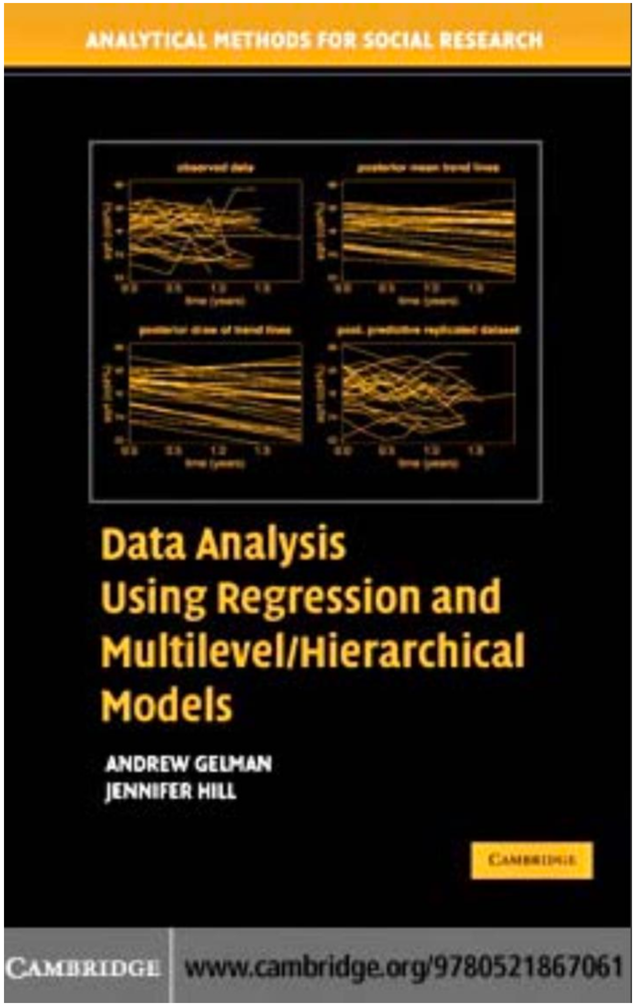
Lecture 5:

Classification

Homework (after Eastern)

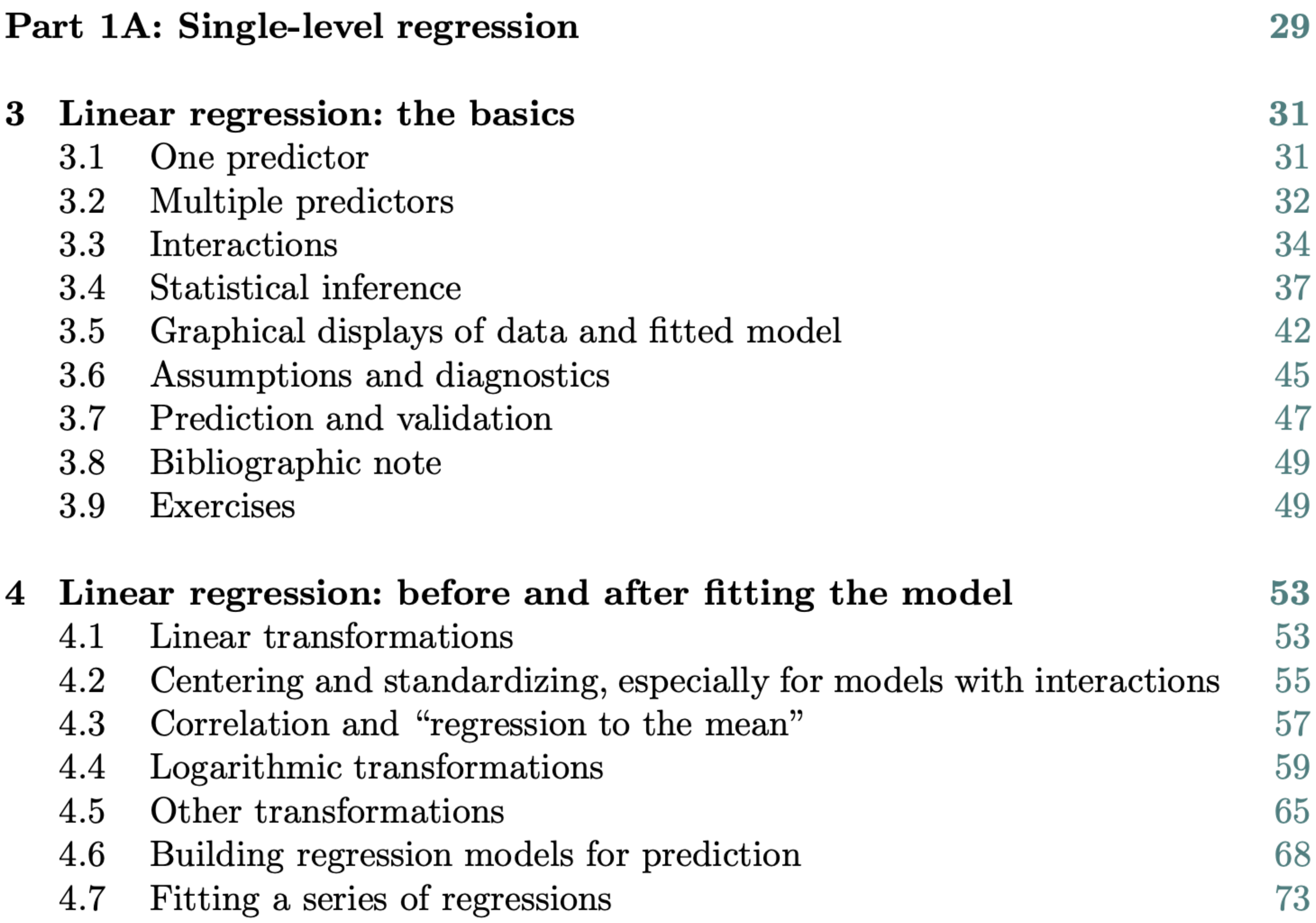


ANALYTICAL METHODS FOR SOCIAL RESEARCH
observed data
posterior mean trend lines
posterior draw of trend lines
post. predictive replicated dataset
time (years)
Data Analysis Using Regression and Multilevel/Hierarchical Models
ANDREW GELMAN
JENNIFER HILL
CAMBRIDGE
CAMBRIDGE
www.cambridge.org/9780521867061

Homework (after Eastern)

Passwords

Homework (after Eastern)



Part 1A: Single-level regression **29**

3 Linear regression: the basics **31**

3.1 One predictor 31

3.2 Multiple predictors 32

3.3 Interactions 34

3.4 Statistical inference 37

3.5 Graphical displays of data and fitted model 42

3.6 Assumptions and diagnostics 45

3.7 Prediction and validation 47

3.8 Bibliographic note 49

3.9 Exercises 49

4 Linear regression: before and after fitting the model **53**

4.1 Linear transformations 53

4.2 Centering and standardizing, especially for models with interactions 55

4.3 Correlation and “regression to the mean” 57

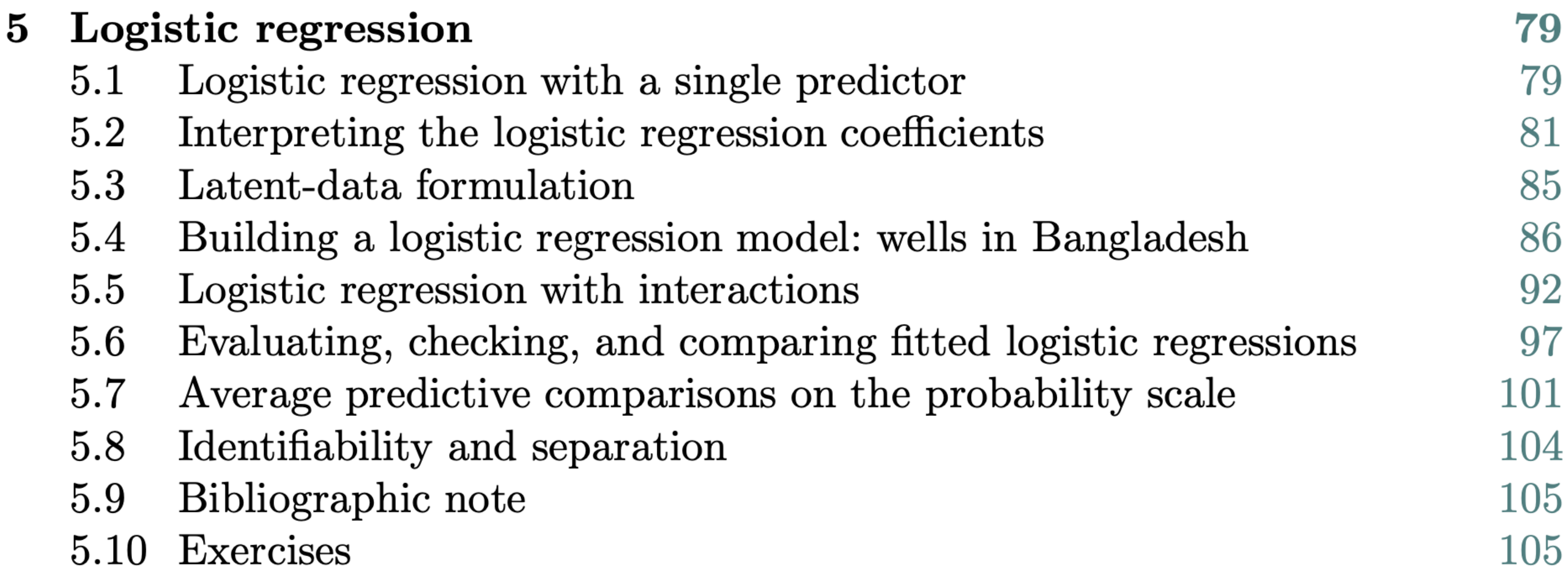
4.4 Logarithmic transformations 59

4.5 Other transformations 65

4.6 Building regression models for prediction 68

4.7 Fitting a series of regressions 73

Homework (after Eastern)



5 Logistic regression **79**

5.1 Logistic regression with a single predictor 79

5.2 Interpreting the logistic regression coefficients 81

5.3 Latent-data formulation 85

5.4 Building a logistic regression model: wells in Bangladesh 86

5.5 Logistic regression with interactions 92

5.6 Evaluating, checking, and comparing fitted logistic regressions 97

5.7 Average predictive comparisons on the probability scale 101

5.8 Identifiability and separation 104

5.9 Bibliographic note 105

5.10 Exercises 105

Classification fundamentals

Classification: definition

Given

- a collection of class labels
- a collection of data objects labelled with a class label

Find a descriptive profile of each class, which will allow the assignment of unlabeled objects to the appropriate class

Definitions

Training set

Collection of labeled data objects used to learn the classification model

Test set

Collection of labeled data objects used to validate the classification model

Classification techniques

1. Decision trees
2. Classification rules
3. Association rules
4. Neural Networks
5. Naïve Bayes and Bayesian Networks
6. k-Nearest Neighbours (k-NN)
7. Support Vector Machines (SVM)

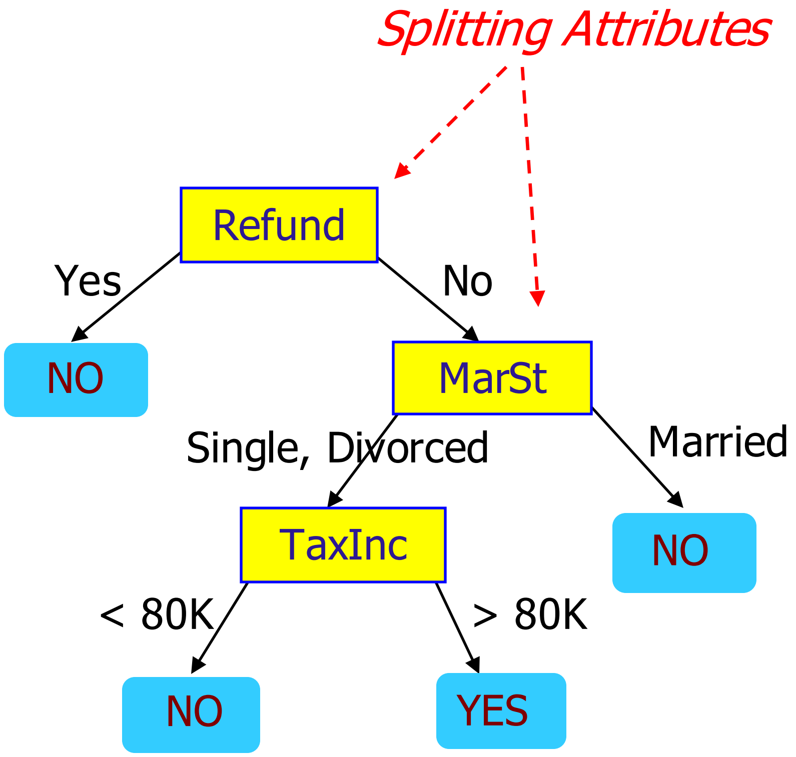
…

Decision trees

Example of decision tree

	categorical	categorical	continuous	class
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**

Training Data



Splitting Attributes

Refund

Yes

No

NO

MarSt

Single, Divorced

Married

TaxInc

NO

< 80K

> 80K

NO

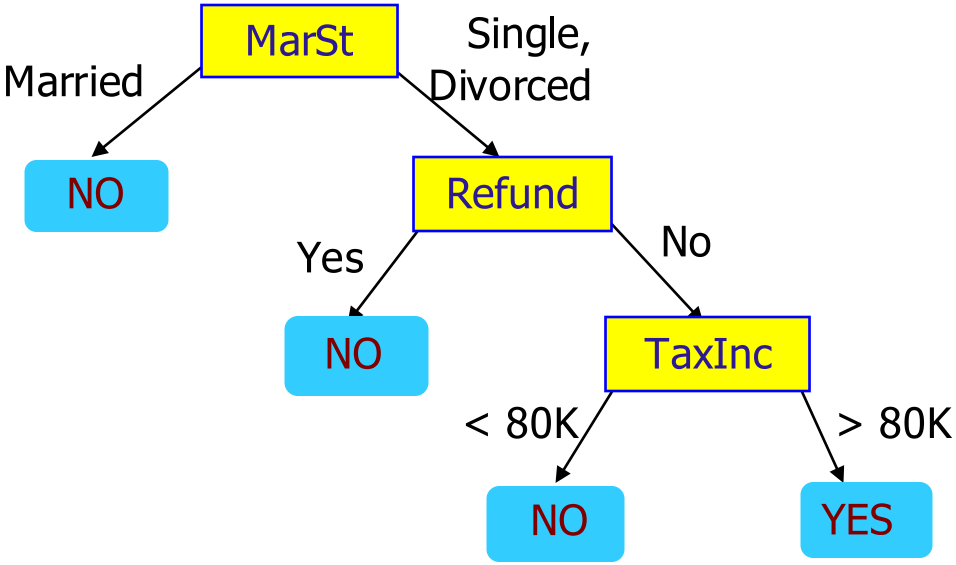
YES

Model: Decision Tree

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

Another example of decision tree

	categorical	categorical	continuous	class
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**



There could be more than one tree that fits the same data!

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

Decision tree induction

Many algorithms to build a decision tree

- Hunt's Algorithm (one of the earliest)
- CART
- ID3, C4.5, C5.0
- SLIQ, SPRINT

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

Decision Tree Based Classification

Advantages

- Inexpensive to construct
- Extremely fast at classifying unknown records
- Easy to interpret for small-sized trees
- Accuracy is comparable to other classification techniques for many simple data sets

Disadvantages

- accuracy may be affected by missing data

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

Evaluation of decision trees

Accuracy

For simple datasets, comparable to other classification techniques

Interpretability

Model is interpretable for small trees

Single predictions are interpretable

Incrementality

Not incremental

- Efficiency
  - Fast model building
  - Very fast classification
- Scalability
  - Scalable both in training set size and attribute number
- Robustness
  - Difficult management of missing data

Random Forest

Ensemble learning technique

- multiple base models are combined
  - to improve accuracy and stability
  - to avoid overfitting

Random forest = set of decision trees

- a number of decision trees are built at training time
- the class is assigned by majority voting

Bibliography: Hastie, Tibshirani, Friedman, The Elements of Statistical Learning, Springer, 2009

Random Forest

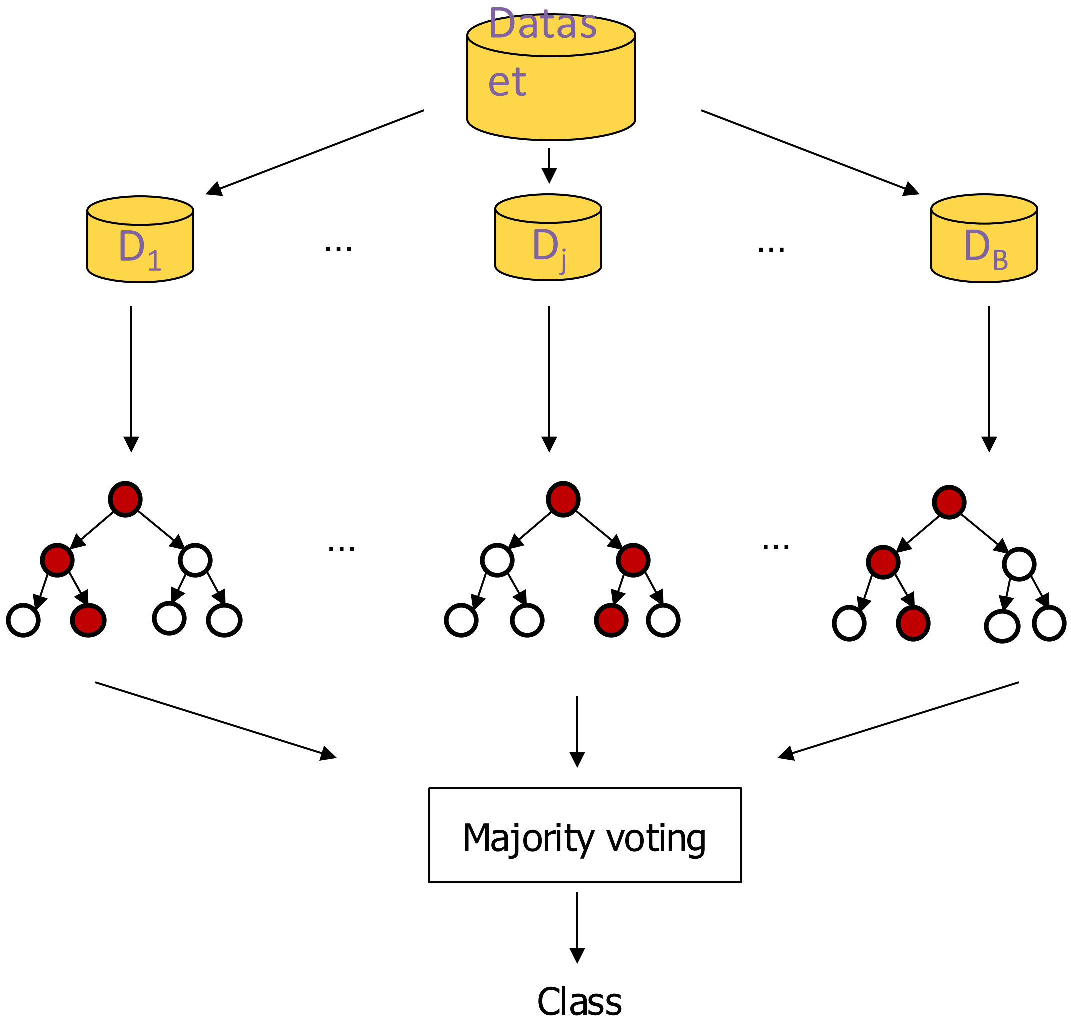
Original Training data

Random subsets

Multiple decision trees

For each subset, a tree is learned on a *random* set of features

Aggregating classifiers



Bootstrap aggregation

Given a training set *D* of *n* instances, it selects B times a *random* sample with replacement from D and trains trees on these dataset samples

```
For b = 1, ..., B
    Sample with replacement $n'$ training examples, $n' \leq n$
        A dataset subset $D_b$ is generated
    Train a classification tree on $D_b$
```

Random Forest – Algorithm Recap

- Given a training set D of n instances with p features
- For b = 1, ..., B
  - Sample randomly with replacement n' training examples. A subset D_b is generated
  - Train a classification tree on D_b
    - During the tree construction, for each candidate split
      - $m \ll p$ random features are selected (typically m ≈ $\sqrt{p}$)
      - the best split is computed among these m features
- Class is assigned by majority voting among the B predictions

Random Forest

Strong points

- higher accuracy than decision trees
- fast training phase
- robust to noise and outliers
- provides global feature importance, i.e. an estimate of which features are important in the classification

Weak points

- results can be difficult to interpret
  - A prediction is given by hundreds of trees
    - but at least we have an indication through feature importance

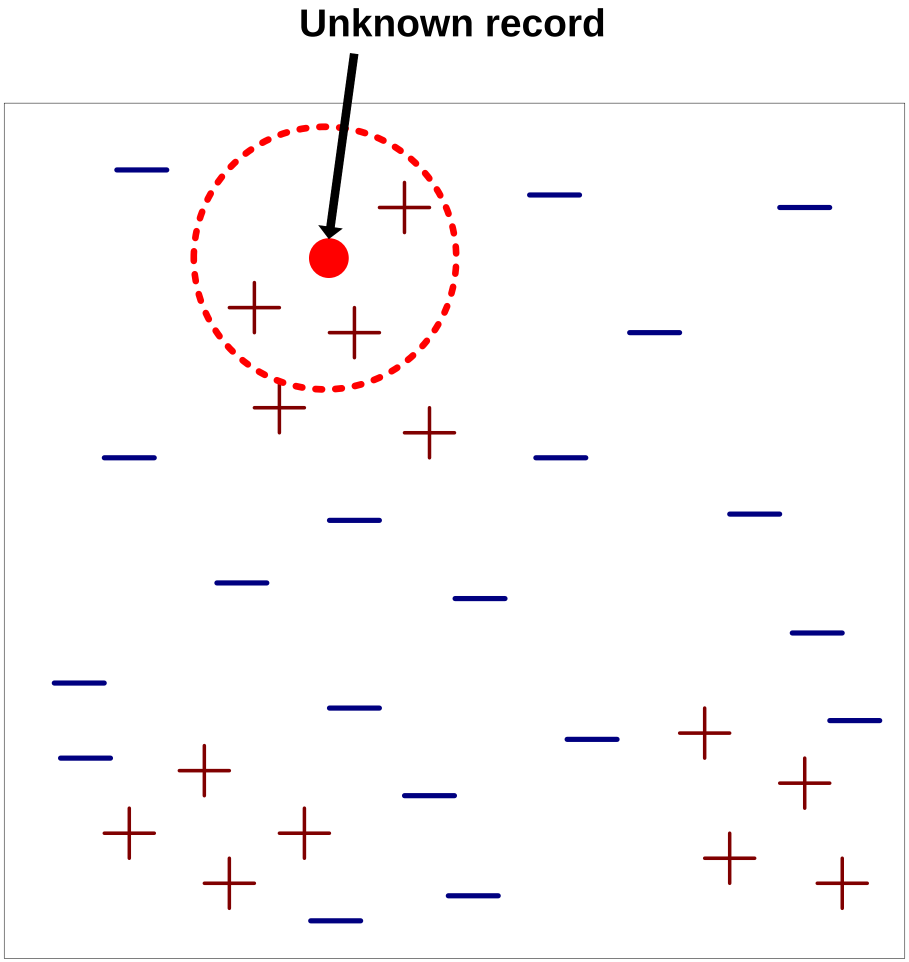
Evaluation of random forests

- Accuracy
  - Higher than decision trees
- Interpretability
  - Model and prediction are not interpretable
    - A prediction may be given by hundreds of trees
  - Provide global feature importance
    - an estimate of which features are important in the classification
- Incrementality
  - Not incremental

- Efficiency
  - Fast model building
  - Very fast classification
- Scalability
  - Scalable both in training set size and attribute number
- Robustness
  - Robust to noise and outliers

K-Nearest Neighbor

Nearest-Neighbor Classifiers

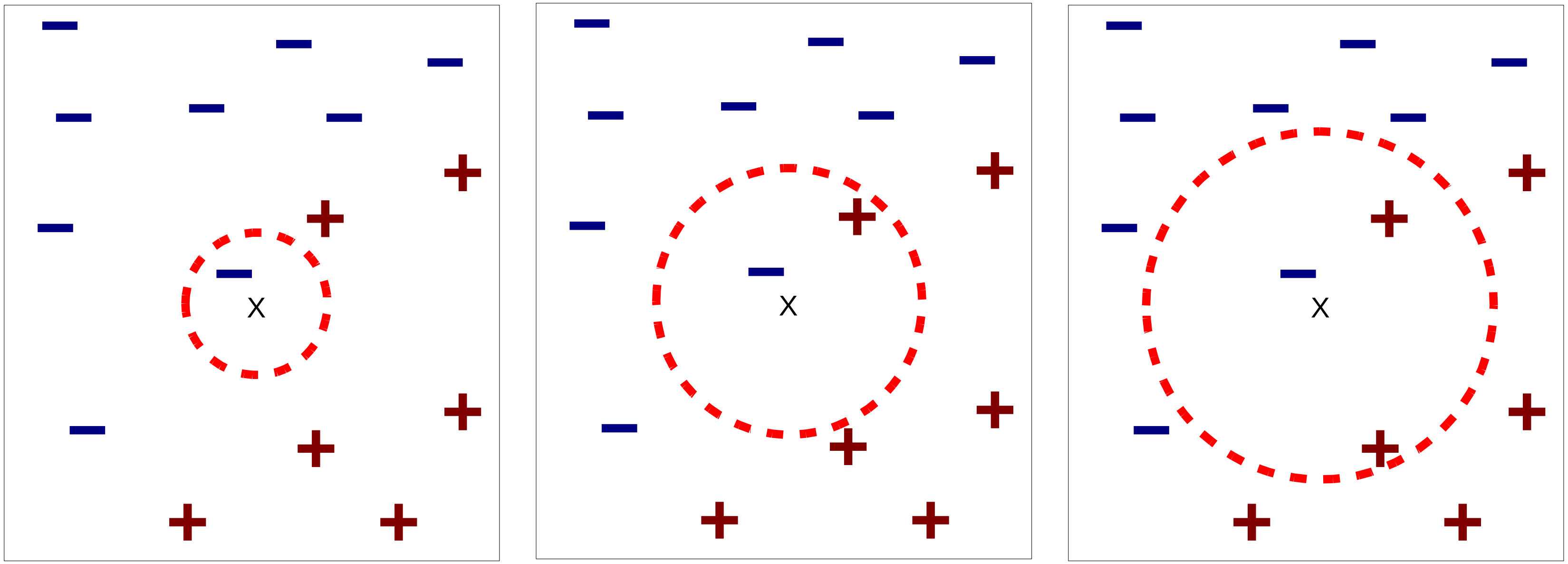


- Requires
  - The set of stored records
  - Distance Metric to compute distance between records
  - The value of *k*, the number of nearest neighbors to retrieve

- To classify an unknown record
  - Compute distance to other training records
  - Identify *k* nearest neighbors
  - Use class labels of nearest neighbors to determine the class label of unknown record (e.g., by taking majority vote)

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

Definition of Nearest Neighbor



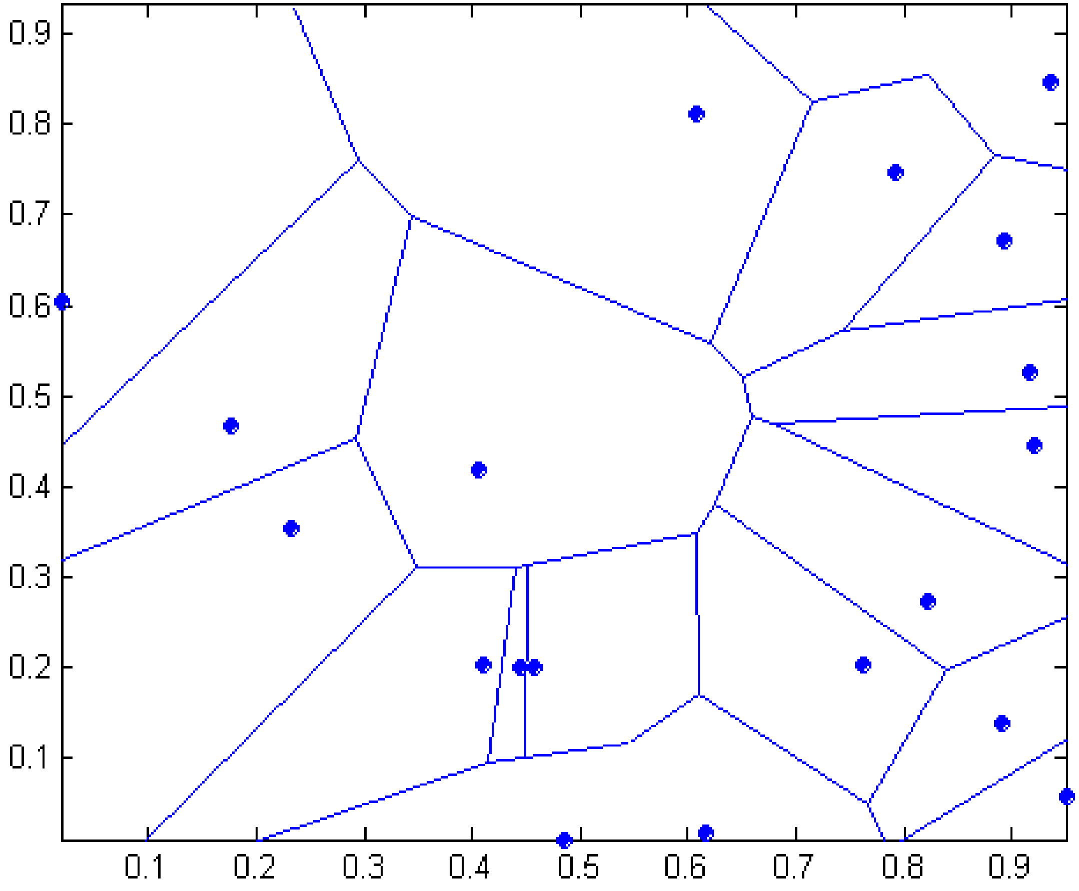
(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

K-nearest neighbors of a record x are data points that have the k smallest distance to x

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

1 nearest-neighbor

Voronoi Diagram



From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

Nearest Neighbor Classification

Compute distance between two points

- Euclidean distance

$$d(p,q)=\sqrt{\sum_i (p_i - q_i)^2}$$

Determine the class from nearest neighbor list

- take the majority vote of class labels among the k-nearest neighbors
- Weigh the vote according to distance
  - weight factor, $w = 1/d^2$

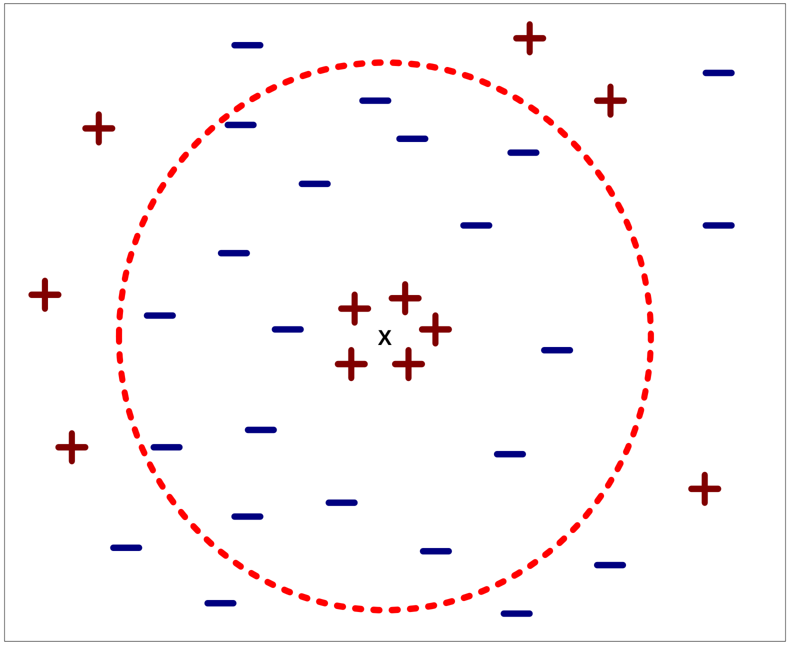
From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

Nearest Neighbor Classification

Choosing the value of k:

If k is too small, sensitive to noise points

If k is too large, neighborhood may include points from other classes



From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

Nearest Neighbor Classification

Scaling issues

Attribute domain should be normalized to prevent distance measures from being dominated by one of the attributes

Example: height [1.5m to 2.0m] vs. income [$10K to $1M]

Problem with distance measures

High dimensional data

curse of dimensionality

Evaluation of KNN

Accuracy

Comparable to other classification techniques for simple datasets

Interpretability

Model is not interpretable

Single predictions can be ”described” by neighbors

Incrementality

Incremental

Training set *must* be available

- Efficiency
  - (Almost) no model building
  - Slower classification, requires computing distances
- Scalability
  - Weakly scalable in training set size
  - Curse of dimensionality for increasing attribute number
- Robustness
  - Depends on distance computation

Bayesian Classification (FINISH POINT)

Elena Baralis

Politecnico di Torino

Bayes theorem

Let C and X be random variables

$$P(C,X) = P(C|X)\ P(X)$$

$$P(C,X) = P(X|C)\ P(C)$$

Hence

$$P(C|X)\ P(X) = P(X|C)\ P(C)$$

and also

$$P(C|X) = P(X|C)\ P(C)\ /\ P(X)$$

Bayesian classification: Example

Outlook	Temperature	Humidity	Windy	Class
sunny	hot	high	false	N
sunny	hot	high	true	N
overcast	hot	high	false	P
rain	mild	high	false	P
rain	cool	normal	false	P
rain	cool	normal	true	N
overcast	cool	normal	true	P
sunny	mild	high	false	N
sunny	cool	normal	false	P
rain	mild	normal	false	P
sunny	mild	normal	true	P
overcast	mild	high	true	P
overcast	hot	normal	false	P
rain	mild	high	true	N

From: Han, Kamber,"Data mining; Concepts and Techniques", Morgan Kaufmann 2006

Bayesian classification: Example

outlook	
P(sunny|p) = 2/9	P(sunny|n) = 3/5
P(overcast|p) = 4/9	P(overcast|n) = 0
P(rain|p) = 3/9	P(rain|n) = 2/5
temperature	
P(hot|p) = 2/9	P(hot|n) = 2/5
P(mild|p) = 4/9	P(mild|n) = 2/5
P(cool|p) = 3/9	P(cool|n) = 1/5
humidity	
P(high|p) = 3/9	P(high|n) = 4/5
P(normal|p) = 6/9	P(normal|n) = 1/5
windy	
P(true|p) = 3/9	P(true|n) = 3/5
P(false|p) = 6/9	P(false|n) = 2/5

P(p) = 9/14
P(n) = 5/14

From: Han, Kamber,"Data mining; Concepts and Techniques", Morgan Kaufmann 2006

Bayesian classification: Example

Data to be labeled

$$X = \langle \text{rain, hot, high, false} \rangle$$

For class p

$$P(X|p)\cdot P(p) =$$

$= P(rain|p)\cdot P(hot|p)\cdot P(high|p)\cdot P(false|p)\cdot P(p) =$

$3/9\cdot 2/9\cdot 3/9\cdot 6/9\cdot 9/14 = 0.010582$

For class n

$$P(X|n)\cdot P(n) =$$

$= P(rain|n)\cdot P(hot|n)\cdot P(high|n)\cdot P(false|n)\cdot P(n) =$

$2/5\cdot 2/5\cdot 4/5\cdot 2/5\cdot 5/14 = 0.018286$

From: Han, Kamber,"Data mining; Concepts and Techniques", Morgan Kaufmann 2006

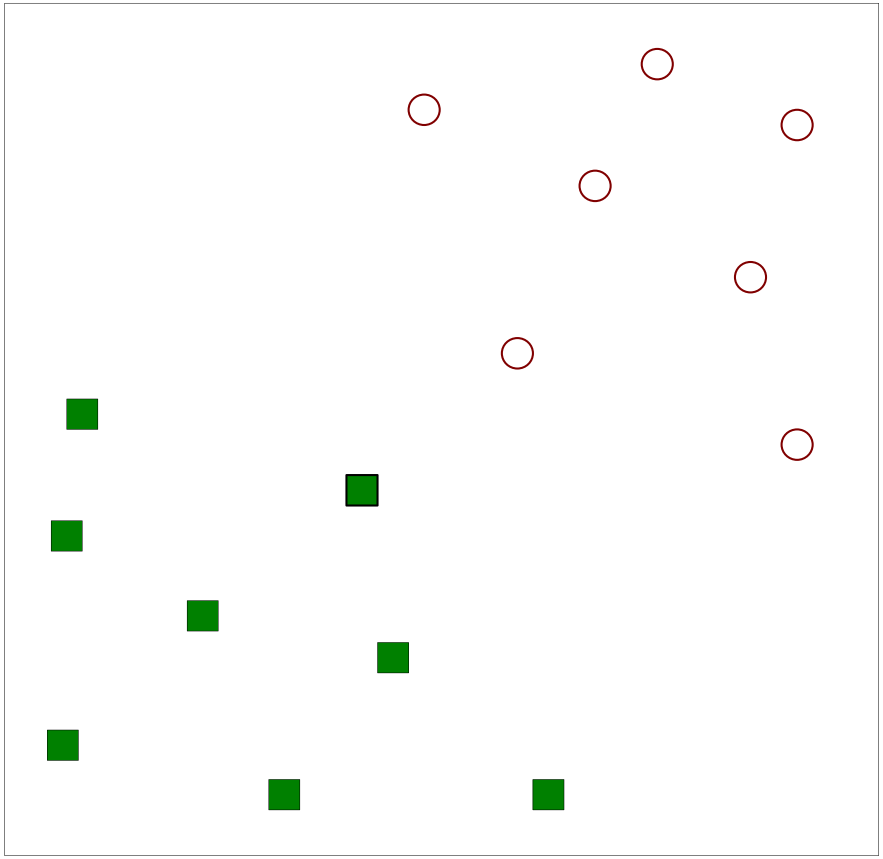
Evaluation of Naïve Bayes Classifiers

- Accuracy
  - Similar or lower than decision trees
    - Naïve hypothesis simplifies model
- Interpretability
  - Model and prediction are not interpretable
    - The weights of contributions in a single prediction may be used to explain
- Incrementality
  - Fully incremental
  - Does *not* require availability of training data

- Efficiency
  - Fast model building
  - Very fast classification
- Scalability
  - Scalable both in training set size and attribute number
- Robustness
  - Affected by attribute correlation

Support Vector Machines

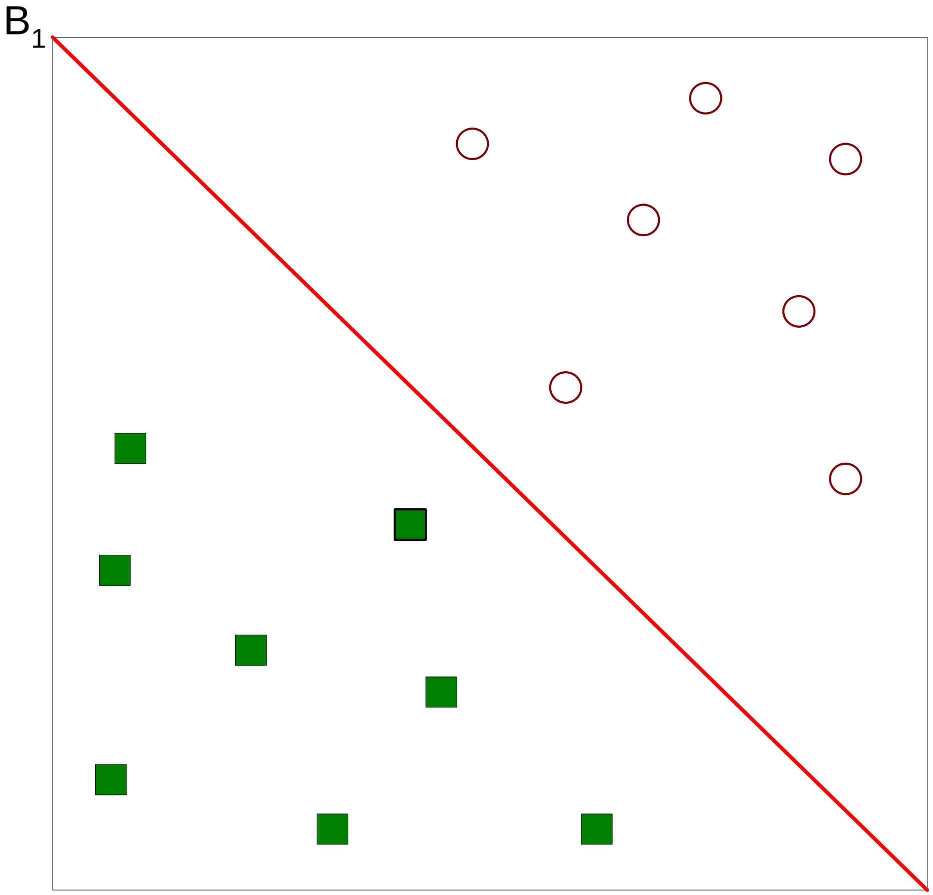
Support Vector Machines



Find a linear hyperplane (decision boundary) that will separate the data

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

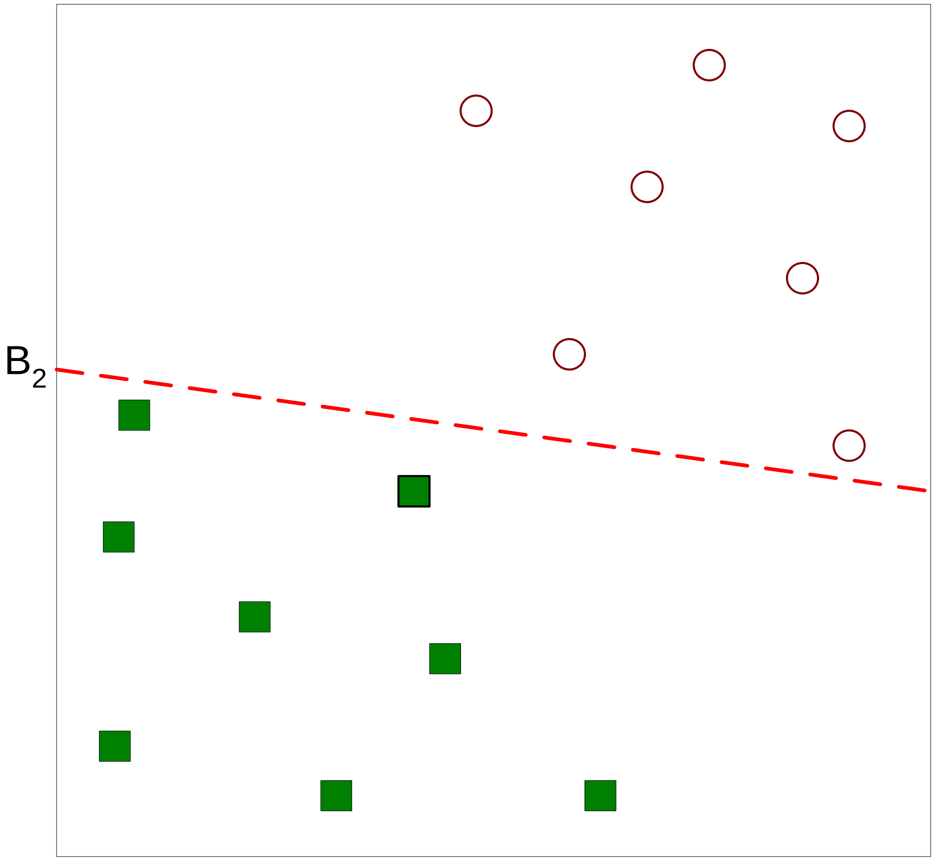
Support Vector Machines



One Possible Solution

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

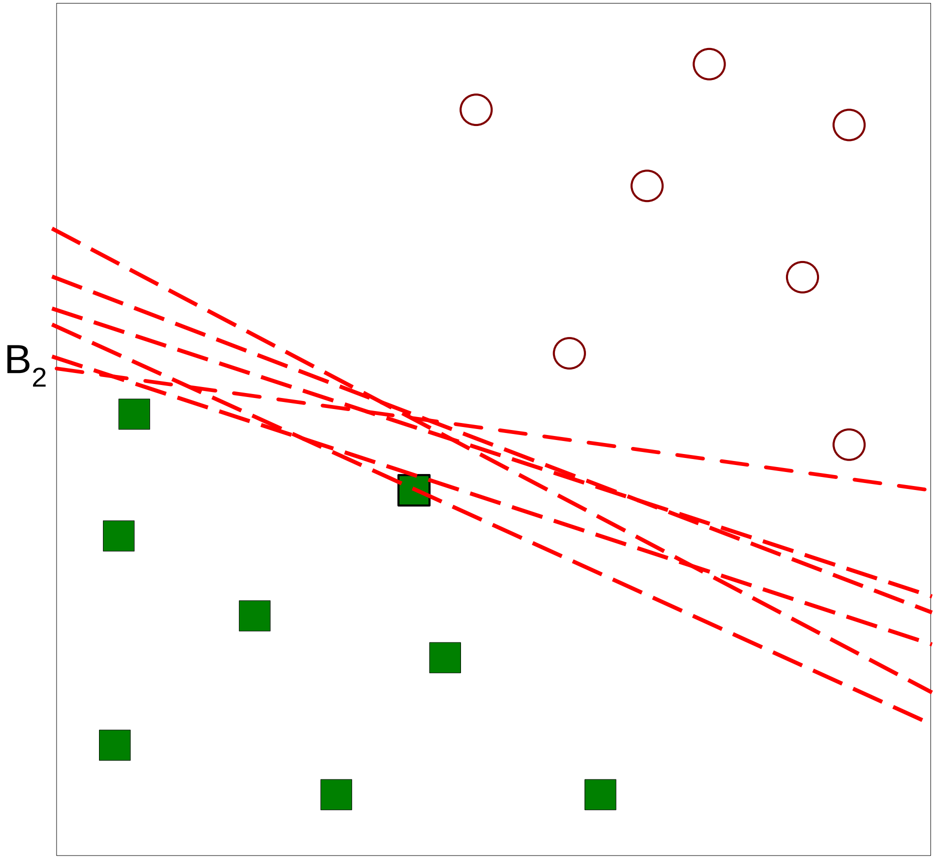
Support Vector Machines



Another possible solution

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

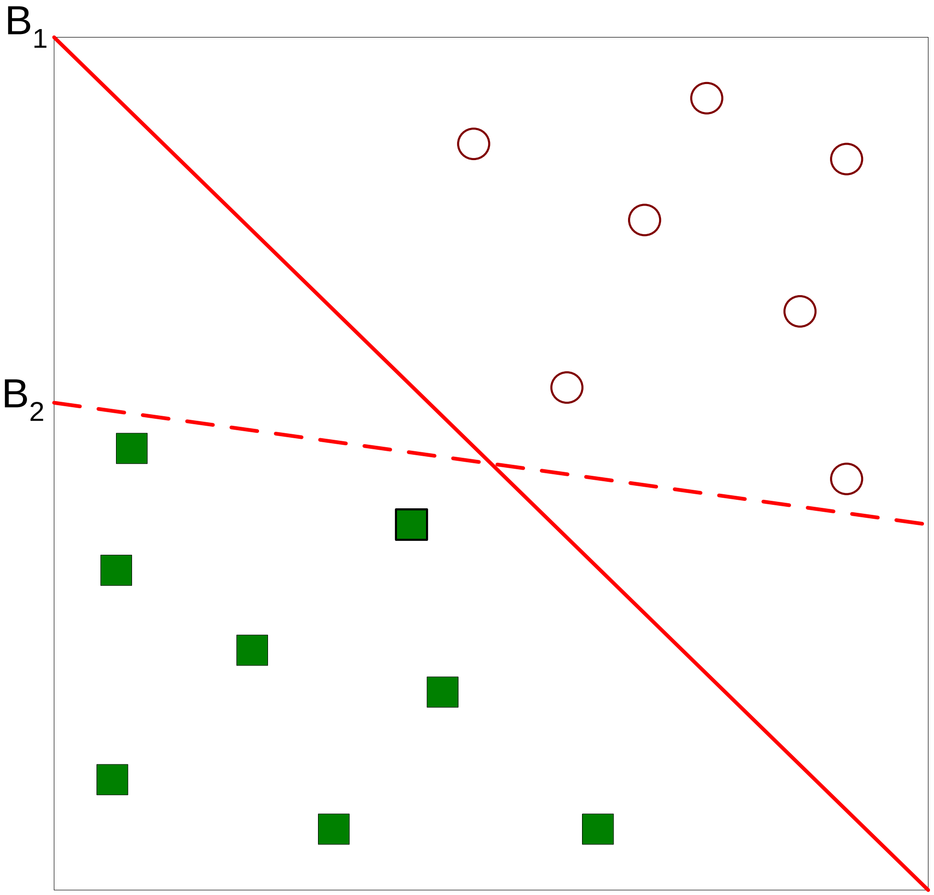
Support Vector Machines



Other possible solutions

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

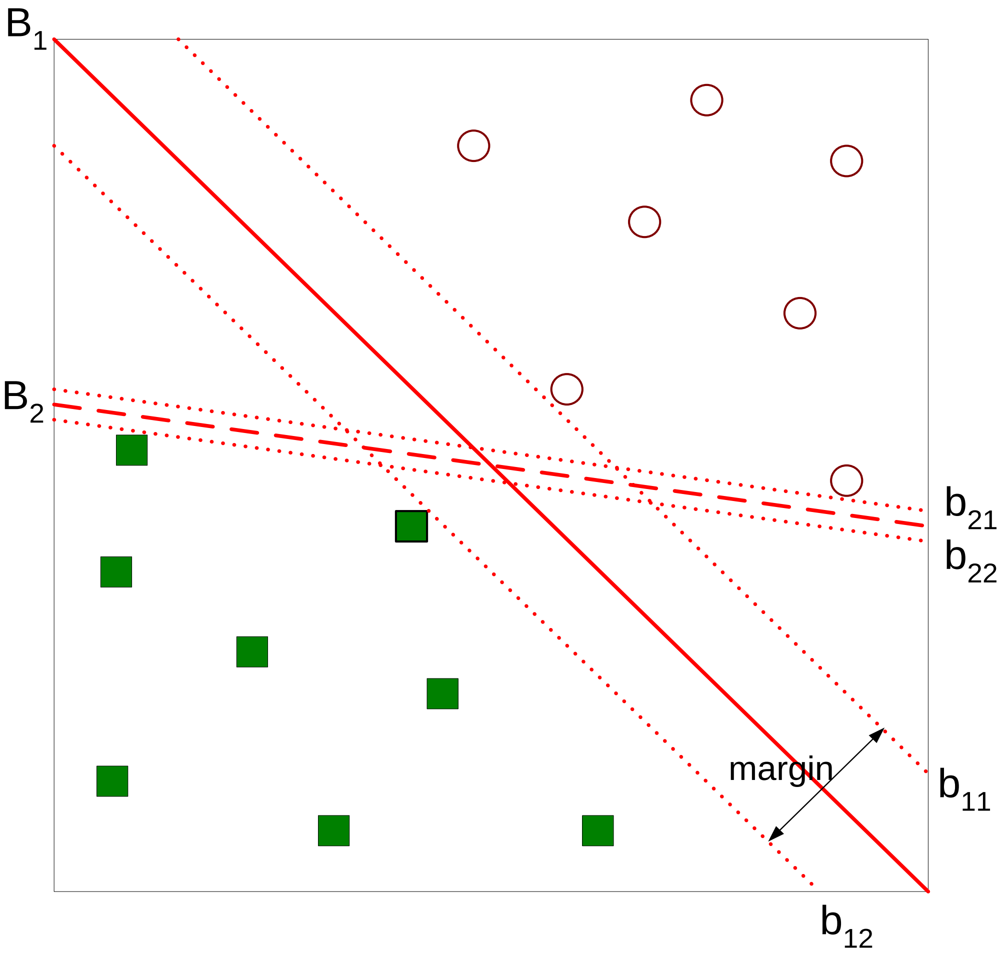
Support Vector Machines



Which one is better? B1 or B2?
How do you define better?

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

Support Vector Machines

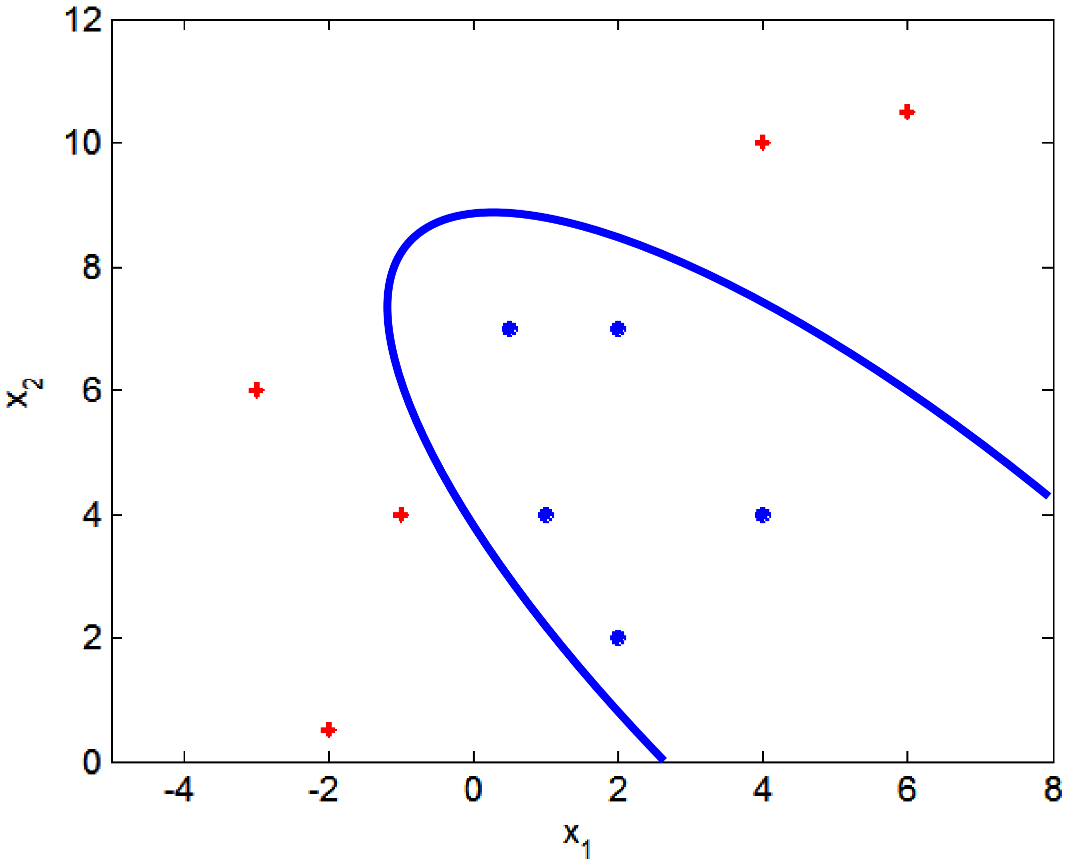


Find hyperplane maximizes the margin => B1 is better than B2

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

Nonlinear Support Vector Machines

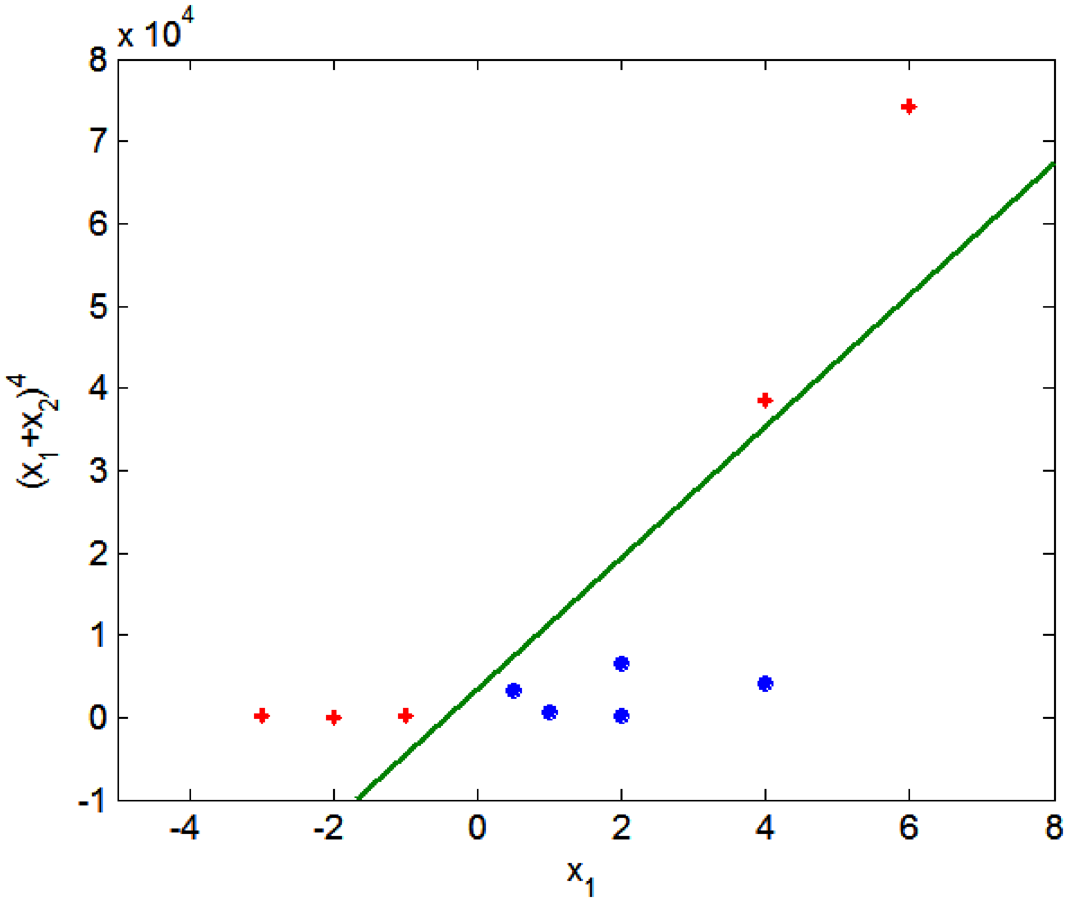
What if decision boundary is not linear?



From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

Nonlinear Support Vector Machines

Transform data into higher dimensional space



From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

Evaluation of Support Vector Machines

Accuracy
- Among best performers

Interpretability
- Model and prediction are not interpretable
  - Black box model

Incrementality
- Not incremental

- Efficiency
  - Model building requires significant parameter tuning
  - Very fast classification
- Scalability
  - Medium scalable both in training set size and attribute number
- Robustness
  - Robust to noise and outliers

Artificial Neural Networks

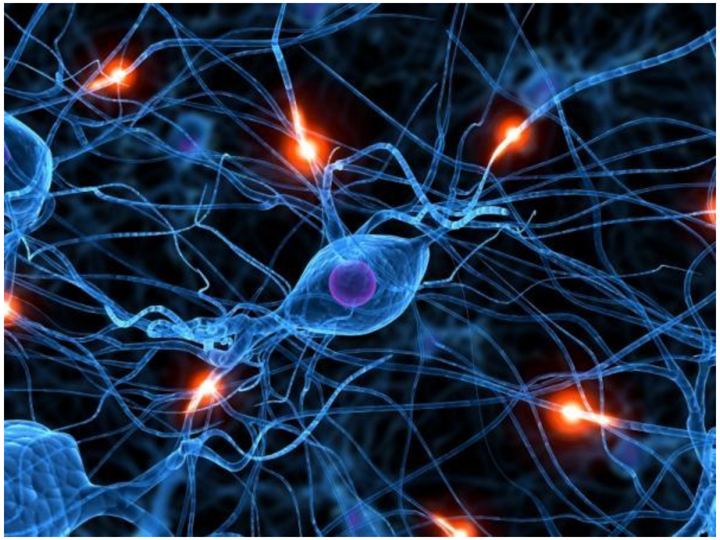
Elena Baralis

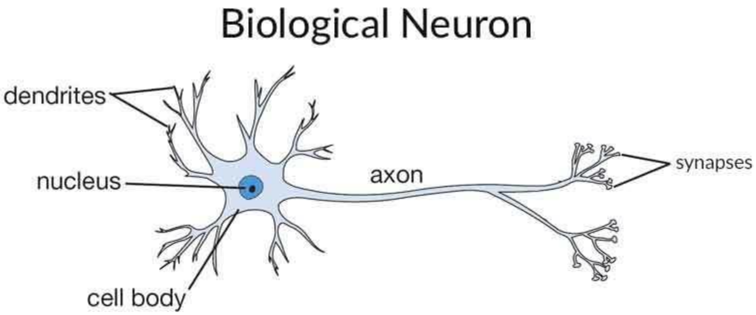
Politecnico di Torino

Artificial Neural Networks

Inspired to the structure of the human brain

- Neurons as elaboration units
- Synapses as connection network

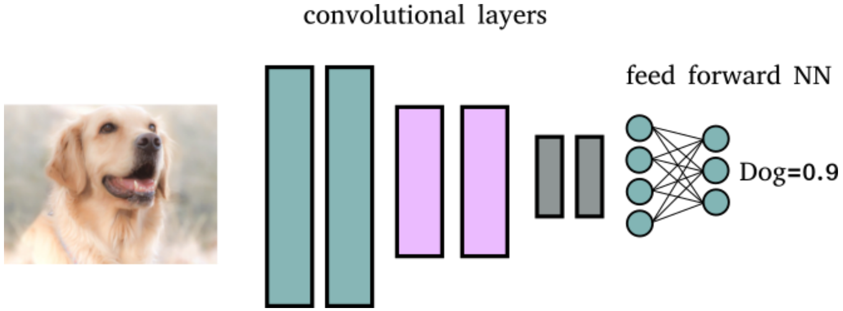




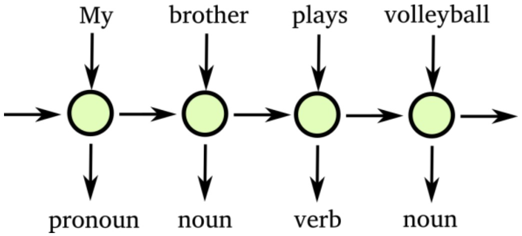
Artificial Neural Networks

Different tasks, different architectures

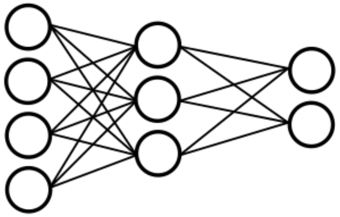
image understanding: convolutional NN (CNN)



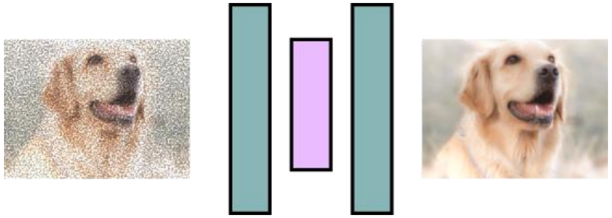
time series analysis: recurrent NN (RNN)



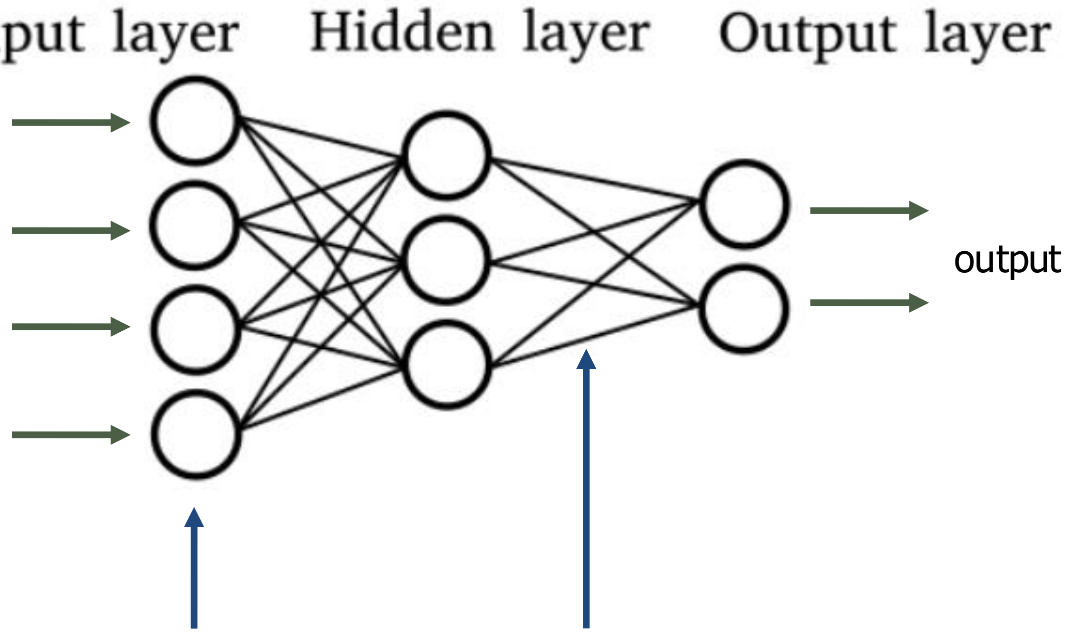
numerical vectors classification: feed forward NN (FFNN)



denoising: auto-encoders



Feed Forward Neural Network



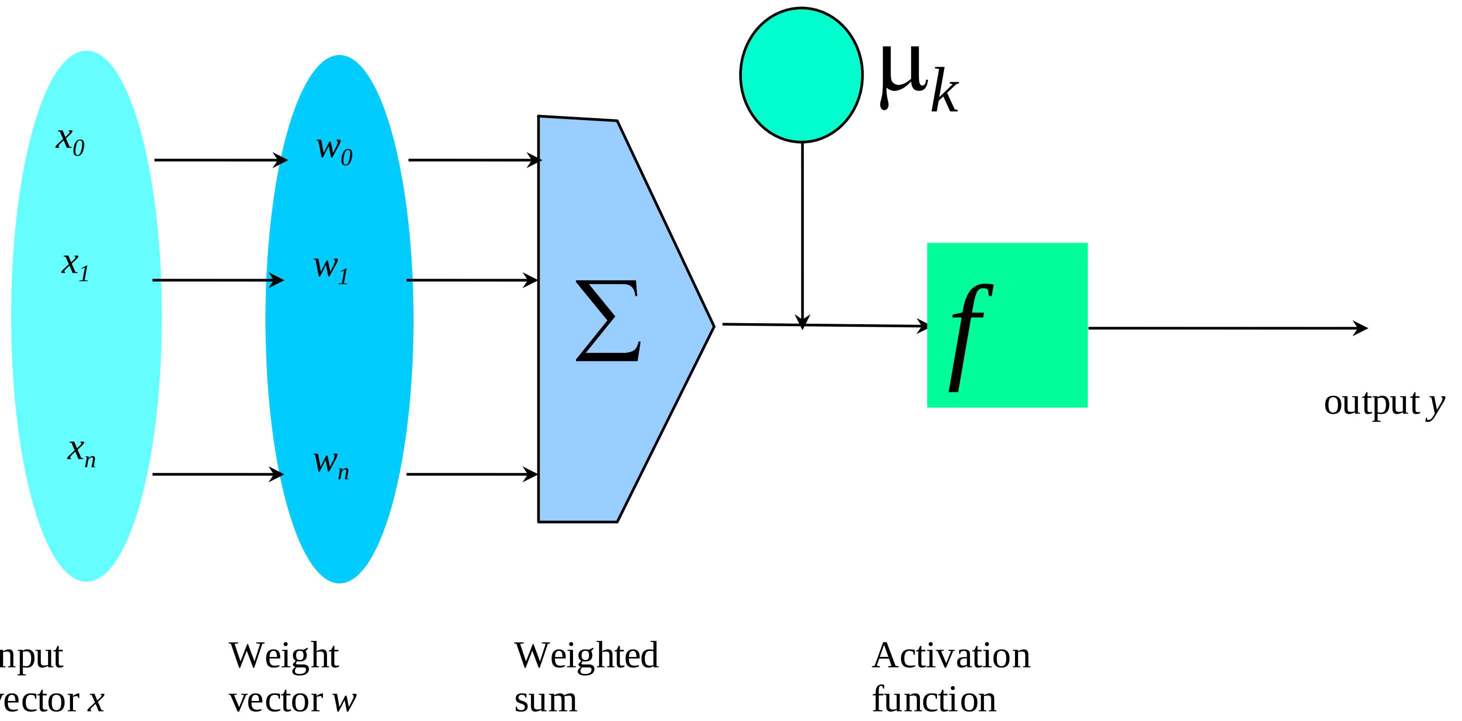
input vector (x_i)

output vector

neuron

weighted connection
(w_{ij})

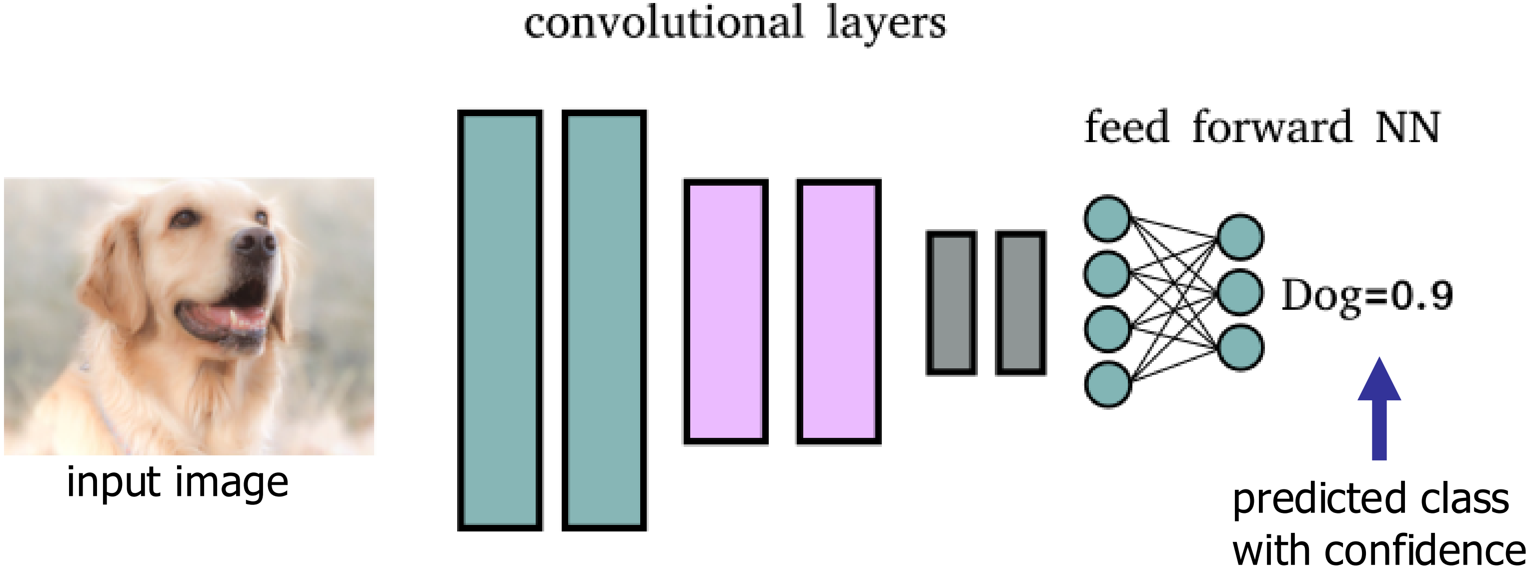
Structure of a neuron



From: Han, Kamber,"Data mining; Concepts and Techniques", Morgan Kaufmann 2006

Convolutional Neural Networks

- Allow automatically extracting **features** from images and performing **classification**



convolutional layers

feed forward NN

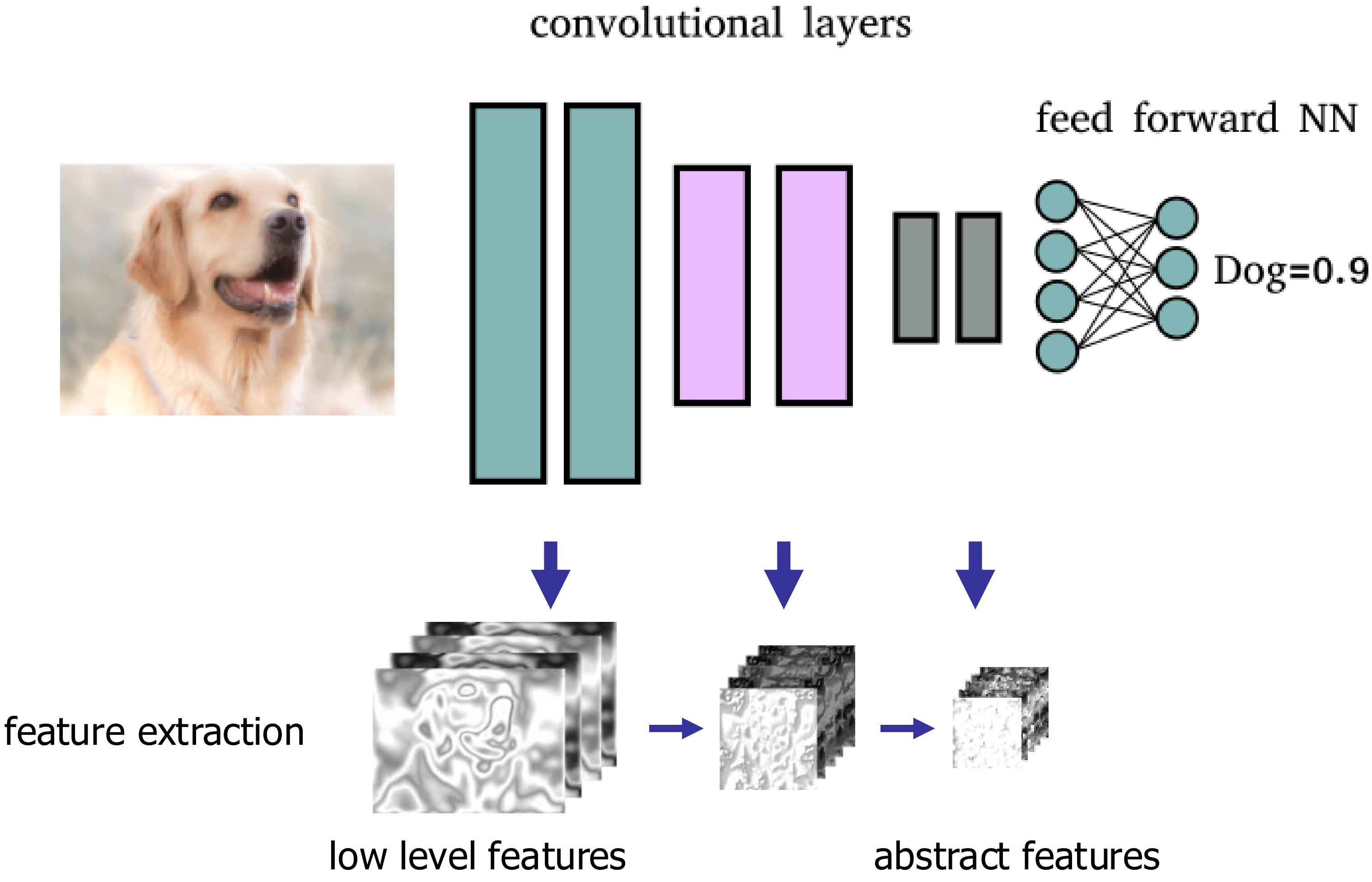
Dog=0.9

input image

predicted class with confidence

Convolutional Neural Network (CNN) Architecture

Convolutional Neural Networks



convolutional layers

feed forward NN

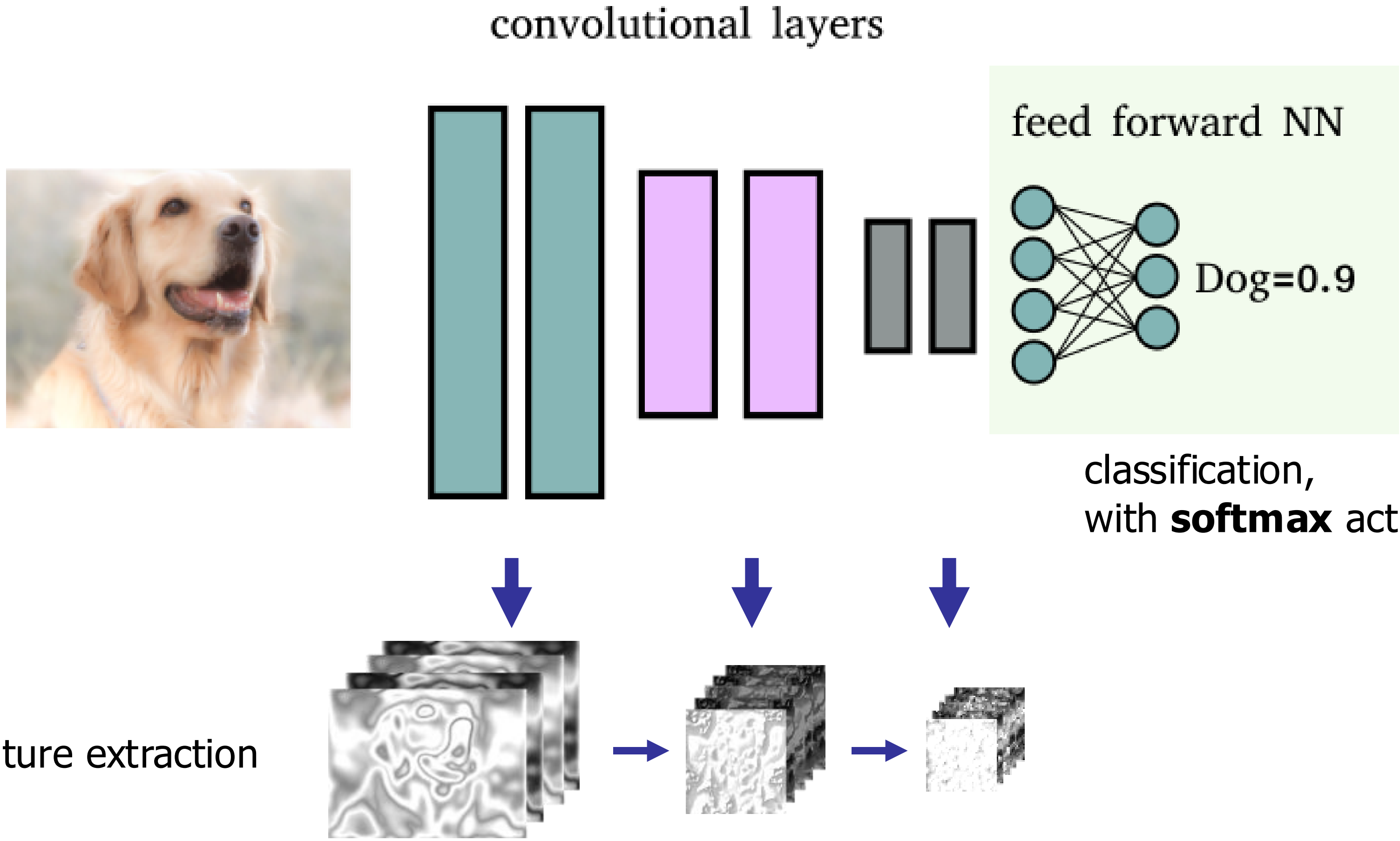
Dog=0.9

feature extraction

low level features

abstract features

Convolutional Neural Networks



classification,
with **softmax** activation

feature extraction

low level features

abstract features

Convolutional Neural Networks

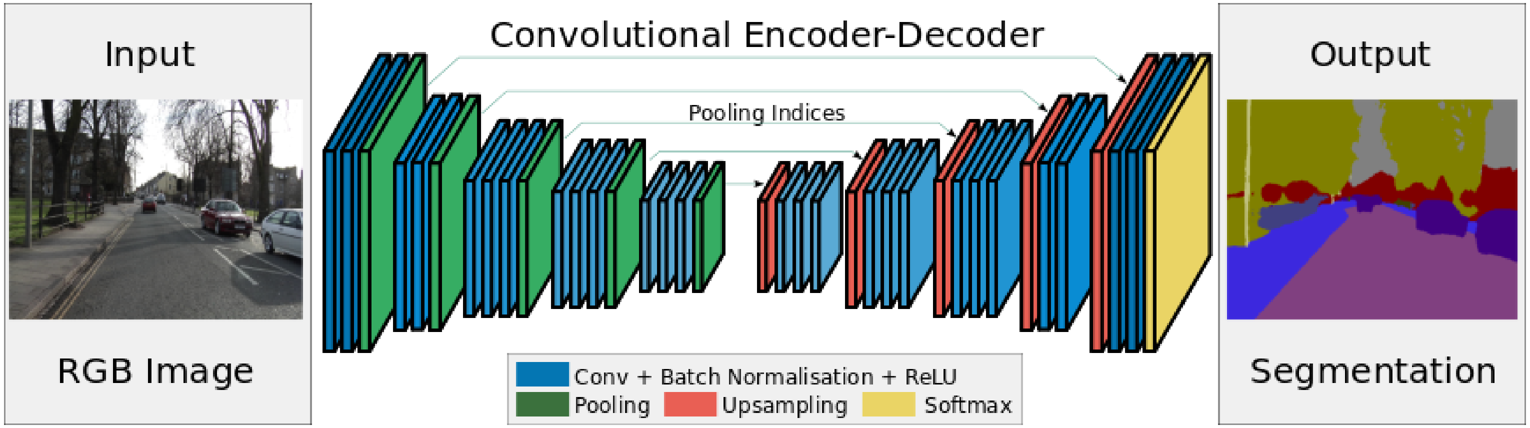
Semantic segmentation CNNs

allow assigning a class to each pixel of the input image

composed of 2 parts

encoder network: convolutional layers to extract abstract features

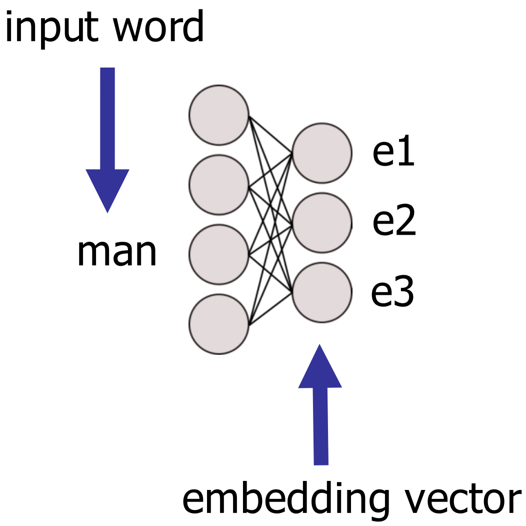
decoder network: deconvolutional layers to obtain the output image from the extracted features



SegNet neural network

Word Embeddings (Word2Vec)

- Word *embeddings* associate words to n-dimensional vectors
  - trained on big text collections to model the word distributions in different sentences and contexts
  - able to capture the *semantic* information of each word
  - words with similar *meaning* share vectors with similar characteristics



input word

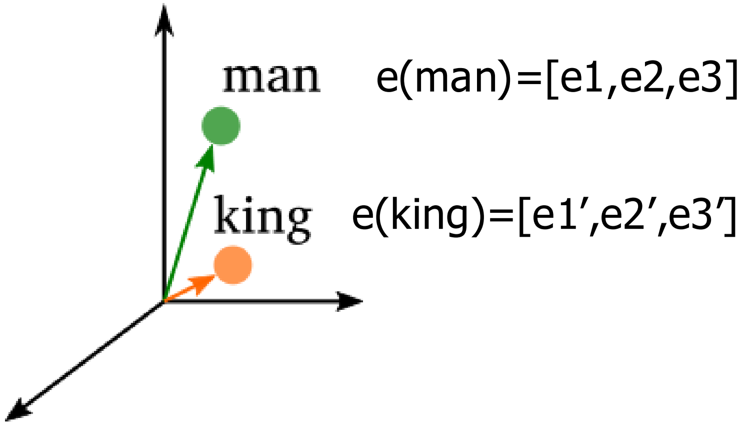
man

e1

e2

e3

embedding vector

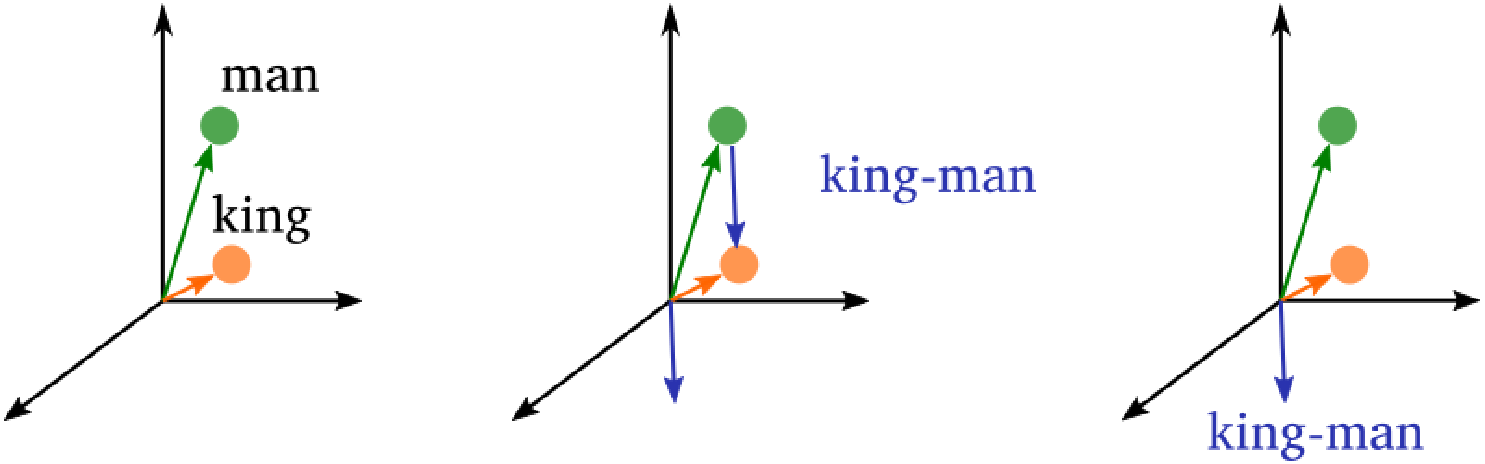


man e(man)=[e1,e2,e3]

king e(king)=[e1′,e2′,e3′]

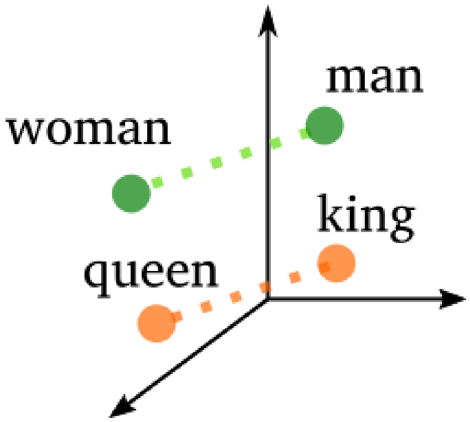
Word Embeddings (Word2Vec)

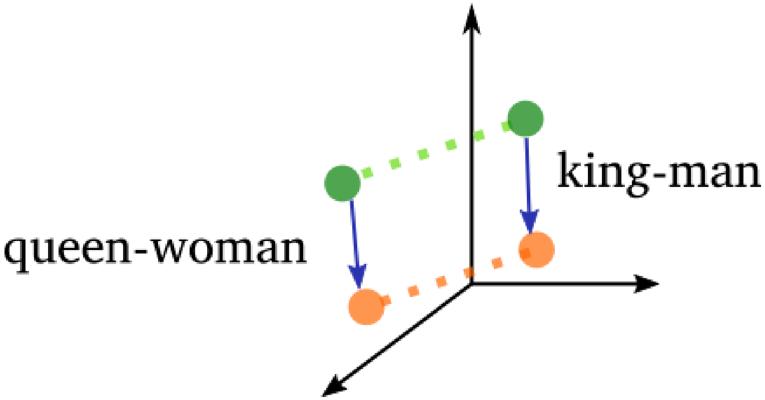
- Since each word is represented with a vector, operations among words (e.g. difference, addition) are allowed



Word Embeddings (Word2Vec)

- Semantic relationships among words are captured by vector positions





king - man = queen - woman
king - man + woman = queen

Model evaluation

Elena Baralis

Politecnico di Torino

Model evaluation

Methods for performance evaluation

- Partitioning techniques for training and test sets

Metrics for performance evaluation

- Accuracy, other measures

Techniques for model comparison

- ROC curve

Methods for performance evaluation

Objective

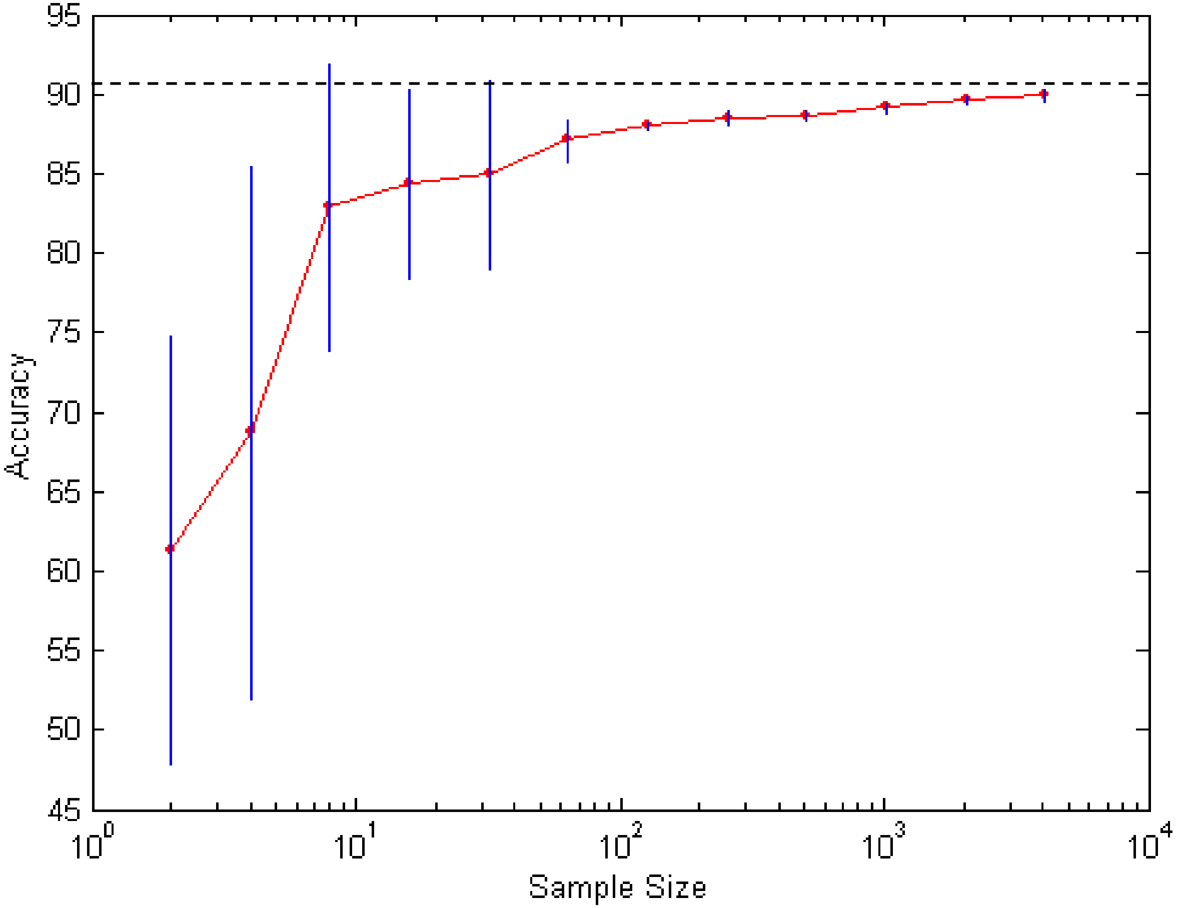
- reliable estimate of performance

Performance of a model may depend on other factors besides the learning algorithm

- Class distribution
- Cost of misclassification
- Size of training and test sets

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

Learning curve



- Learning curve shows how accuracy changes with varying training sample size
- Requires a sampling schedule for creating learning curve:
  - Arithmetic sampling (Langley, et al)
  - Geometric sampling (Provost et al)

Effect of small sample size:

- Bias in the estimate
- Variance of estimate

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

Partitioning data

Several partitioning techniques

- holdout
- cross validation

Stratified sampling to generate partitions

- without replacement

Bootstrap

- Sampling with replacement

Methods of estimation

Partitioning labeled data for training, validation and test

Several partitioning techniques

- holdout
- cross validation

Stratified sampling to generate partitions

- without replacement

Bootstrap

- Sampling with replacement

Holdout

Fixed partitioning

- Typically, may reserve 80% for training, 20% for test
- Other proportions may be appropriate, depending on the dataset size

Appropriate for large datasets

- may be repeated several times
  - repeated holdout

Cross validation

Cross validation

partition data into k disjoint subsets (i.e., folds)

k-fold: train on k-1 partitions, test on the remaining one

- repeat for all folds

reliable accuracy estimation, not appropriate for very large datasets

Leave-one-out

cross validation for k=n

only appropriate for very small datasets

Model performance estimation

Model training step

Building a new model

Model validation step

Hyperparameter tuning

Algorithm selection

Model test step

Estimation of model performance

Model performance estimation

Typical dataset size

- Training set 60% of labeled data
- Validation set 20% of labeled data
- Test set 20% of labeled data

Splitting labeled data

- Use hold-out to split in
  - training+validation
  - test
- Use cross validation to split in
  - training
  - validation

Metrics for model evaluation

Evaluate the predictive accuracy of a model

Confusion matrix

binary classifier

	PREDICTED CLASS		
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	a	b
	Class=No	c	d

a: TP (true positive)

b: FN (false negative)

c: FP (false positive)

d: TN (true negative)

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

Accuracy

Most widely-used metric for model evaluation

$$\text{Accuracy} = \frac{\text{Number of correctly classified objects}}{\text{Number of classified objects}}$$

Not always a reliable metric

Accuracy

For a binary classifier

	PREDICTED CLASS		
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

$$\text{Accuracy} = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN}$$

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

Limitations of accuracy

Consider a binary problem

- Cardinality of Class 0 = 9900
- Cardinality of Class 1 = 100

Model

$$() \rightarrow \textit{class 0}$$

- Model predicts everything to be class 0
  - accuracy is 9900/10000 = 99.0 %

Accuracy is misleading because the model does not detect any class 1 object

Limitations of accuracy

Classes may have different importance

- Misclassification of objects of a given class is more important
- e.g., ill patients erroneously assigned to the healthy patients class

Accuracy is not appropriate for

- unbalanced class label distribution
- different class relevance

Class specific measures

- Evaluate separately for each class C

$$\text{Recall (r)} = \frac{\text{Number of objects correctly assigned to C}}{\text{Number of objects belonging to C}}$$

$$\text{Precision (p)} = \frac{\text{Number of objects correctly assigned to C}}{\text{Number of objects assigned to C}}$$

- Maximize

$$\text{F - measure (F)} = \frac{2rp}{r + p}$$

Class specific measures

- For a binary classification problem
  - on the confusion matrix, for the positive class

$$\text{Precision (p)} = \frac{a}{a+c}$$

$$\text{Recall (r)} = \frac{a}{a+b}$$

$$\text{F - measure (F)} = \frac{2rp}{r+p} = \frac{2a}{2a+b+c}$$

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

ROC (Receiver Operating Characteristic)

Developed in 1950s for signal detection theory to analyze noisy signals

- characterizes the trade-off between positive hits and false alarms

ROC curve plots

- TPR, True Positive Rate (on the y-axis)

  TPR = TP/(TP+FN)

- against
- FPR, False Positive Rate (on the x-axis)

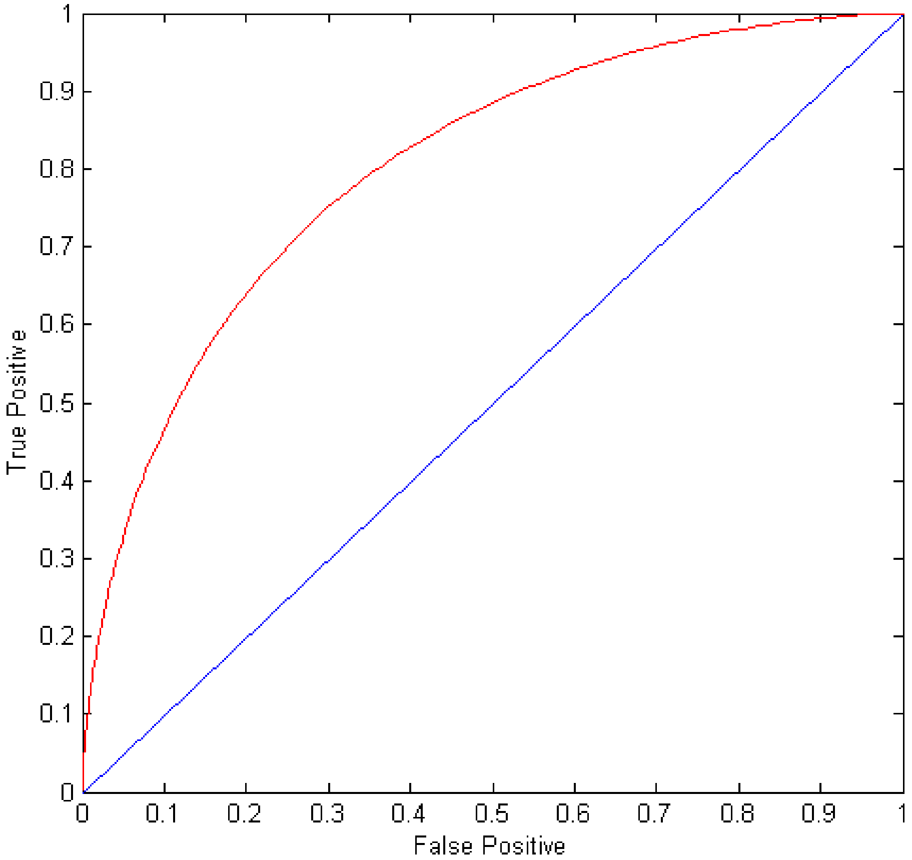
  FPR = FP/(FP + TN)

ROC curve

(FPR, TPR)
(0,0): declare everything
    to be negative class
(1,1): declare everything
    to be positive class
(0,1): ideal

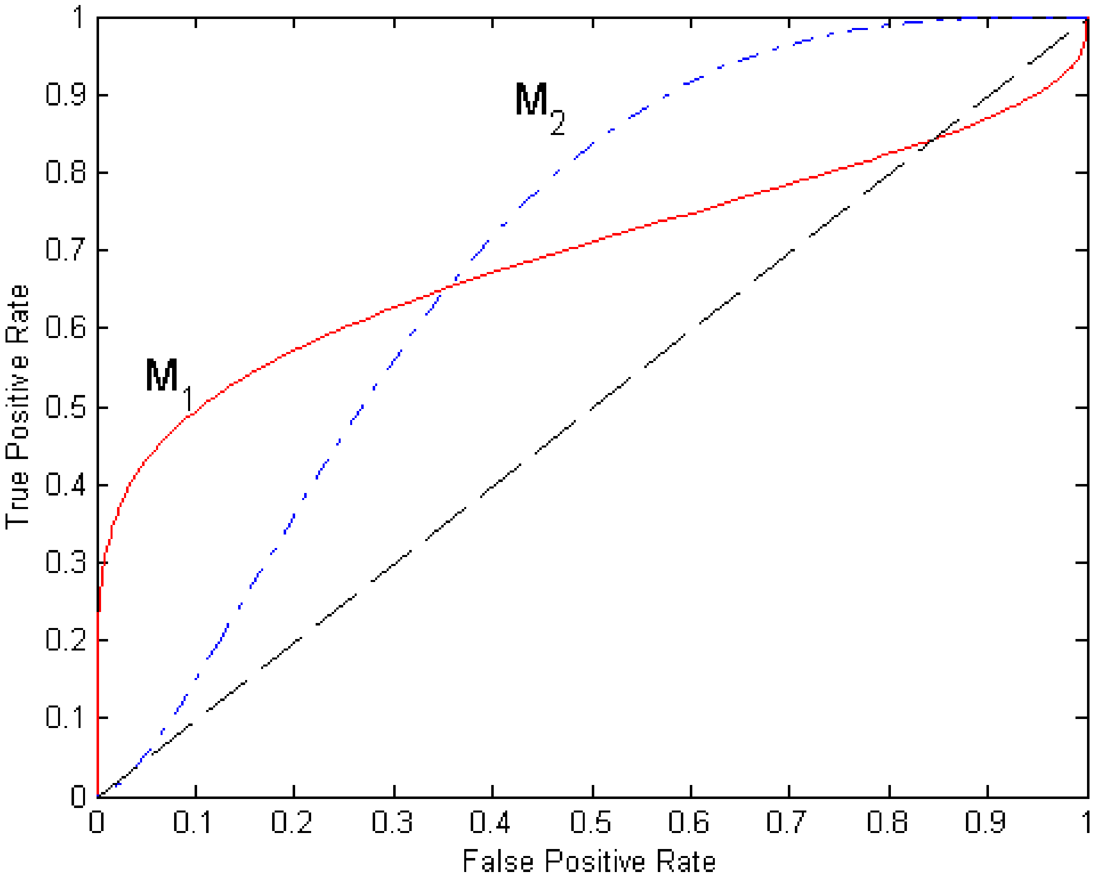
Diagonal line

- Random guessing
- Below diagonal line
  - prediction is opposite of the tr class



From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

Using ROC for Model Comparison



- No model consistently outperforms the other
  - M_1 is better for small FPR
  - M_2 is better for large FPR
- Area under ROC curve
  - Ideal

    Area = 1.0
  - Random guess

    Area = 0.5

From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

Next slides taken from MIT Course of “Data Science”

There are Three Kinds of Lies

LIES

DAMNED LIES

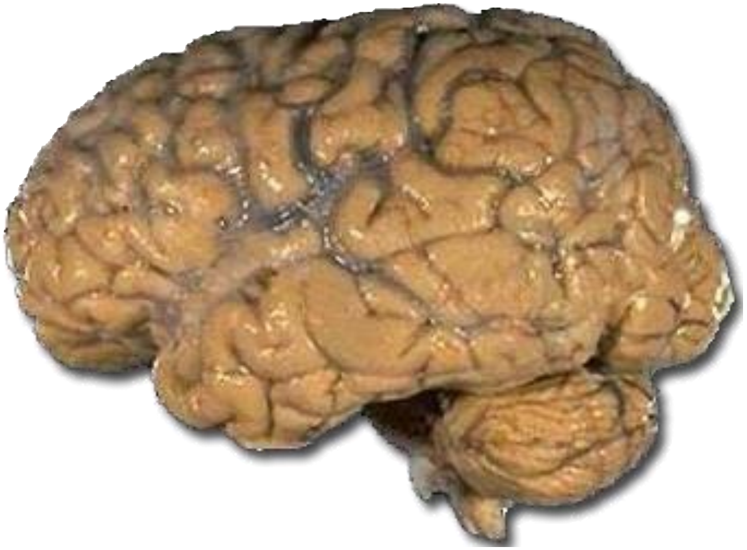
and

STATISTICS

Humans and Statistics

Human Mind

Statistics

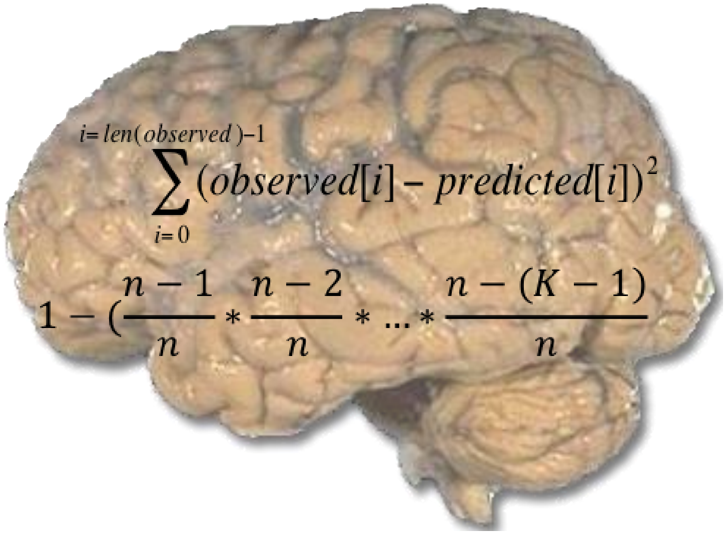


$$\sum_{i=0}^{i=len(observed)-1}(observed[i]-predicted[i])^2$$

$$1-(\frac{n-1}{n}*\frac{n-2}{n}*\ldots*\frac{n-(K-1)}{n}$$

Humans and Statistics

“If you can't prove what you want to prove, demonstrate something else and pretend they are the same thing. In the daze that follows the collision of statistics with the human mind, hardly anyone will notice the difference.” – *Darrell Huff*

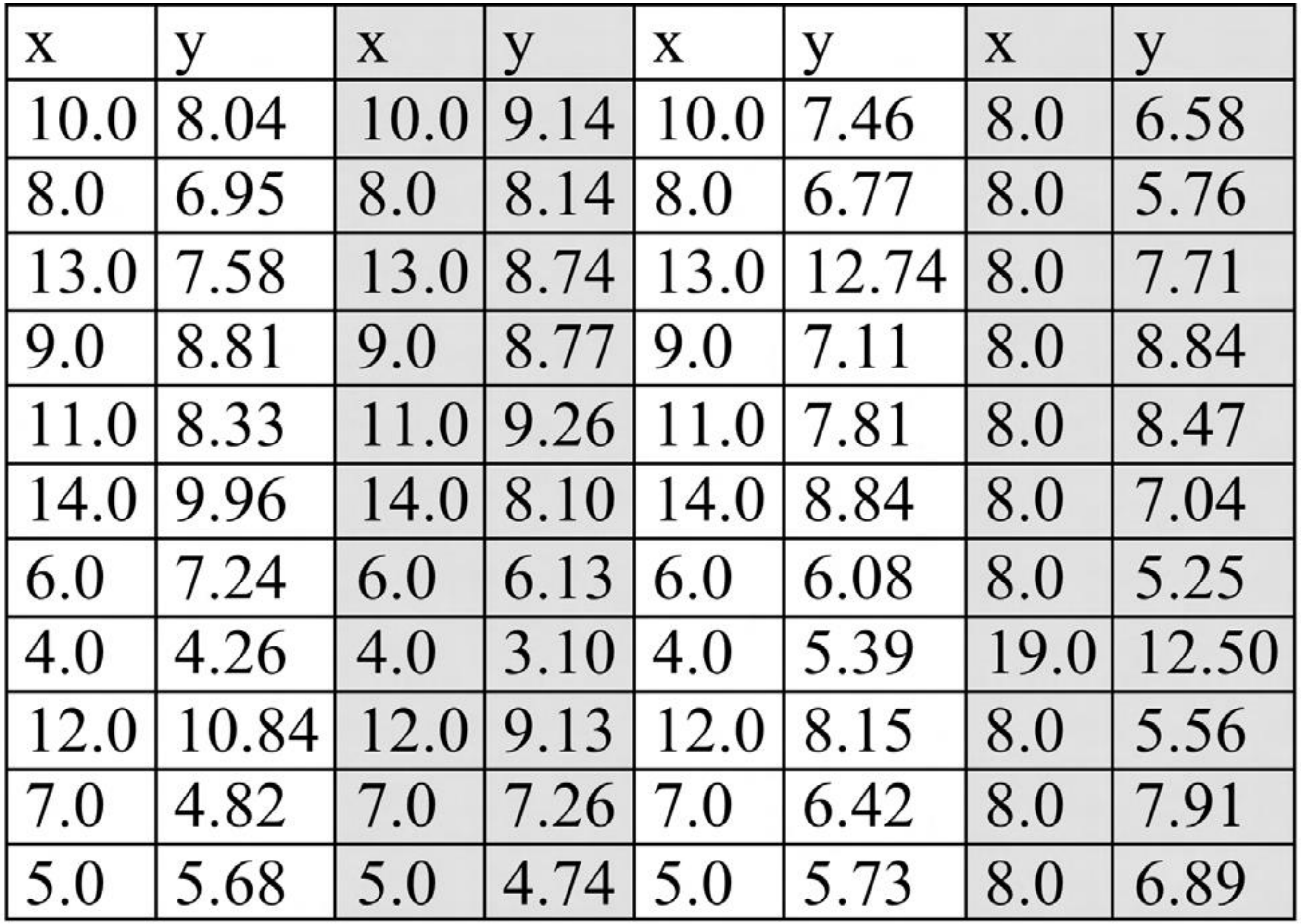


$$\sum_{i=0}^{i=len(observed)-1}(observed[i]-predicted[i])^2$$

$$1-(\frac{n-1}{n}*\frac{n-2}{n}*\ldots*\frac{n-(K-1)}{n}$$

Anscombe’s Quartet

- Four groups each containing 11 x, y pairs

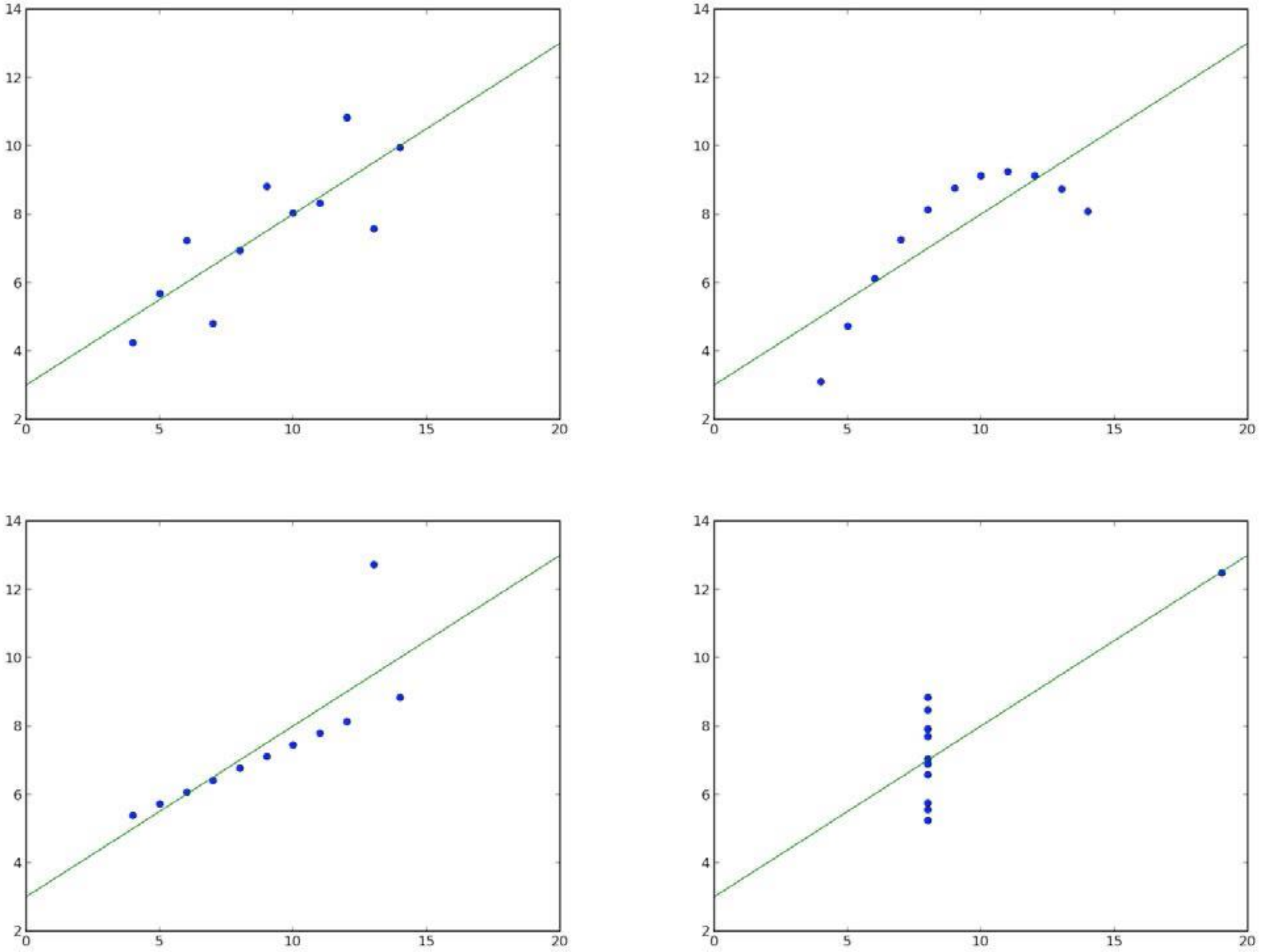


x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Summary Statistics

- Summary statistics for groups identical
  - Mean x = 9.0
  - Mean y = 7.5
  - Variance of x = 10.0
  - Variance of y = 3.75
  - Linear regression model: y = 0.5x + 3
- Are four data sets really similar?

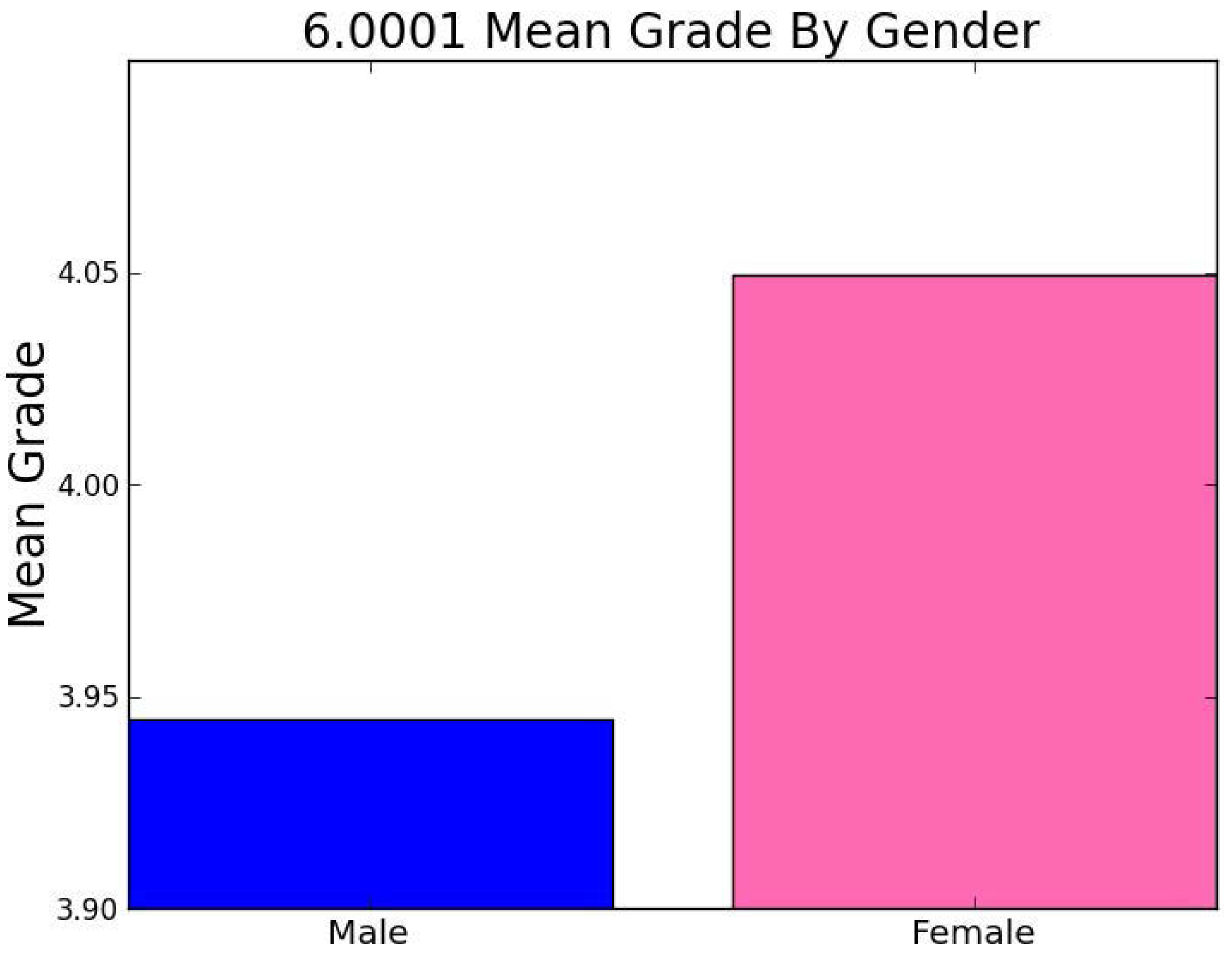
Let’s Plot the Data



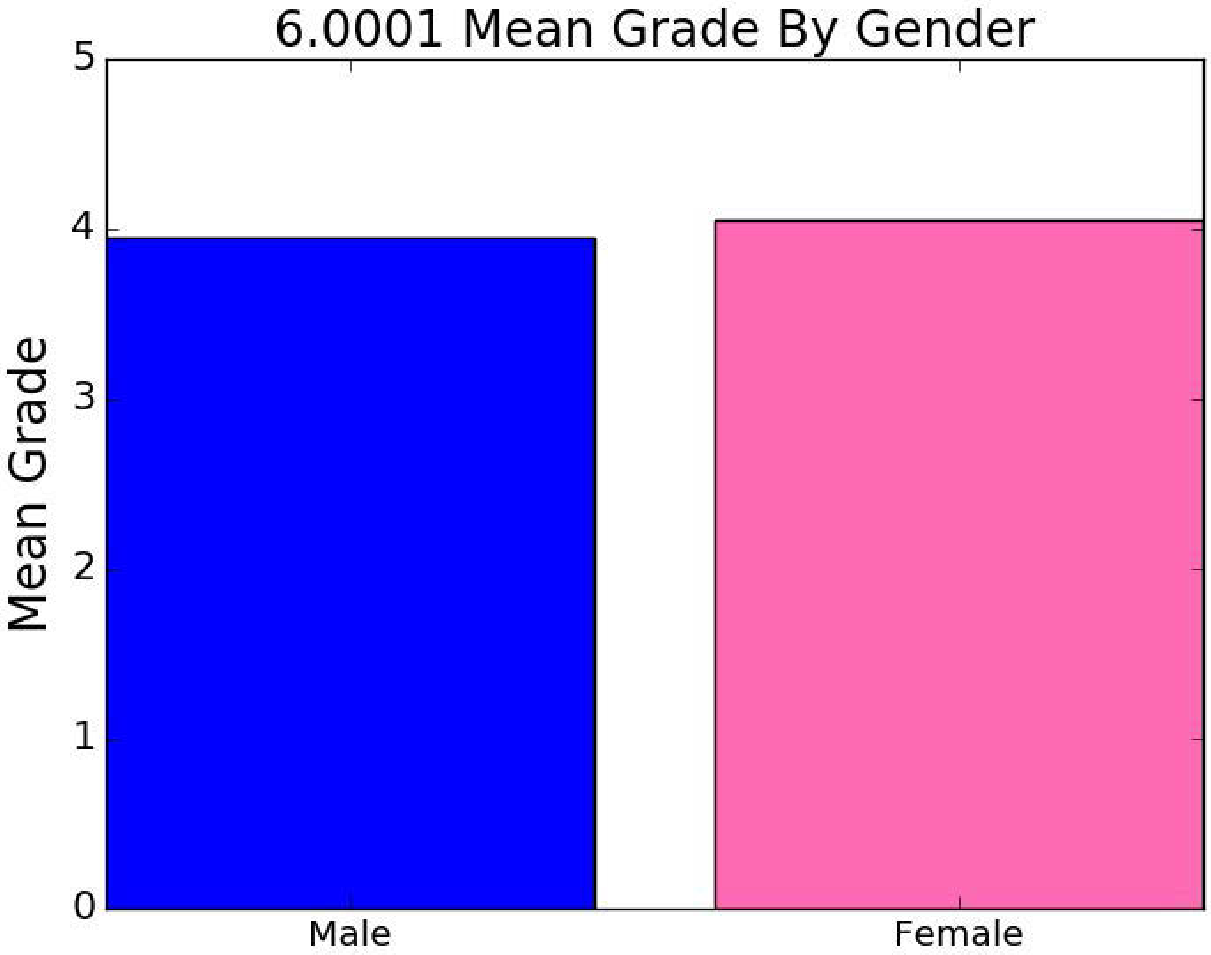
Moral: Statistics about the data is not the same as the data

Moral: Use visualization tools to look at the data itself

Lying with Pictures

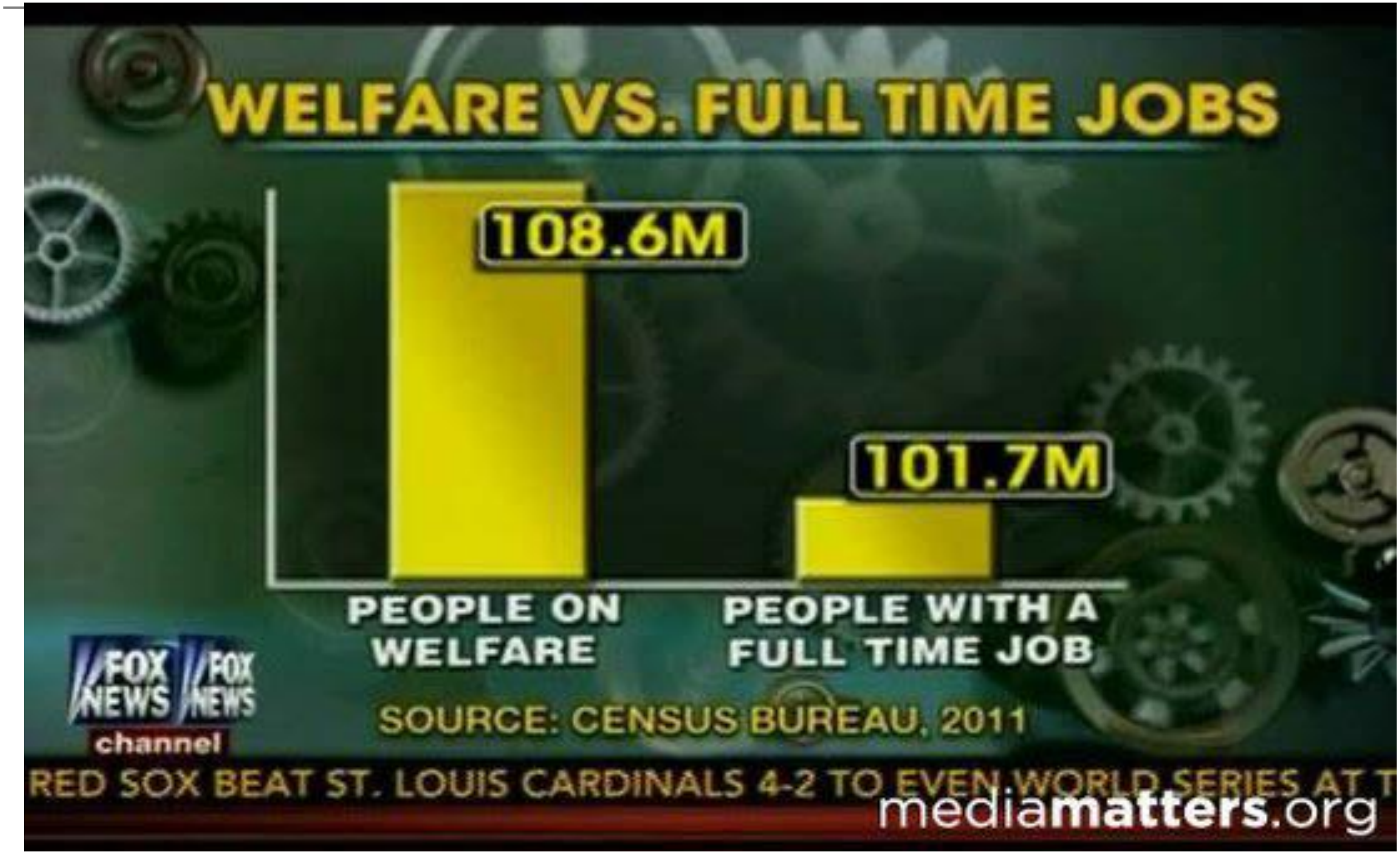


Telling the Truth with Pictures



Moral: Look carefully at the axes labels and scales

Lying with Pictures



Moral: Ask whether the things being compared are actually comparable

Garbage In, Garbage Out

“On two occasions I have been asked [by members of Parliament], ‘Pray, Mr. Babbage, if you put into the machine wrong figures, will the right answers come out?’ I am not able rightly to apprehend the kind of confusion of ideas that could provoke such a question.” – Charles Babbage (1791-1871)

Calhoun’s Response to Errors in Data

“there were so many errors they balanced one another, and led to the same conclusion as if they were all correct.”

Was it the case that the measurement errors are unbiased and independent of each of other, and therefore almost identically distributed on either side of the mean?

No, later analysis showed that the errors were not random but systematic.

“it was the census that was insane and not the colored people.”— James Freeman Clarke

Moral: Analysis of bad data can lead to dangerous conclusions.

Sampling

- All statistical techniques are based upon the assumption that by sampling a subset of a population we can infer things about the population as a whole
- As we have seen, *if random sampling is used*, one can make meaningful mathematical statements about the expected relation of the sample to the entire population
- Easy to get random samples in simulations
- Not so easy in the field, where some examples are more convenient to acquire than others

Non-representative Sampling

- "Convenience sampling" not usually random, e.g.,
  - Survivor bias, e.g., course evaluations at end of course or grading final exam in 6.0002 on a strict curve
  - Non-response bias, e.g., opinion polls conducted by mail or online
- When samples not random and independent, we can still do things like computer means and standard deviations, but **we should not draw conclusions from them** using things like the empirical rule and central limit theorem.
- Moral: Understand how data was collected, and whether assumptions used in the analysis are satisfied. If not, be wary.