BITS Pilani presentation

BITS Pilani
Pilani Campus

Tanmay Tulsidas Verlekar
CSIS

innovate achieve lead

BITS Pilani

Pilani Campus

Applied Machine Learning SE ZG568 / SS ZG568

Lecture No. 3

Data

A data set can often be viewed as a collection of data objects. In turn, data objects are described by a number of attributes that capture the basic characteristics of an object.

Table 2.1. A sample data set containing student information.

Student ID	Year	Grade Point Average (GPA)	...
	⋮		
1034262	Senior	3.24	...
1052663	Sophomore	3.51	...
1082246	Freshman	3.62	...
	⋮		

A measurement scale is a rule (function) that associates a numerical or symbolic value with an attribute of an object.

The Type of an Attribute

Table 2.2. Different attribute types.

Attribute Type		Description	Examples	Operations
Categorical (Qualitative)	Nominal	The values of a nominal attribute are just different names; i.e., nominal values provide only enough information to distinguish one object from another. $(=, \neq)$	zip codes, employee ID numbers, eye color, gender	mode, entropy, contingency correlation, χ^2 test
	Ordinal	The values of an ordinal attribute provide enough information to order objects. $(<, >)$	hardness of minerals, $\{good, better, best\}$, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Numeric (Quantitative)	Interval	For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. $(+, -)$	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviatic Pearson's correlation, t and F tests
	Ratio	For ratio variables, both differences and ratios are meaningful. $(*, /)$	temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current	geometric mean, harmonic mean, percent variation

1. **Distinctness** $=$ and $\neq$
2. **Order** $<, \leq, >,$ and $\geq$
3. **Addition** $+$ and $-$
4. **Multiplication** $*$ and $/$

Discrete: A discrete attribute has a finite or countably infinite set of values.

Binary attributes are a special case of discrete attributes and assume only two values.

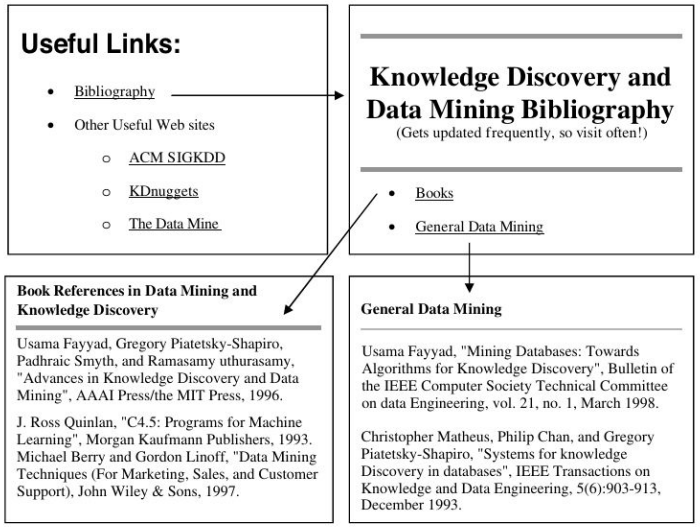
Continuous: A continuous attribute is one whose values are real numbers.

Types of Datasets

General Characteristics of Data Sets: dimensionality, sparsity, and resolution.

Tid	Refund	Marital Status	Taxable Income	Defaulted Borrower
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**

(a) Record data.



(a) Linked Web pages.

Ordered Data

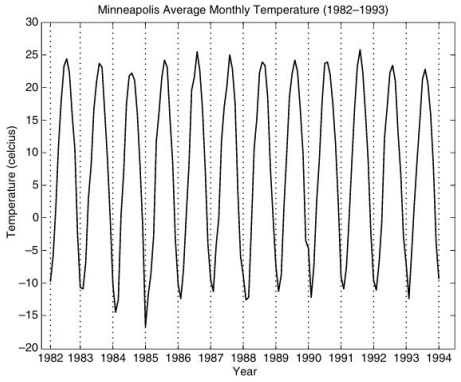
Time	Customer	Items Purchased
t1	C1	A, B
t2	C3	A, C
t2	C1	C, D
t3	C2	A, D
t4	C2	E
t5	C1	A, E

Customer	Time and Items Purchased
C1	(t1: A,B) (t2:C,D) (t5:A,E)
C2	(t3: A, D) (t4: E)
C3	(t2: A, C)

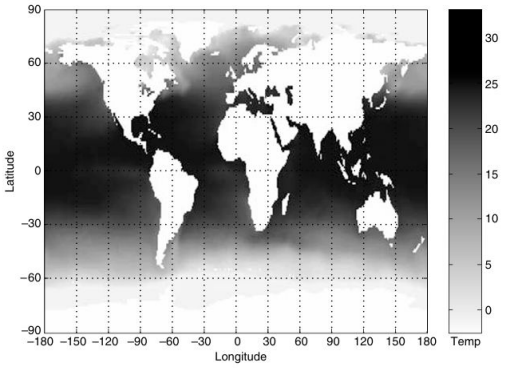
(a) Sequential transaction data.

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

(b) Genomic sequence data.



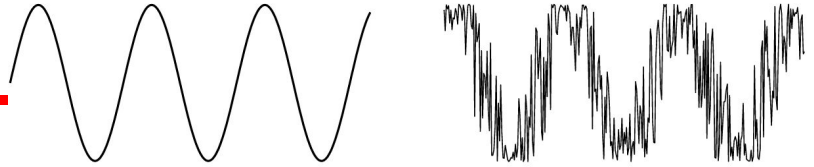
(c) Temperature time series.



(d) Spatial temperature data.

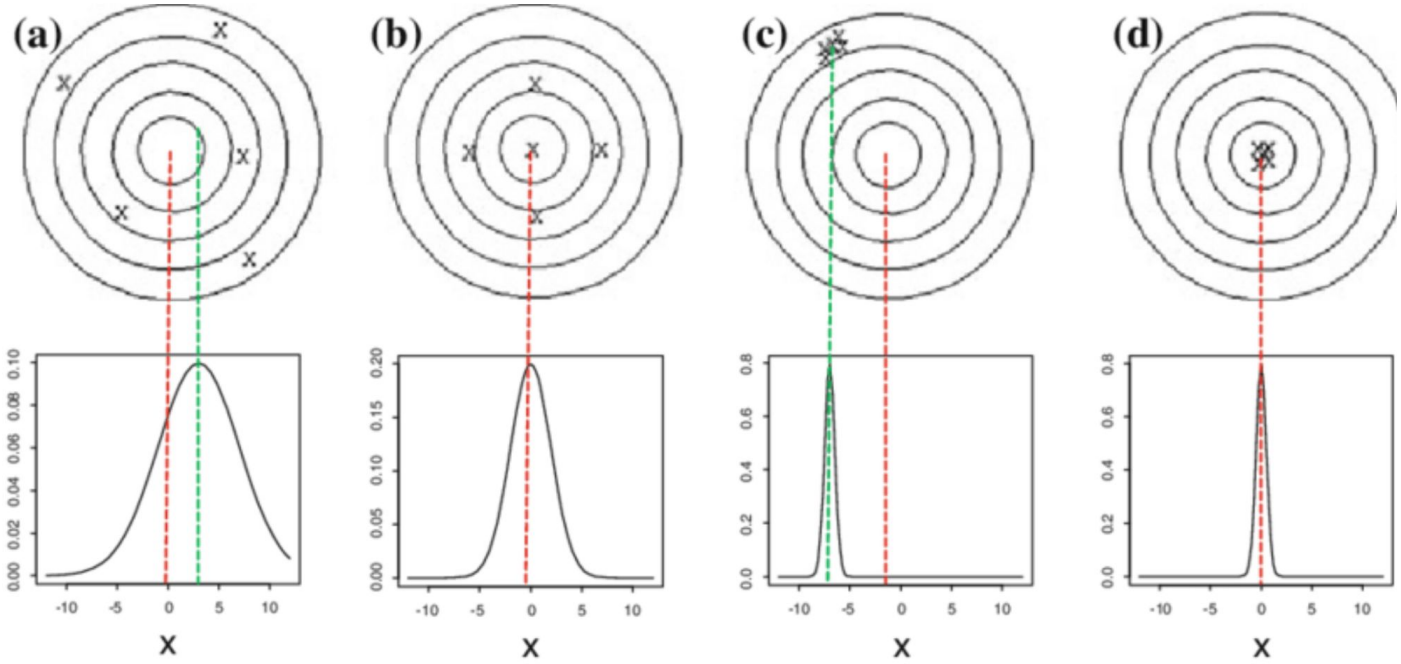
Figure 2.4. Different variations of ordered data.

Data Quality



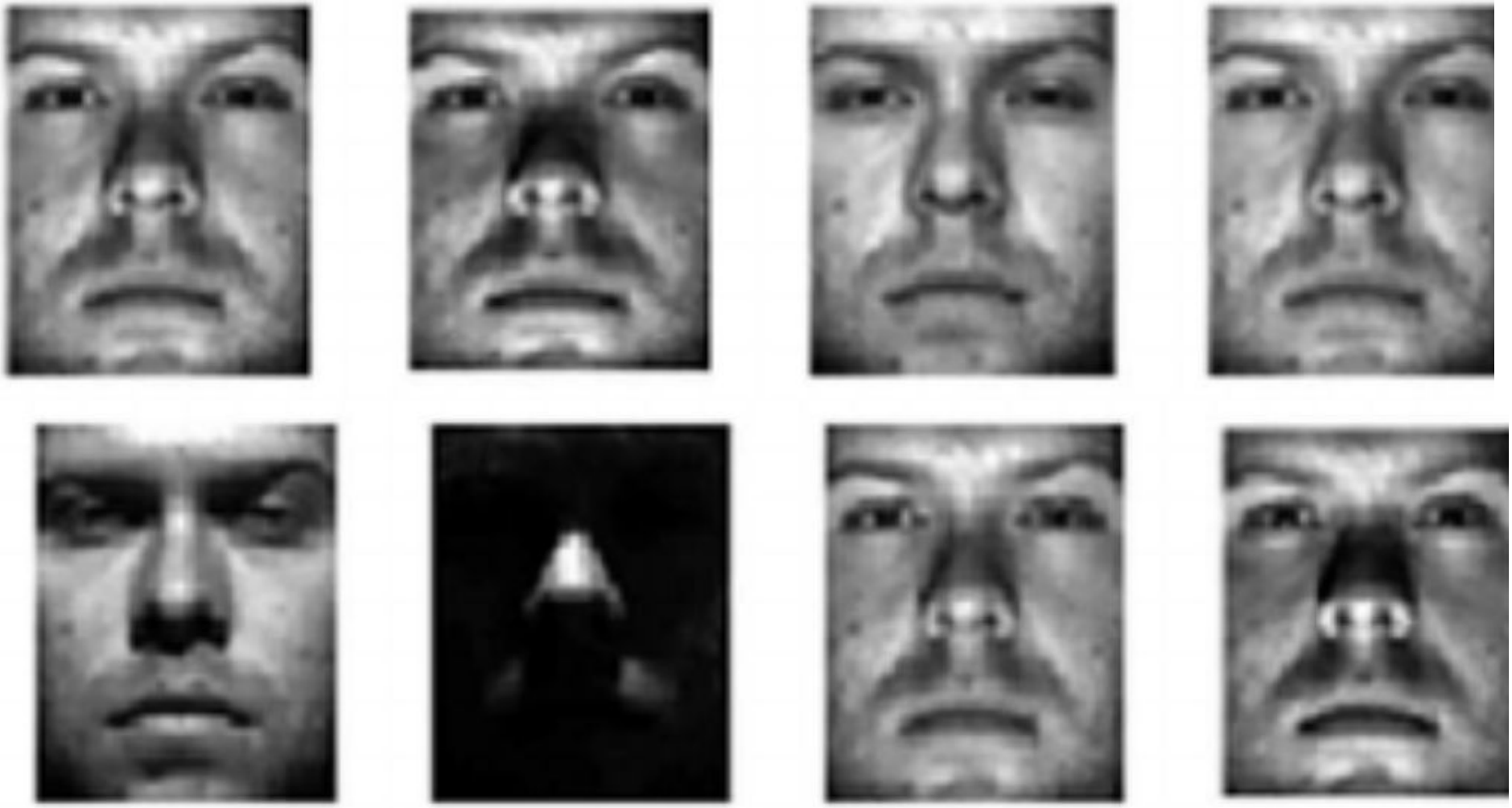
(a) Time series. (b) Time series with noise.

Figure 2.5. Noise in a time series context.



1 Illustration of the concepts of accuracy, precision, and bias. a High bias, low precision) low accuracy; b low bias, low precision) low accuracy; c high bias, high precision) low accuracy; d low bias, high precision) high accuracy. The vertical red line shows the true value (target value). The vertical green line shows the prediction bias. The shape of the curve shows the precision: a curve more concentrated at the mean implies higher precision (low PEV), while a curve less concentrated at the mean implies less precision (high PEV). PEV, prediction error variance

Data Quality: noise vs outlier



How to handle Missing Values

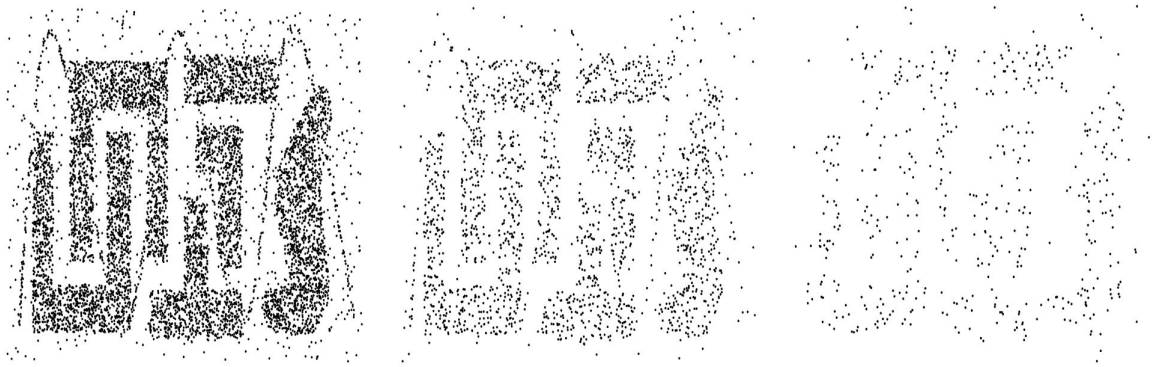
Eliminate Data Objects or Attributes .

Estimate Missing Values.

Ignore the Missing Value during Analysis.

Data Preprocessing

- Aggregation
- Sampling: random vs stratified



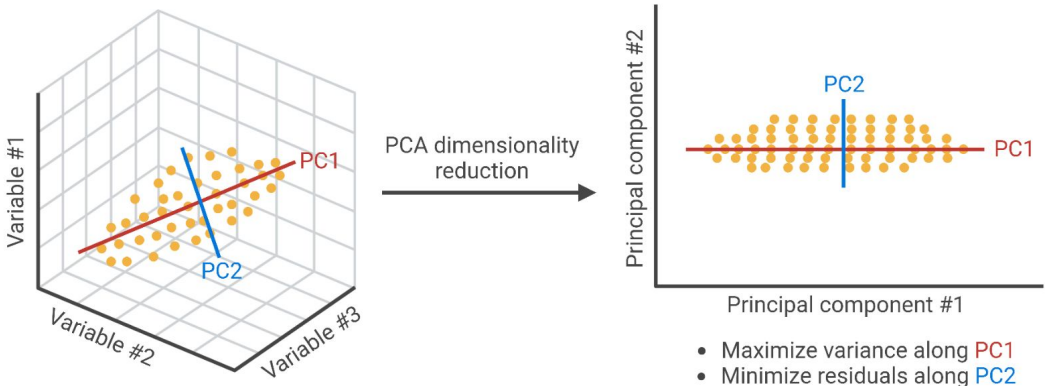
(a) 8000 points (b) 2000 points (c) 500 points

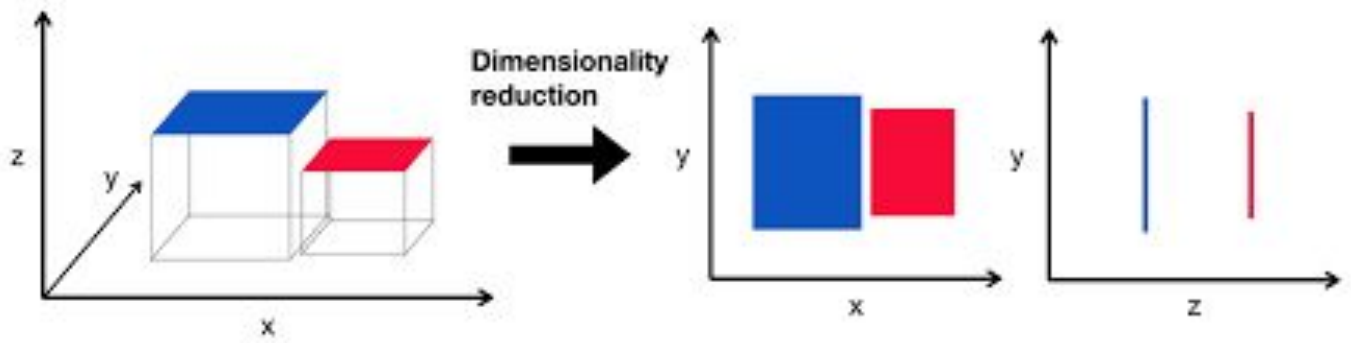
Figure 2.9. Example of the loss of structure with sampling.

Adaptive or progressive sampling schemes : start with a small sample, and then increase the sample size until a sample of sufficient size has been obtained.

Data Preprocessing

- Dimensionality reduction
- Feature subset selection
- Feature creation





Data Preprocessing

- Discretization and binarization

Table 2.6. Conversion of a categorical attribute to five asymmetric binary attributes.

Categorical Value	Integer Value	x_1	x_2	x_3	x_4	x_5
awful	0	1	0	0	0	0
poor	1	0	1	0	0	0
OK	2	0	0	1	0	0
good	3	0	0	0	1	0
great	4	0	0	0	0	1

Data Preprocessing

Data Transformation:

Min-max scaling: values are shifted and rescaled so that they end up ranging from 0 to 1.

Standardization: first it subtracts the mean value (so standardized values always have a zero mean), and then it divides by the standard deviation so that the resulting distribution has unit variance.

Similarity and dissimilarity

Table 2.7. Similarity and dissimilarity for simple attributes

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases}$	$s = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \neq y \end{cases}$
Ordinal	$d = \lvert x - y \rvert / (n-1)$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - d$
Interval or Ratio	$d = \lvert x - y \rvert$	$s = -d$, $s = \frac{1}{1+d}$, $s = e^{-d}$, $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

General Approach to Solving a Classification Problem

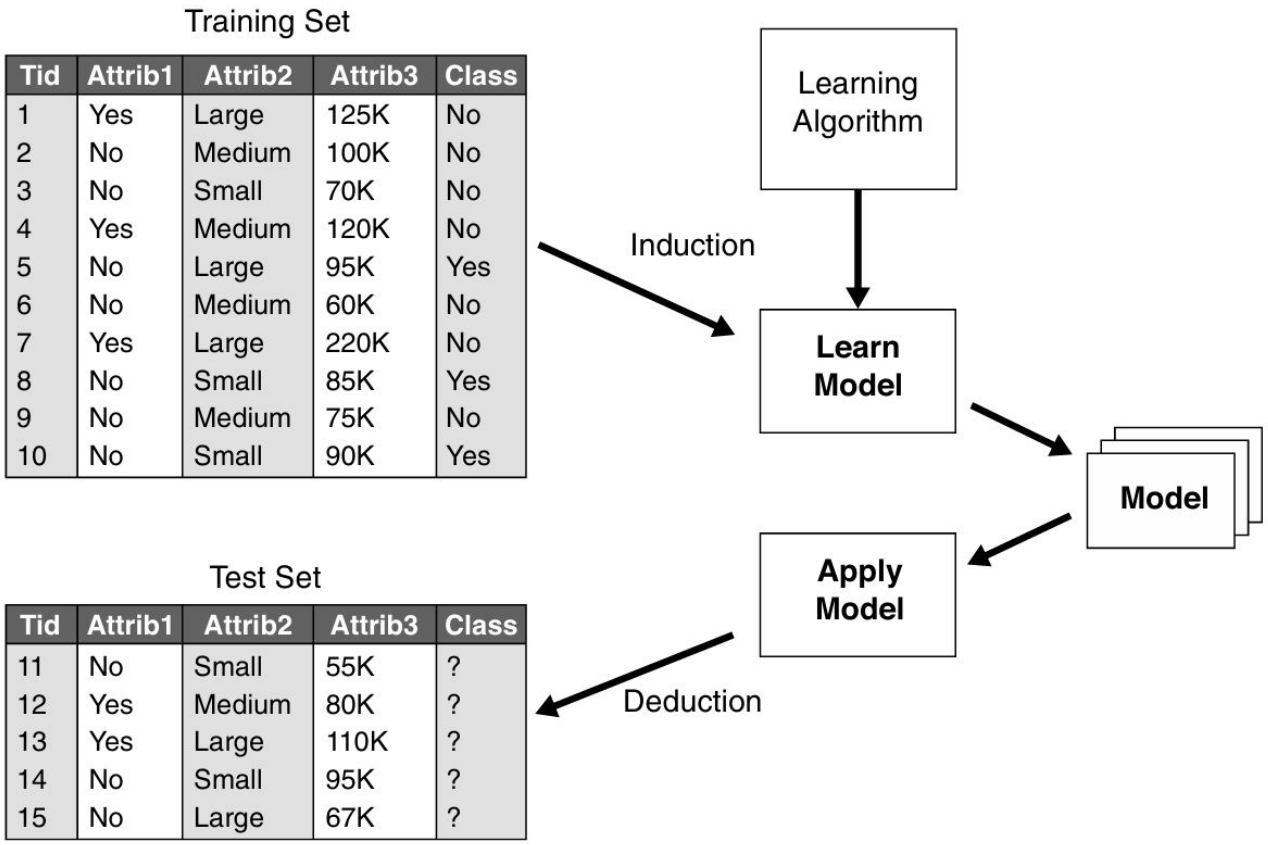
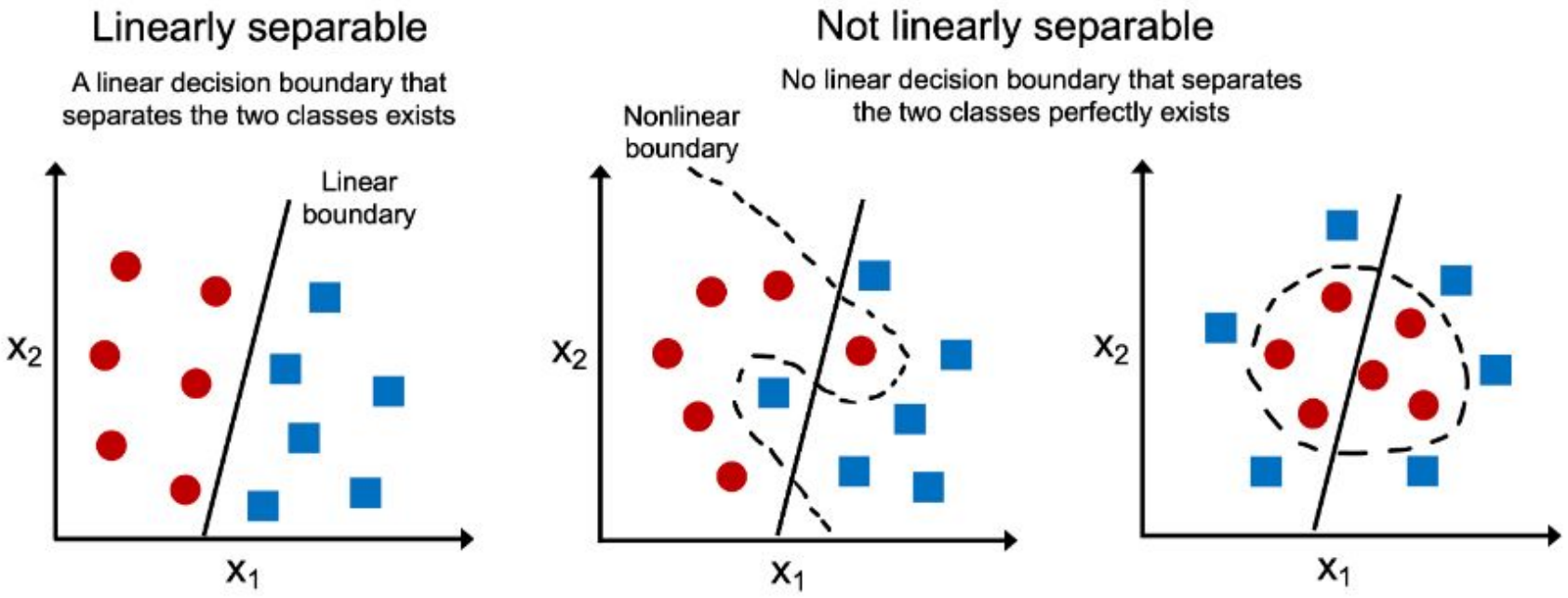


Figure 4.3. General approach for building a classification model.

Select and Train a Model



Fine-Tune Your Model

Grid Search: fiddle with the hyperparameters manually, until you find a great combination of hyperparameter values.

Randomized Search: instead of trying out all possible combinations, it evaluates a given number of random combinations.

Ensemble Methods: combine the models that perform best.

Confusion matrix

$$\text{precision} = \frac{TP}{TP + FP}$$

$$\text{recall} = \frac{TP}{TP + FN}$$

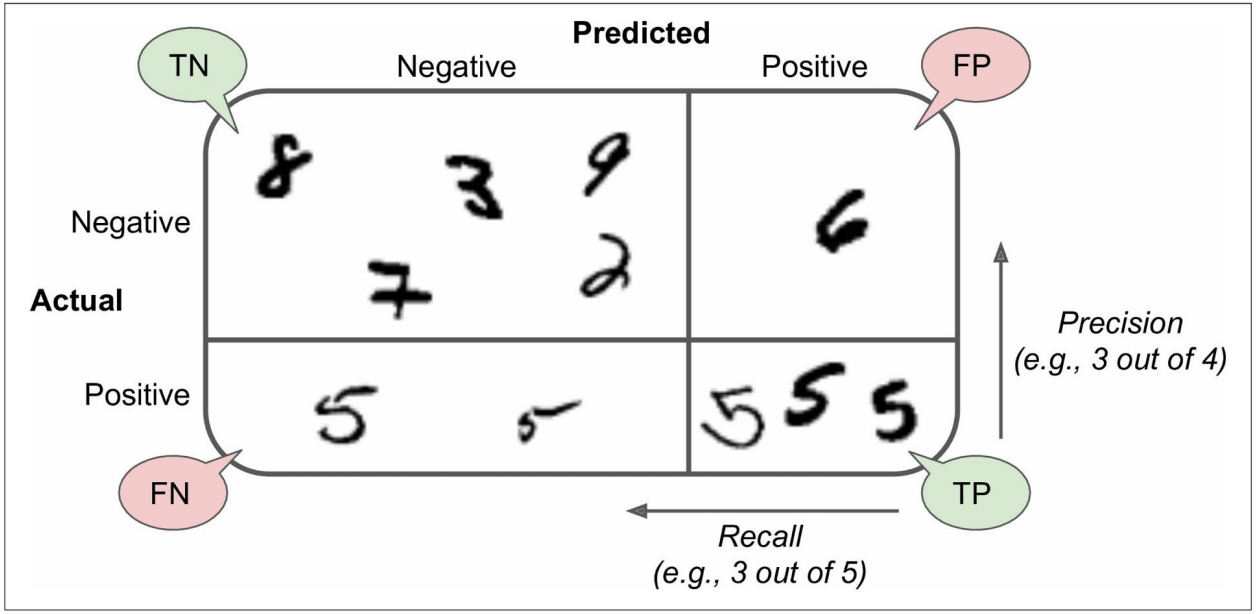


Figure 3-2. An illustrated confusion matrix

$$F_1 = \frac{2}{\frac{1}{\text{precision}} + \frac{1}{\text{recall}}} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = \frac{TP}{TP + \frac{FN + FP}{2}}$$

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

Precision/Recall Tradeof

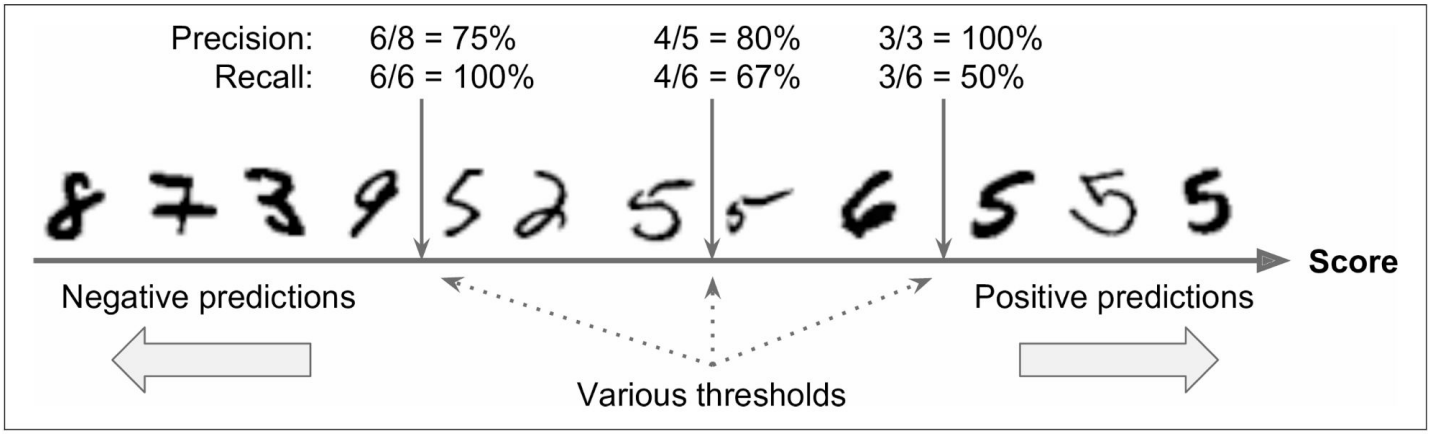


Figure 3-3. Decision threshold and precision/recall tradeoff

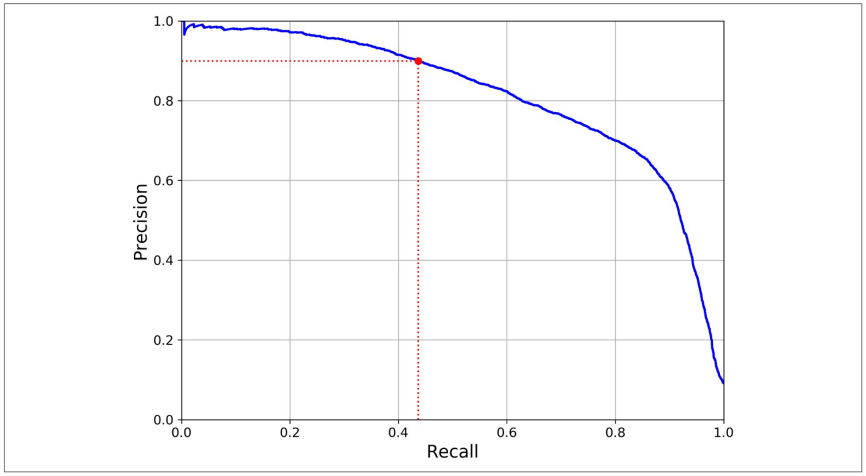


Figure 3-5. Precision versus recall

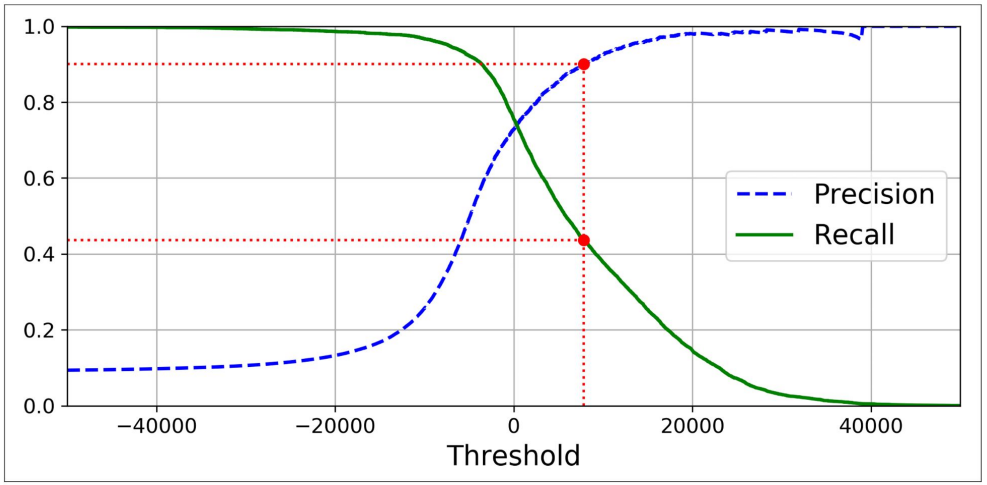


Figure 3-4. Precision and recall versus the decision threshold

ROC

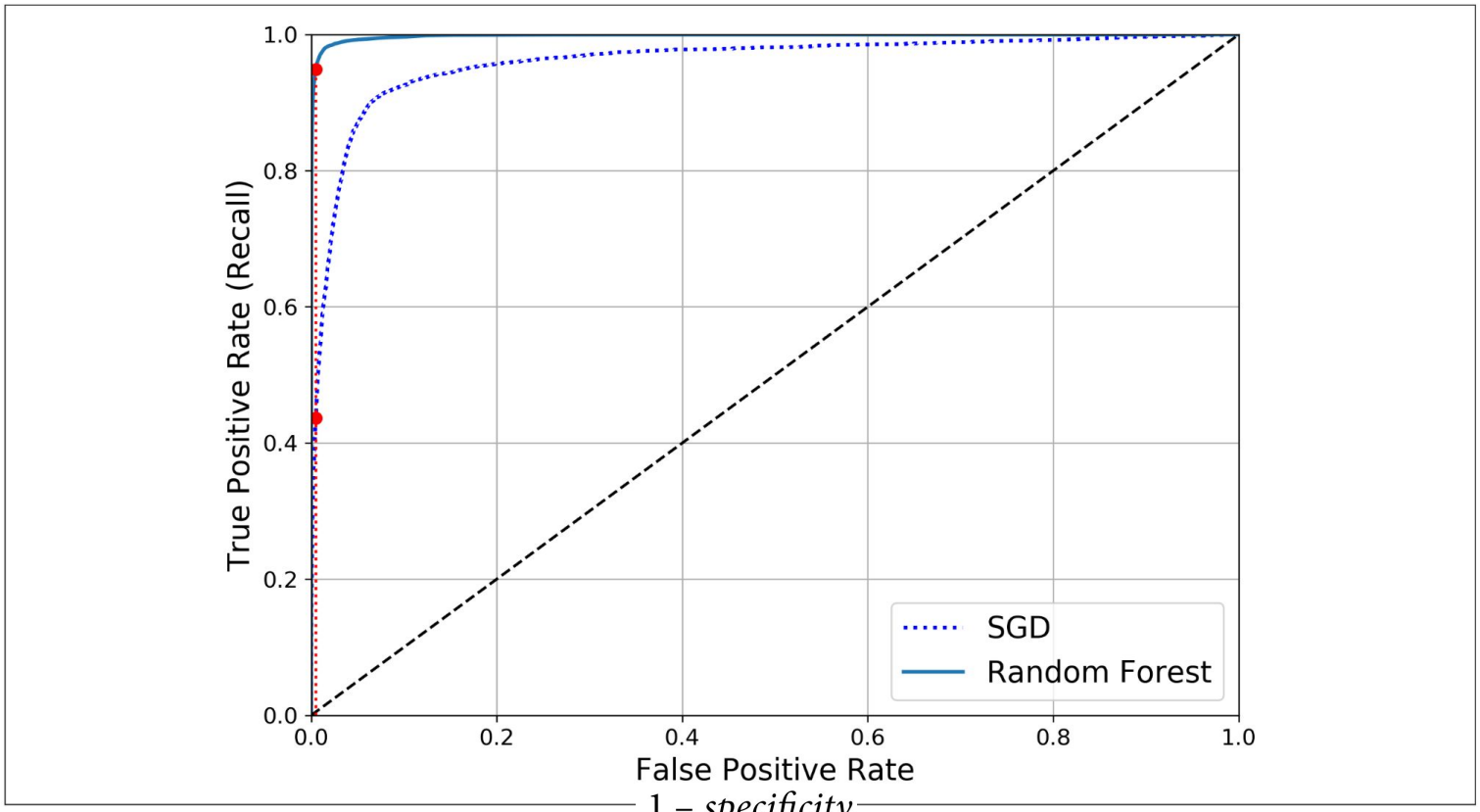
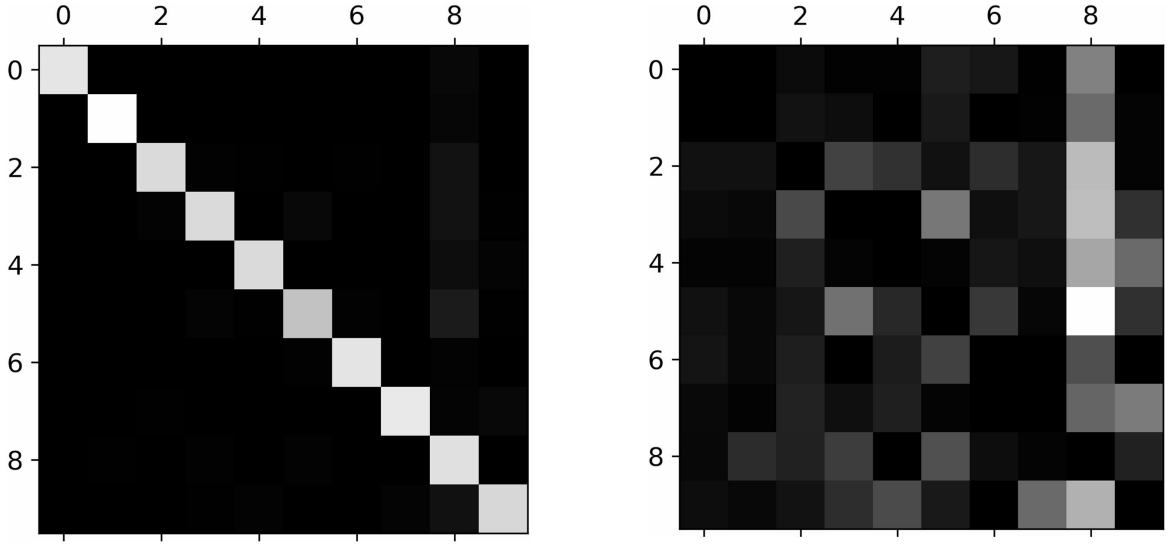


Figure 3-7. Comparing ROC curves

true negative rate $specificity = \frac{tn}{tn + fp}$

Multiclass Classification

Some algorithms are capable of handling multiple classes directly. Others are strictly binary classifiers. They can handle multiclass following: one-versus-all (OvA) or one-versus-one (OvO)



Other types: Multilabel and Multioutput.

Multilabel and Multioutput

Multi-class image classification	Multi-label image classification	Object Detection	Instance Segmentation
			
CAT	CAT, DOG	DOG, DOG, CAT	DOG, DOG, CAT