive** of these methods: classification is achieved by following a path up the tree, from its **root**, through its **branches**, and ending at its **leaves**.

Decision Trees

To make a **prediction** for a new instance, follow the path down the tree, and read the prediction directly once a leaf is reached.

Creating the tree and traversing it might be **time-consuming** if there are too many variables.

Prediction accuracy can be a concern in trees whose growth is **unchecked**. In practice, the criterion of **purity** at the leaf-level is linked to bad prediction rates for new instances.

- other criteria are often used to prune trees, which may lead to **impure** leaves (i.e. with non-trivial entropy).

Decision Tree Algorithm (ID3)

Task: grow a decision tree using a training set (a subset of the data for which the correct classification of the target is known).

Overview:

1. Split the training data (**parent**) set into (**children**) subsets, using the different levels of a particular attribute
2. Compute the **information gain** for each subset
3. Select the **most advantageous** split, gain-wise
4. Repeat for each node until some **leaf** criterion is met (each item in the leaf has the same classification is one possibility)

Information Gain

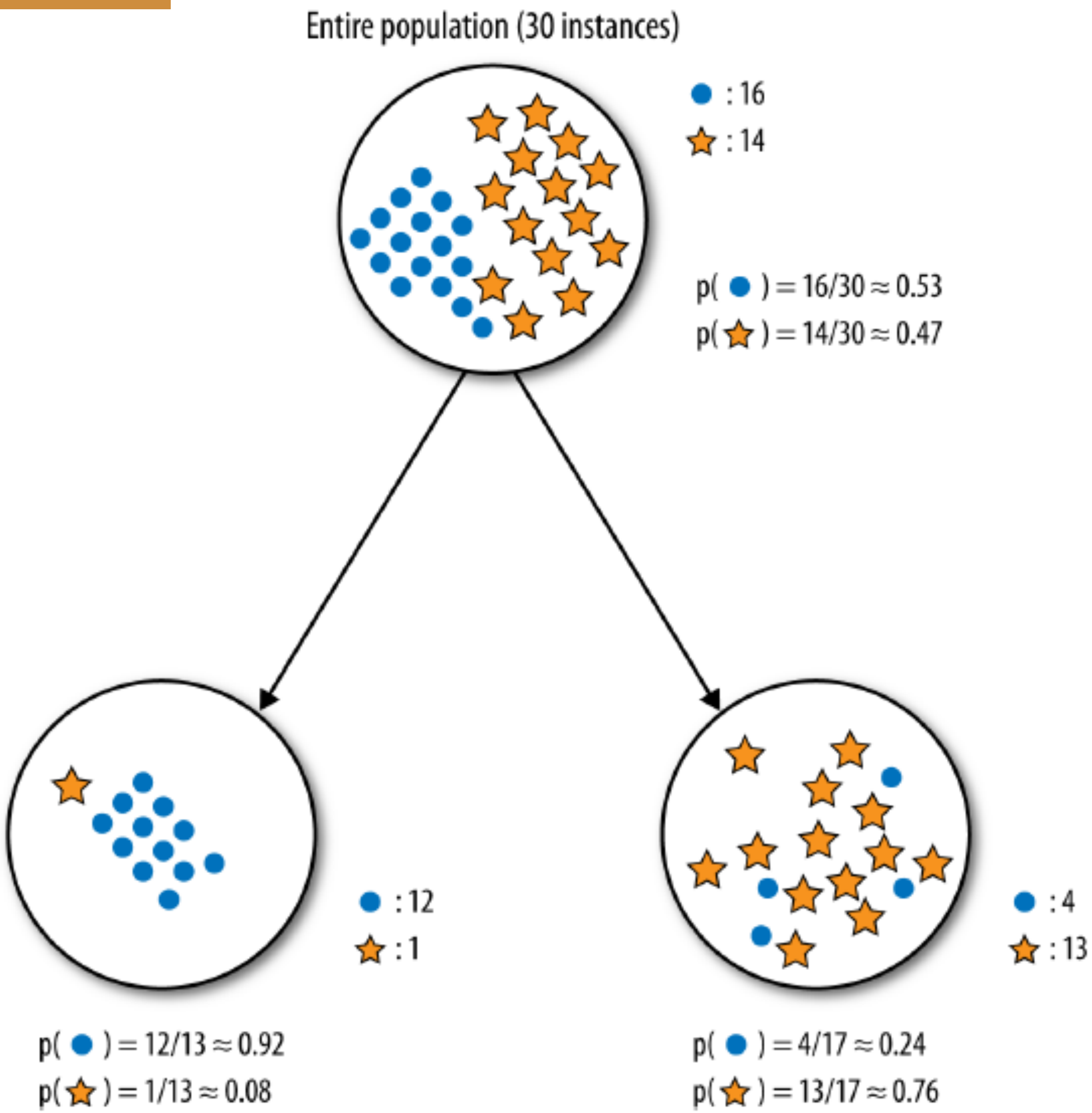
Entropy is a measure of disorder in a set S. Let p_i be the % of observations in S belonging to category i, for $i = 1, \dots, n$. The entropy of S is given by

$$E(S) = -p_1 \log p_1 - p_2 \log p_2 - \cdots - p_n \log n.$$

If the **parent set** S consisting of m records is split into k **children sets** $C_1, \dots, C_k$ containing $q_1, \dots, q_k$ records (resp.), then the **information gain** from the split is given by

$$\mathrm{IG}(S; C) = E(S) - \frac{1}{m}[q_1 E(C_1) + \cdots + q_k E(C_k)].$$

[Provost & Fawcett, *Data Science for Business*]



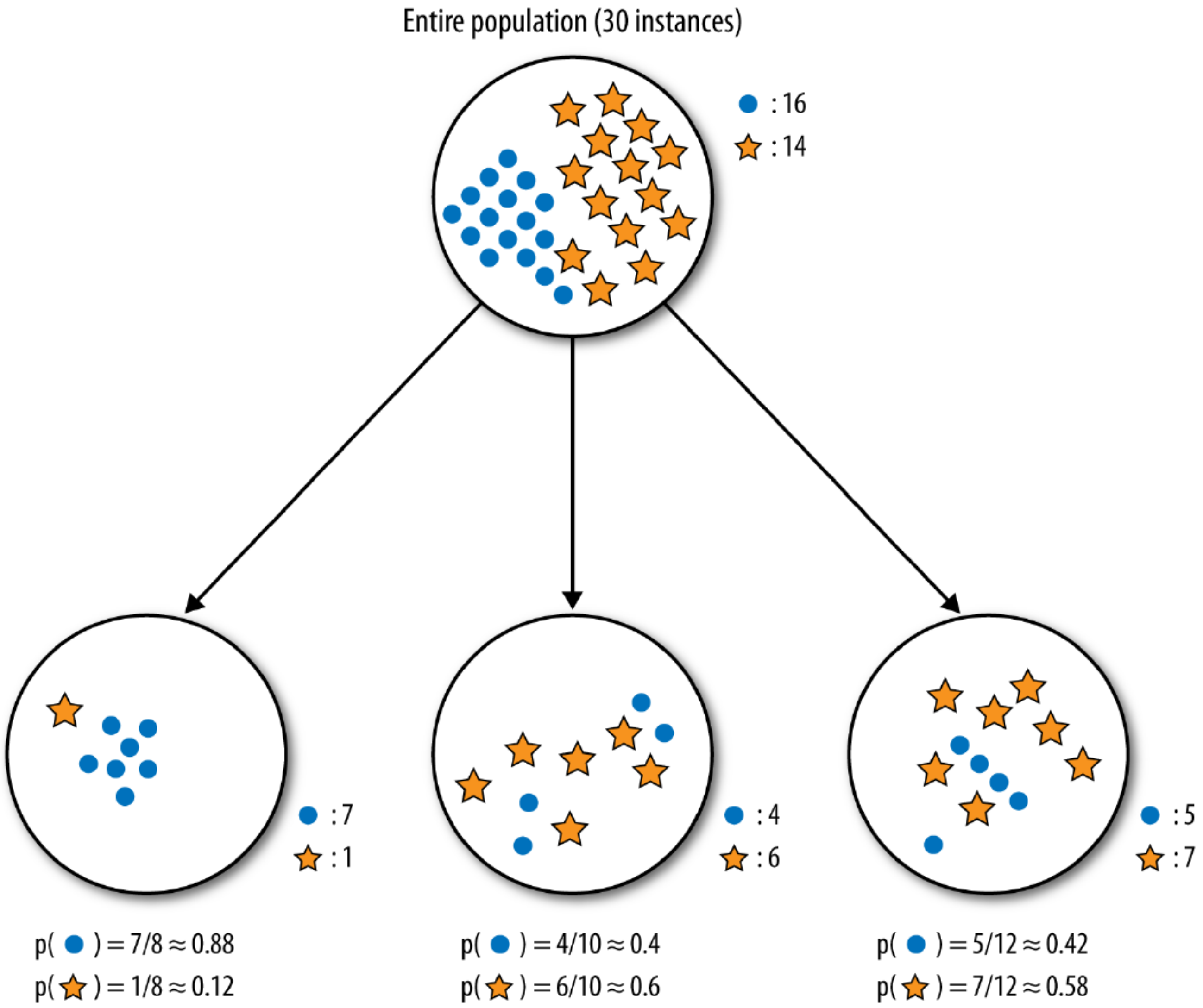
$$E(S) = -p_\circ \log p_\circ - p_* \log p_*$$
$$= -\frac{16}{30}\log\frac{16}{30} - \frac{14}{30}\log\frac{14}{30} \approx 0.99$$

$$E(L) = -p_\circ \log p_\circ - p_* \log p_*$$
$$= -\frac{12}{13}\log\frac{12}{13} - \frac{1}{13}\log\frac{1}{13} \approx 0.39$$

$$E(R) = -p_\circ \log p_\circ - p_* \log p_*$$
$$= -\frac{4}{17}\log\frac{4}{17} - \frac{13}{17}\log\frac{13}{17} \approx 0.79$$

$$\mathrm{IG} = E(S) - \tfrac{1}{30}[q_L E(L) + q_R E(R)]$$
$$\approx 0.99 - \tfrac{1}{30}[13(0.39) + 17(0.79)]$$
$$\approx \mathbf{0.37}$$

[Provost & Fawcett, *Data Science for Business*]



$$E(S) = -p_\circ \log p_\circ - p_* \log p_* = -\frac{16}{30}\log\frac{16}{30} - \frac{14}{30}\log\frac{14}{30} \approx 0.99$$

$$E(L) = -p_\circ \log p_\circ - p_* \log p_* = -\frac{7}{8}\log\frac{7}{8} - \frac{1}{8}\log\frac{1}{8} \approx 0.54$$

$$E(C) = -p_\circ \log p_\circ - p_* \log p_* = -\frac{4}{10}\log\frac{4}{10} - \frac{6}{10}\log\frac{6}{10} \approx 0.97$$

$$E(R) = -p_\circ \log p_\circ - p_* \log p_* = -\frac{5}{12}\log\frac{5}{12} - \frac{7}{12}\log\frac{7}{12} \approx 0.98$$

$$\mathrm{IG} = E(S) - \frac{1}{30}[q_L E(L) + q_C E(C) + q_R E(R)] \approx 0.99 - \frac{1}{30}[8(0.54) + 10(0.97) + 12(0.98)] \approx \mathbf{0.13}$$

Decision Trees Strengths

White box model

- predictions can always be explained by following the appropriate paths

Can be used with **incomplete** datasets

Built-in feature selection

- less relevant features don't tend to be used as splitting features

Makes **no assumption** about

- independence, constant variance, underlying distributions, co-linearity

Decision Trees Limitations

Not as accurate as other algorithms (usually)

Not robust: small changes in the training dataset can lead to a completely different tree, with a completely different predictions

Particularly vulnerable to **overfitting** in the absence of **pruning**

- pruning procedures are typically convoluted

Optimal decision tree learning is **NP-complete**

Biased towards categorical features with **high** number of levels

Decision Trees Notes

Splitting metrics:

- information gain, Gini impurity, variance reduction, etc.

Common variants:

- Iterative Dichotomiser 3, C4.0, C4.5, CHAID, MARS, conditional inference trees, CART

Decision trees can also be combined together using boosting algorithms (**AdaBoost**) or **Random Forests**, providing a type of voting procedure (Ensemble Learning).

Suggested Reading

Decision Trees and Other Algorithms

Data Understanding, Data Analysis, Data Science
Volume 3: Spotlight on Machine Learning

19. Introduction to Machine Learning
    - 19.4 Classification and Regression
      - Classification Algorithms
      - Decision Tree
      - Toy Example: Kyphosis Dataset
    - 19.7 R Examples
      - Classification: Kyphosis Dataset

21. Focus on Classification and Supervised Learning
    - 21.2 Simple Classifiers
    - 21.3 Rare Occurrences
    - 21.4 Other Supervised Approaches
    - 21.5 Ensemble Learning

Exercises

Decision Trees and Other Algorithms

1. Go over the kyphosis classification example found in DUDADS (see suggested reading). Repeat the process with the titanic dataset (you may wish to visualize the dataset first) in order to build a decision tree that will help you determine if a passenger survived the sinking or not.

Exercises

Decision Trees and Other Algorithms

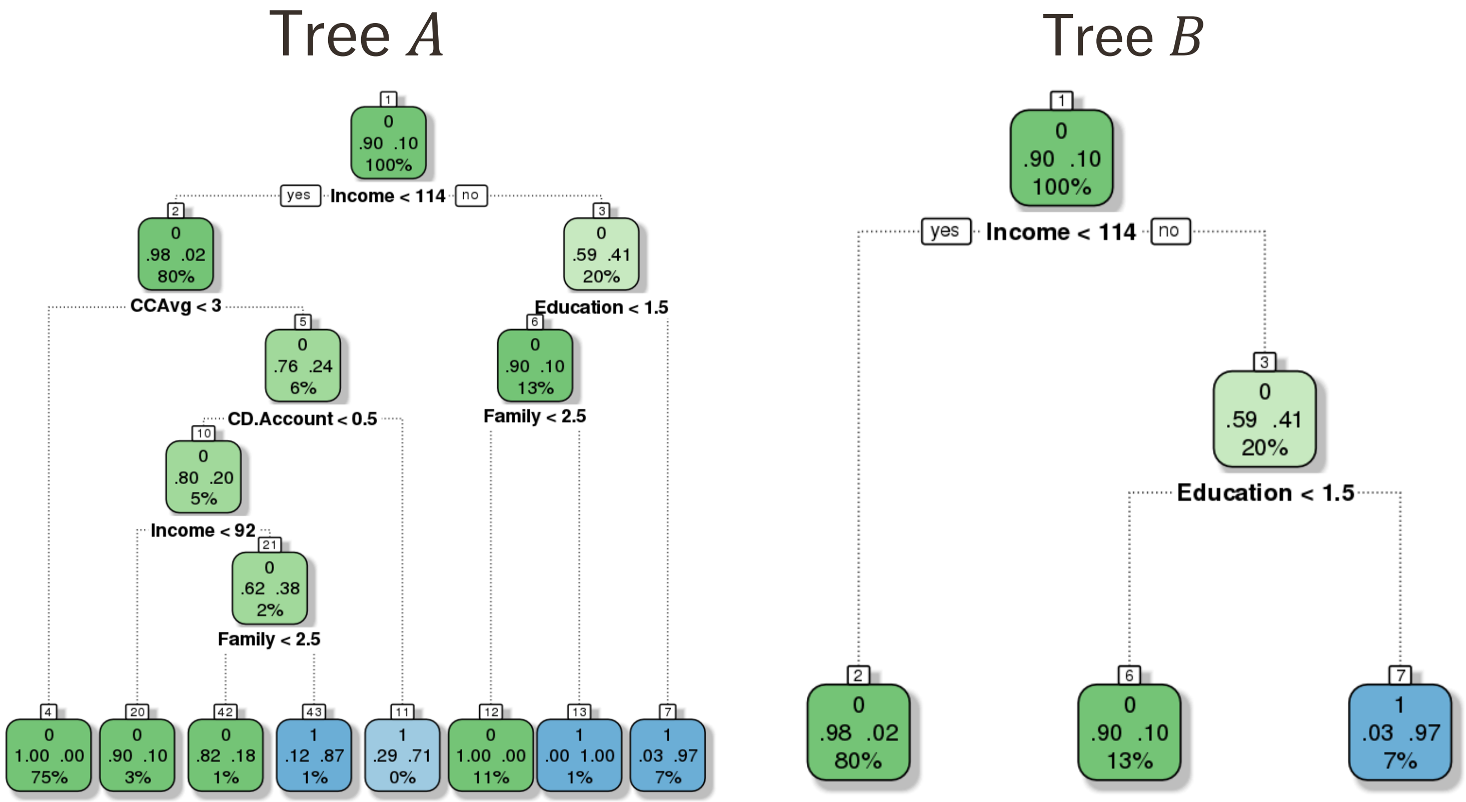
2. UniversalBank is looking at converting its **liability** customers (i.e., customers who only have deposits at the bank) into **asset** customers (i.e., customers who have a loan with the bank). In a previous campaign, *UniversalBank* was able to convert 9.6% of 5000 of its liability customers into asset customers. The marketing department would like to understand what combination of factors make a customer more likely to accept a personal loan, in order to better design the next conversion campaign.

   The dataset contains data on 5000 customers, including the following measurements: age, years of professional experience, yearly income (in $K), family size, value of mortgage with the bank, whether the client has a certificate of deposit with the bank, a credit card, etc.

Exercises

Decision Trees and Other Algorithms

2. (cont.) We build 2 decision trees on a training subset of 3000 records to predict whether a customer is likely to accept a personal loan (1) or not (0).



Exercises

Decision Trees and Other Algorithms

a. How many variables are used in the construction of tree A? Of tree B?

b. Is the following decision rule valid or not for tree A:
   IF (Income ≥ 114) AND (Education ≥ 1.5)
   THEN (Personal Loan $= 1$)?

c. Is the following decision rule valid or not for tree B:
   IF (Income < 92) AND (CCAvg ≥ 3)
   AND (CD.Account < 0.5)
   THEN (Personal Loan $= 0$)?

d. What prediction would tree A make for a customer with:
   - yearly income of 94,000$USD (`Income = 94`),
   - 2 kids (`Family = 4`),
   - no certificate of deposit with the bank (`CD.Account = 0`),
   - a credit card interest rate of 3.2% (`CCAvg = 3.2`), and
   - a graduate degree in Engineering (`Education = 3`).

e. What about tree B?

Predicted

Actual

Classes	A	B	C	D	Total
A	50	10	30	20	110
B	15	20	30	15	80
C	20	10	30	40	100
D	15	15	30	50	110
Total	100	55	120	125	800

6. Performance Evaluation

Model Selection

As a consequence of the **No-Free-Lunch Theorem**, no single classifier can be the best performer for every problem.

Model selection must take into account:

- the **nature** of the available data
- the **relative frequencies of the classification sub-groups**
- the **stated classification goals**
- how easily the model lends itself to **interpretation** and **statistical analysis**
- how much **data preparation** is required

Model Selection

Model selection must take into account (continued):

- whether it can accommodate various data types and missing observations
- whether it performs well with large datasets, and
- whether it is **robust** against small data departures from theoretical assumptions.

Past success is not a guarantee of future success – it is the analyst's responsibility to try a **variety of models**.

But how can the "**best**" model be selected?

Classification Errors

When attempting to determine what kind of music a new customer would prefer, there is no real **cost** in making a mistake; if, on the other hand, the classifier attempts to determine the presence or absence of cancerous cells in lung tissue, mistakes are **more consequential**.

Several metrics can be used to assess a classifier's performance, depending on the context.

Binary classifiers are simpler and have been studied far longer than multi-level classifiers; consequently, a larger body of evaluation metrics is available for these classifiers.

Binary Classifiers

		Predicted: Category I	Predicted: Category II	Total
Actuals	Category I	TP	FN	AP
	Category II	FP	TN	AN
	Total	PP	PN	T

TP, TN, FP, FN: **True Positives**, **True Negatives**, **False Positives**, and **False Negatives**, respectively.

Perfect classifiers would have $FP, FN = 0$, but that rarely ever happens in practice (and not ideal, in a way).

Metrics:

- sensitivity = $TP/(TP + FN)$
- specificity = $TN/(FP + TN)$
- precision = $TP/(TP + FP)$
- recall = $TP/(TP + FN)$
- negative predictive value = $TN/(TN + FN)$
- false positive rate = $FP/(FP + TN)$
- false discovery rate = $FP/(FP + TP)$
- false negative rate = $FP/(FN + TP)$
- accuracy = $(TP + TN)/T$

Other metrics:

F_1-score, ROC AUC, informedness, markedness, Matthews' Correlation Coefficient (MCC), etc.

Session 2

		Predicted A	Predicted B	Total	
Actuals	A	54	10	64	79.0%
	B	6	11	17	21.0%
Total		60	21	81	
		74.1%	25.9%		

Classification Rates	
Sensitivity:	0.84
Specificity:	0.65
Precision:	0.90
Negative Predictive Value:	0.52
False Positive Rate:	0.35
False Discovery Rate:	0.10
False Negative Rate:	0.16

Performance Metrics	
Accuracy:	0.80
F1-Score:	0.87
Informedness (ROC):	0.49
Markedness:	0.42
M.C.C.:	0.46
Pearson's chi2:	0.01
Hist. Stat:	0.10

		Predicted A	Predicted B	Total	
Actuals	A	54	0	54	66.7%
	B	16	11	27	33.3%
Total		70	11	81	
		86.4%	13.6%		

Classification Rates	
Sensitivity:	1.00
Specificity:	0.41
Precision:	0.77
Negative Predictive Value:	1.00
False Positive Rate:	0.59
False Discovery Rate:	0.23
False Negative Rate:	0.00

Performance Metrics	
Accuracy:	0.80
F1-Score:	0.87
Informedness (ROC):	0.41
Markedness:	0.77
M.C.C.:	0.56
Pearson's chi2:	0.33
Hist. Stat:	0.40

Both classifiers have an accuracy of 80%; the second classifier makes some wrong predictions for A, but never for B; the first classifier makes mistakes for both classes. The second classifier mistakenly predicts occurrence A as B on 16 occasions, but the first one only does so 6 times. Which one is best depends on the **cost of misclassification**.

Multi-Level Classifiers

It is preferable to select metrics that generalize more readily to **multi-level classifiers.**

Accuracy: proportion of correct predictions amid all the observations

- value ranges from 0% to 100%
- the higher the accuracy, the better the match
- a predictive model with high accuracy may be useless thanks to the **Accuracy Paradox**

Matthews Correlation Coefficient (MCC): useful even when the classes are of very different sizes

- correlation coefficient between actual and predicted classifications
- range varies from -1 to 1
- if $\mathrm{MCC} = 1$, predicted and actual responses are identical
- if $\mathrm{MCC} = 0$, the classifier performs no better than a random prediction ("flip of a coin").

Multi-Level Classifiers

MCC: 69.7%
Accuracy: 78.3%
Pearson: 0.13161
Hist: 30.0%

			Predicted							
			Maltreatment			Risk				
			Unfounded	Suspected	Substantiated	No	Yes	Unknown	*Total*	
Actuals	Maltreatment	Unfounded	4,577	-	-	198	6	-	*4,781*	*29.2%*
		Suspected	-	965	-	29	2	-	*995*	*6.1%*
		Substantiated	-	-	6,187	116	35	2	*6,339*	*38.7%*
	Risk	No	894	-	763	949	19	9	*2,632*	*16.1%*
		Yes	123	-	520	122	111	5	*880*	*5.4%*
		Unknown	212	-	303	184	21	24	*745*	*4.6%*
		Total	*5,805*	*965*	*7,772*	*1,597*	*194*	*40*	***16,372***	
			35.5%	*5.9%*	*47.5%*	*9.8%*	*1.2%*	*0.2%*		

Regression Performance Evaluation

For numerical targets y with predictions $\hat{y}$, metrics include:

- **mean squared** and **mean absolute errors**

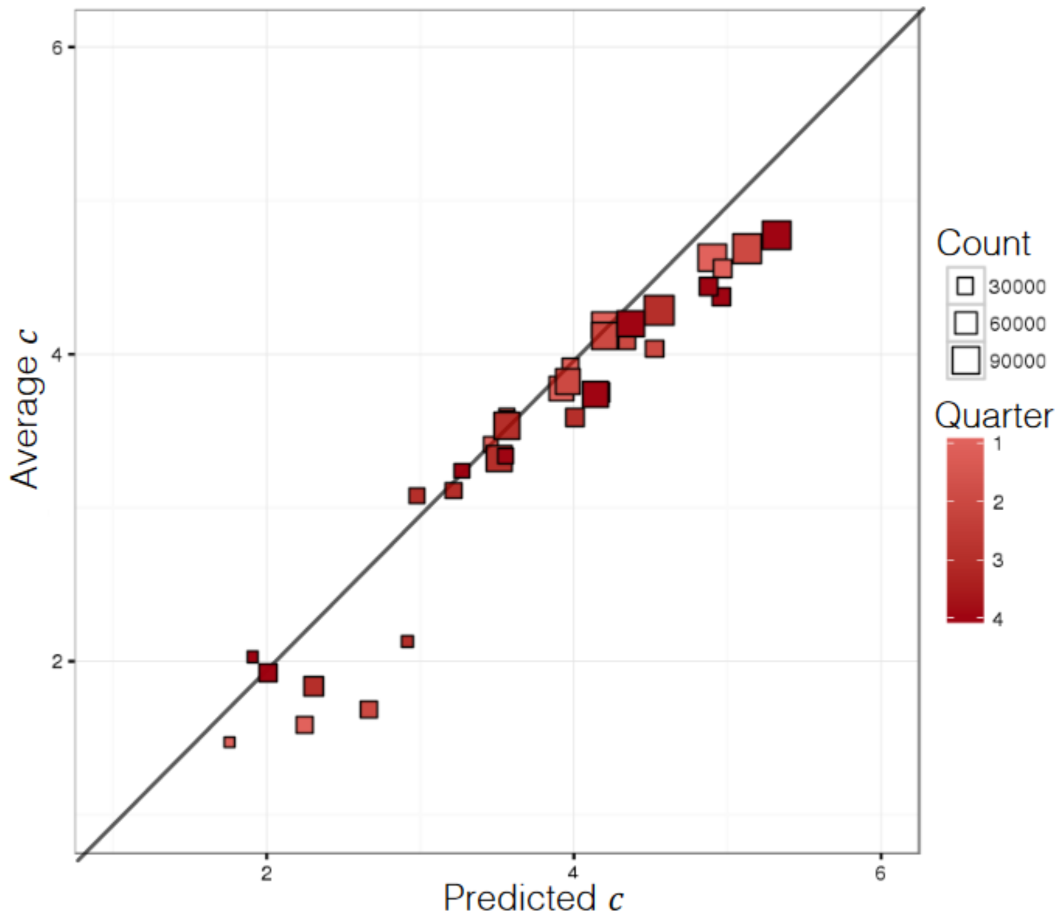
$$\mathrm{MSE} = \mathrm{mean}\{(\hat{y}_i - y_i)^2\},\ \mathrm{MAE} = \mathrm{mean}\{|\hat{y}_i - y_i|\}$$

- **normalized mean squared** and **normalized mean absolute errors**

$$\mathrm{NMSE} = \frac{\mathrm{mean}\{(\hat{y}_i - y_i)^2\}}{\mathrm{mean}\{(\bar{y} - y_i)^2\}},\mathrm{NMAE} = \frac{\mathrm{mean}\{|\hat{y}_i - y_i|\}}{\mathrm{mean}\{|\bar{y} - y_i|\}}$$

- **mean average percentage error** $\mathrm{MAPE} = \mathrm{mean}\left\{\frac{|\hat{y}_i - y_i|}{y_i}\right\}$
- **correlation** $\rho_{\hat{y},y}$
- etc.

Regression Performance Evaluation



Suggested Reading

Performance Evaluation

Data Understanding, Data Analysis, Data Science
Volume 3: Spotlight on Machine Learning

19. Introduction to Machine Learning
    19.4 Classification and Regression
    - Performance Evaluation

20. Regression and Value Estimation
    20.1 Statistical Learning
    - Model Evaluation

21. Focus on Classification and Supervised Learning
    21.1 Overview
    - Model Evaluation

Exercises

Performance Evaluation

We continue the UniversalBank example. The confusion matrices for the predictions of trees A and B on the remaining 2000 testing observations are shown below.

1. Using the appropriate matrices, compute the performance evaluation metrics for each of the trees (on the testing set).
2. If customers who would not accept a personal loan get irritated when offered a personal loan, what tree should *the* marketing group use to maintain good customer relations?

Tree A

		Predicted A	Predicted B	Total	
Actuals	A	1792	19	1811	90.55%
	B	18	171	189	9.45%
Total		1810	190	2000	
		90.50%	9.50%		

Tree B

		Predicted A	Predicted B	Total	
Actuals	A	1801	10	1811	90.55%
	B	64	125	189	9.45%
Total		1865	135	2000	
		93.25%	6.75%		