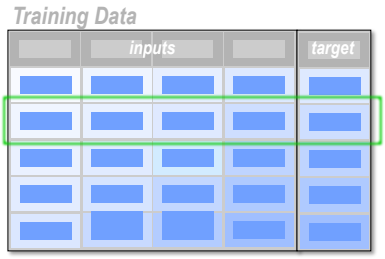
Predictive Modeling Training Data



Training data case: categorical or numeric input and target measurements

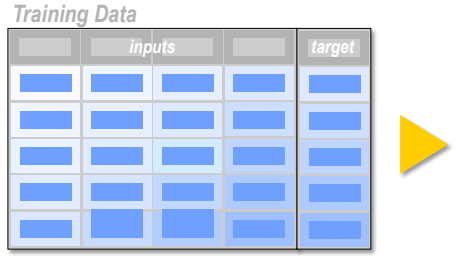
Examples of categorical variables:
race, sex, age group, educational level etc

1. Number of times pregnant.
2. Plasma glucose concentration 2 hours in an oral glucose tolerance test.
3. Diastolic blood pressure (mm Hg).
4. Triceps skin fold thickness (mm).
5. 2-Hour serum insulin (mu U/ml).
6. Body mass index (BMI).
7. Diabetes pedigree function.
8. Age (years).
9. Class, onset of diabetes within five years.

- **Predictive Modeling** (a.k.a. supervised prediction or supervised learning)
- **Training data :** training cases, examples, instances, or records
- **Variables :** inputs, predictors, features, explanatory or independent variables
- **Targets :** response, outcome, or dependent variable

1 ...

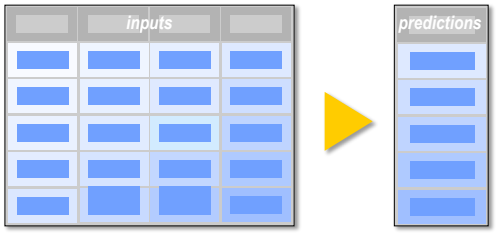
Predictive Model



Predictive model: a concise representation of the input and target association

3 ...

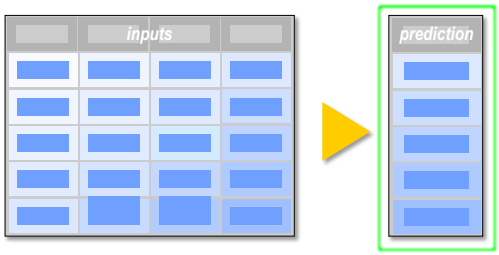
Predictive Model



Predictions: output of the predictive model given a set of input measurements

6 ...

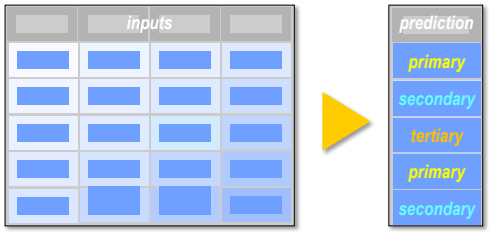
Three Prediction Types



- decisions
- rankings
- estimates

8 ...

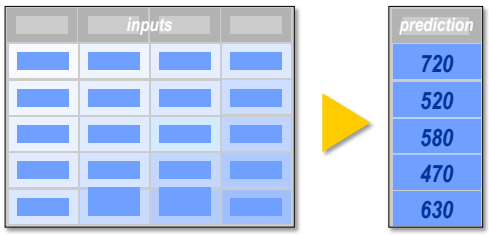
Decision Predictions



A predictive model uses input measurements to make the best decision for each case.

10 ...

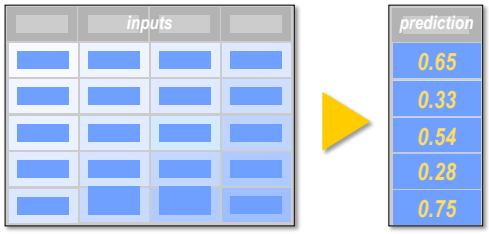
Ranking Predictions



A predictive model uses input measurements to optimally rank each case.

12 ...

Estimate Predictions



A predictive model uses input measurements to optimally estimate the target value.

14 ...

Quiz – Correct Answer

Match the Predictive Modeling Application to the Decision Type.

Predictive Modeling Application	Decision Type
C House Sales	A. Decision
B Risk Profiling	B. Ranking
B Product Ranking (e-Commerce)	C. Estimate
A Cyber Intrusion Detection	
C Revenue Forecasting	
A Voice/Image Recognition	

17

Data Quality: Why Preprocess the Data?

Measures for data quality: A multidimensional view

- Accuracy: correct or wrong, accurate or not
- Completeness: not recorded, unavailable, …
- Consistency: some modified but some not, dangling, …
- Timeliness: timely update?
- Believability: how trustable the data are correct?
- Interpretability: how easily the data can be understood?

18

19

Major Tasks in Data Preprocessing

Data cleaning

- Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies

Data integration

- Integration of multiple databases, data cubes, or files

Data reduction

- Dimensionality reduction – data encoding scheme
- To obtain a reduced representation of the original data
  - Data compression techniques
    - E.g., wavelet transforms and principal components analysis
  - Attribute subset selection
    - E.g., removing irrelevant attributes
  - Attribute construction
    - Where a small set of more useful attributes is derived from the original set

20

Data transformation and data discretization

- Rescaling [0, 1]
- Binnarisation
  - All values above the threshold are marked 1 and all equal to or below are marked as 0.
- Standardisation (Gaussian)
  - Means of 0 and STDEV of 1
- Normalisation
  - A distance-based learning algorithms work better
  - e.g., neural networks, nearest-neighbor, clustering
  - Scaled to a smaller range such as a length of 1 (called unit norm or a vector with length of 1 in linear algebra)
- Concept hierarchy generation (e.g., DT)
  - Raw data values for attributes are replaced by ranges or higher conceptual levels
  - Raw value for *age* may be replaced by higher-level concepts, such as youth, adult, or senior

21

Data Cleaning

Data in the Real World Is Dirty: Lots of potentially incorrect data, e.g., instrument faulty, human or computer error, transmission error

- incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
  - e.g., *Occupation*=“ ” (missing data)
- noisy: containing noise, errors, or outliers
  - e.g., *Salary*=“−10” (an error)
- inconsistent: containing discrepancies in codes or names, e.g.,
  - *Age*=“42”, *Birthday*=“03/07/2010”
  - Was rating “1, 2, 3”, now rating “A, B, C”
  - discrepancy between duplicate records
- Intentional (e.g., *disguised missing* data)
  - Jan. 1 as everyone’s birthday?

22

Incomplete (Missing) Data

Data is not always available

- E.g., many tuples have no recorded value for several attributes, such as age, number of times pregnanct, salary

Missing data may be due to

- equipment malfunction
- inconsistent with other recorded data and thus deleted
- data not entered due to misunderstanding
- certain data may not be considered important at the time of entry
- not register history or changes of the data

Missing data may need to be inferred

23

Simple Prediction Illustration – Regressions

$$\text{logit}(\hat{p}) = \hat{w}_0 + \hat{w}_1 x_1 + \hat{w}_2 x_2$$

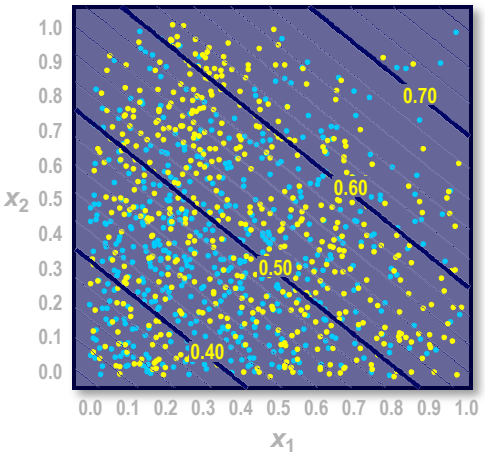
$$\hat{p} = \frac{1}{1 + e^{-\text{logit}(\hat{p})}}$$

Find parameter estimates by *maximizing*

$$\sum \log(\hat{p}_i) + \sum \log(1 - \hat{p}_i)$$

primary outcome training cases — secondary outcome training cases

log-likelihood function



Simple Prediction Illustration – Regressions

$$\text{logit}(\hat{p}) = \hat{w}_0 + \hat{w}_1 x_1 + \hat{w}_2 x_2$$

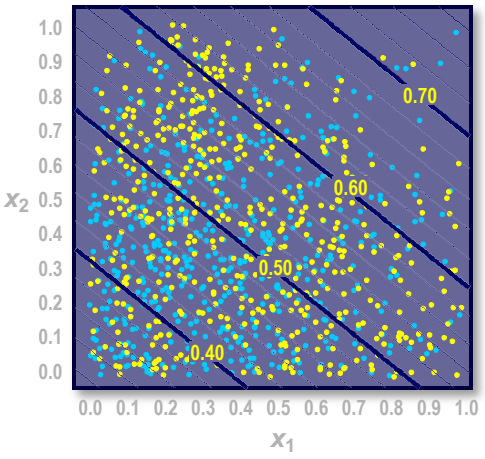
$$\hat{p} = \frac{1}{1 + e^{-\text{logit}(\hat{p})}}$$

Find parameter estimates by *maximizing*

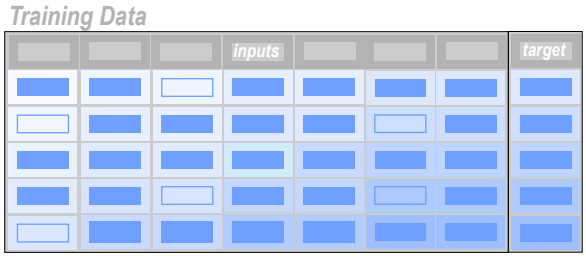
$$\sum \log(\hat{p}_i) + \sum \log(1 - \hat{p}_i)$$

primary outcome training cases — secondary outcome training cases

log-likelihood function



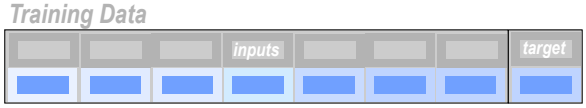
Missing Values and Regression Modeling



Problem 1: Training data cases with missing values on inputs used by a regression model are ignored.

26 ...

Missing Values and Regression Modeling



Consequence: Missing values can significantly reduce your amount of training data for regression modeling!

28 ...

Missing Values and the Prediction Formula

$$\text{logit}(\hat{p}) = -0.81 + 0.92 \cdot x_1 + 1.11 \cdot x_2$$

Predict: (x1, x2) = (0.3, ?)

Problem 2: Prediction formulas cannot score cases with missing values.

29 ...

Missing Value Issues

- Manage missing values.

Problem 1: Training data cases with missing values on inputs used by a regression model are ignored.

Problem 2: Prediction formulas cannot score cases with missing values.

33 ...

Missing Value Remedies

Manage missing values.

- **Ignore the tuple:** usually done when class label is missing (when doing classification) — not effective when the % of missing values per attribute varies considerably
- **Fill in the missing value manually:** tedious + infeasible?
- **Fill in it automatically with**
  - a global constant : e.g., "unknown", a new class?!
  - the attribute mean
  - the attribute mean for all samples belonging to the same class: smarter
  - the most probable value: inference-based such as Bayesian formula or decision tree

Synthetic distribution

Estimation

$x_i = f(x_1, \ldots, x_p)$

35

...

Noisy Data

Noise: random error or variance in a measured variable

Incorrect attribute values may be due to

- faulty data collection instruments
- data entry problems
- data transmission problems
- technology limitation
- inconsistency in naming convention

Other data problems which require data cleaning

- duplicate records
- incomplete data
- inconsistent data

36

36

How to Handle Noisy Data?

Binning

- first sort data and partition into (equal-frequency) bins
- then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.

Regression

- smooth by fitting the data into regression functions

Clustering

- detect and remove outliers

Combined computer and human inspection

- detect suspicious values and check by human (e.g., deal with possible outliers)

37

Simple Discretization: Binning

Equal-width (distance) partitioning

- Divides the range into *N* intervals of equal size: uniform grid
- if *A* and *B* are the lowest and highest values of the attribute, the width of intervals will be: $W = (B - A)/N$.
- The most straightforward, but outliers may dominate presentation
- Skewed data is not handled well

Equal-depth (frequency) partitioning

- Divides the range into *N* intervals, each containing approximately same number of samples
- Good data scaling
- Managing categorical attributes can be tricky

38

Binning Methods for Data Smoothing

Sorted data for price (in pounds): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

- Partition into equal-frequency (**equi-depth**) bins:
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34
- Smoothing by **bin means**:
  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29
- Smoothing by **bin boundaries**:
  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34

39

How might you determine outliers in the data?

- Outliers in the data may be detected by clustering, where similar values are organized into groups, or 'clusters'.
- Values that fall outside of the set of clusters may be considered outliers.
- Alternatively, a combination of computer and human inspection can be used where a predetermined data distribution is implemented to allow the computer to identify possible outliers.
- These possible outliers can then be verified by human inspection with much less effort than would be required to verify the entire initial data set.

41

What other methods are there for data smoothing?

Alternate forms of binning

- smoothing by bin means, medians, modes
- smoothing by bin boundaries
- equi-width bins can be used to implement any of the forms of binning, where the interval range of values in each bin is constant.

Mean=the average.
Median= the middle number. You line up the numbers in order from smallest to largest, and cross one out on each side at a time.
Mode=the most common or frequent number.

42

What other methods are there for data smoothing?

Alternate forms of binning

- smoothing by bin medians
- smoothing by bin boundaries
- equi-width bins can be used to implement any of the forms of binning, where the interval range of values in each bin is constant.

Regression techniques

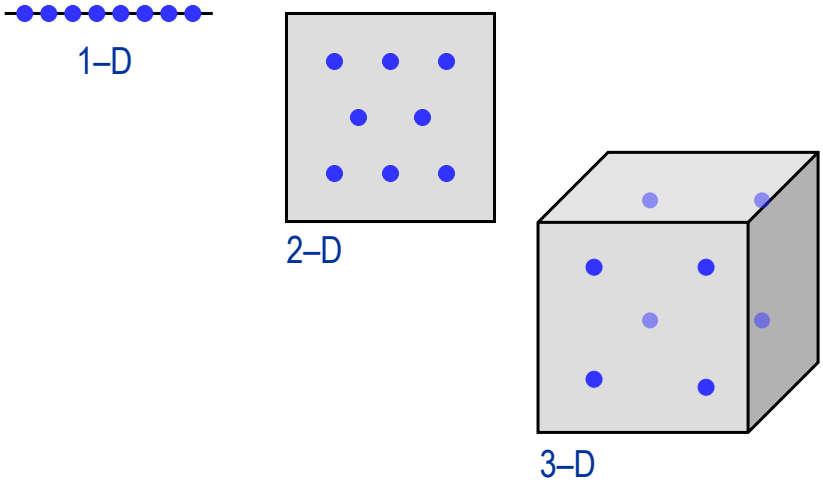
- smooth the data by fitting it to a function such as through linear or multiple regression

Concept hierarchies

- can smooth the data by rolling-up lower level concepts to higher-level concepts

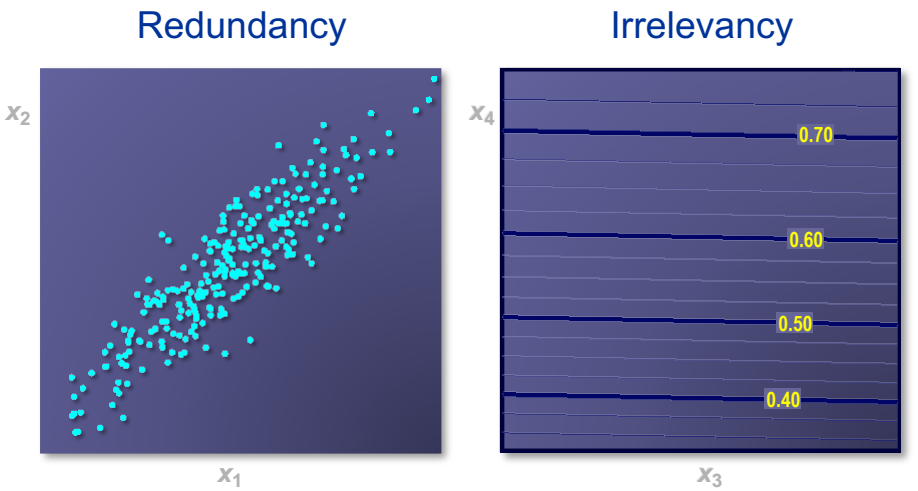
43

The Curse of Dimensionality



44

Input Reduction – Redundancy



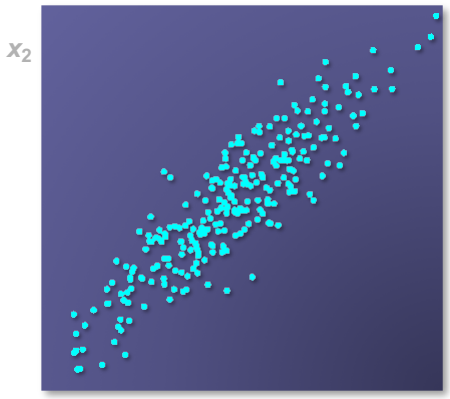
45

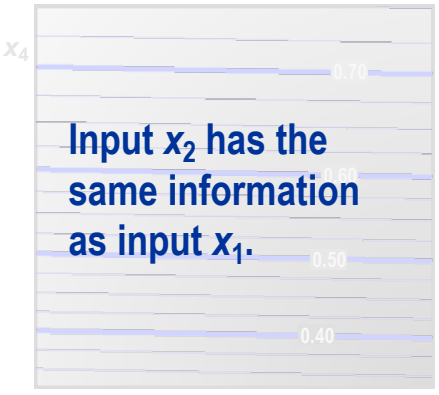
...

Input Reduction – Redundancy

Redundancy

Irrelevancy





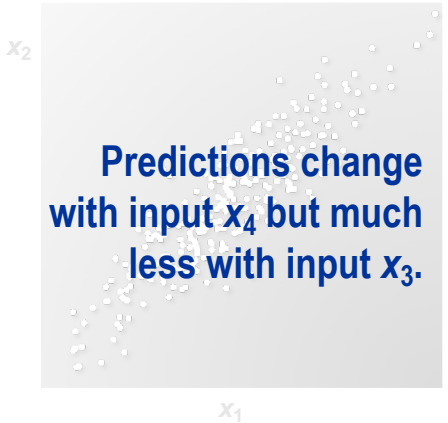
Input x_2 has the same information as input x_1.

46 ...

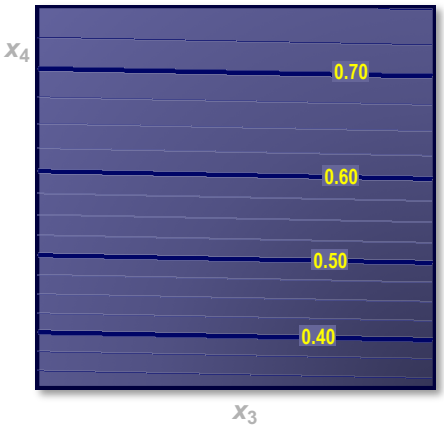
Input Reduction – Irrelevancy

Redundancy

Irrelevancy



Predictions change with input x_4 but much less with input x_3.



47 ...

Lab (414-417)

Visualisation using Pandas and MatPlotLib

Univariate Plots

We will look at three techniques that you can use to understand each attribute of your dataset independently.

- Histograms
- Density Plots
- Box and Whisker Plots

Multivariate Plots

We will look at the examples of two plots that show the interactions between multiple variables in your dataset.

- Correlation Matrix Plot.
- Scatter Plot Matrix.

Data Transformation using SciKit-Learn

- Recale
- Standardise
- Normalise
- Binarise

48

Challenge Question!

(o) [Challenge!] *Run a Decision Tree algorithm on both the raw and the normalised data and compare their results. The output should look like the below.*

```
Normalised Samples
 [[0.034 0.828 0.403 0.196 0.    0.188 0.004 0.28 ]
 [0.008 0.716 0.556 0.244 0.    0.224 0.003 0.261]
 [0.04  0.924 0.323 0.    0.    0.118 0.003 0.162]
 [0.007 0.588 0.436 0.152 0.622 0.186 0.001 0.139]
 [0.    0.596 0.174 0.152 0.731 0.188 0.01  0.144]]
Mean estimated accuracy
 0.6874572795625428
Mean estimated accuracy on normalised data
 0.6378332194121668
```

You may use the below codes.

```
# Decision tree classification
from sklearn.model_selection import KFold
from sklearn.model_selection import cross_val_score
from sklearn.tree import DecisionTreeClassifier
kfold = KFold(n_splits=10, random_state=7)
model = DecisionTreeClassifier()
results = cross_val_score(model, X, Y, cv=kfold)
print("Mean estimated accuracy \n",results.mean())
```

49