Principles of Data Mining

Max Bramer

Introduction

- Supervised learning :

  Clasification, numerical prediction
- Unsupervised learning:

  Association rules, clustering

Data for dataming

- Datasets, ensamble of variables of a *universe of objects* that are of interest

SoftEng	ARIN	HCI	CSA	Project	Class
A	B	A	B	B	Second
A	B	B	B	B	Second
B	A	A	B	A	Second
A	A	A	A	B	First
A	A	B	B	A	First
B	A	A	B	B	Second
....	....	....	....	....	....
A	A	B	A	B	First

Types of variables

- Nominals
- Binary
- Ordinal
- Integer
- Interval-scaled
- Ratio-scaled

categorical corresponding to nominal, binary and ordinal variables

continuous corresponding to integer, interval-scaled and ratio-scaled variables.

Data preparation

- Data cleaning:
  - Missig values
    - Descartar instancias
    - Replace for the most frequent/average value
    - Usar reglas de asociación
  - Reducing the Number of Attributes

Data cleaning

Reducing the Number of Attributes

- Feature reduction- Dimension reduction

Introduction to clasification

- Clasification

Assign objects to one of a number of mutually exhaustive and exclusive categories known as *classes*.

- – People who are at high, medium or low risk of a car accident in the next 12months
- – People who are likely to vote for each of a number of political parties (or none)
- – The likelihood of rain the next day for a weather forecast (very likely, likely, unlikely, very unlikely).

Naive Bayes Classifiers

Atributes are nominal

Usually we are interested in a set of alternative possible events, which are *mutually exclusive* and *exhaustive*, meaning that one and only one must always occur.

In a train example, we might define four mutually exclusive and exhaustive events

- $E1$ – train cancelled
- $E2$ – train ten minutes or more late
- $E3$ – train less than ten minutes late
- $E4$ – train on time or early.
- The probability of an event is usually indicated by a capital letter P, so we might have
- $P(E1) = 0.05$
- $P(E2) = 0.1$
- $P(E3) = 0.15$
- $P(E4) = 0.7$: $P(E1) + P(E2) + P(E3) + P(E4) = 1$

Naive Bayes Classifiers

day	season	wind	rain	class
weekday	spring	none	none	on time
weekday	winter	none	slight	on time
weekday	winter	none	slight	on time
weekday	winter	high	heavy	late
saturday	summer	normal	none	on time
weekday	autumn	normal	none	very late
holiday	summer	high	slight	on time
sunday	summer	normal	none	on time
weekday	winter	high	heavy	very late
weekday	summer	none	slight	on time
saturday	spring	high	heavy	cancelled
weekday	summer	high	slight	on time
saturday	winter	normal	none	late
weekday	summer	high	none	on time
weekday	winter	normal	heavy	very late
saturday	autumn	high	slight	on time
weekday	autumn	none	heavy	on time
holiday	spring	normal	slight	on time
weekday	spring	normal	none	on time
weekday	spring	normal	slight	on time

Figure 3.1 The *train* Dataset

Naive Bayes Classifiers

Problem

day | season | wind | rain | class

weekday	winter	high	heavy	????

$$P(c_i) \times \prod_{j=1}^{n} P(a_j = v_j \mid class = c_i).$$

Conditional and Prior Probabilities: *train* Dataset

	class = on time	class = late	class = very late	class = cancelled
day = weekday	**9/14 = 0.64**	**1/2 = 0.5**	**3/3 = 1**	**0/1 = 0**
day = saturday	2/14 = 0.14	1/2 = 0.5	0/3 = 0	1/1 = 1
day = sunday	1/14 = 0.07	0/2 = 0	0/3 = 0	0/1 = 0
day = holiday	2/14 = 0.14	0/2 = 0	0/3 = 0	0/1 = 0
season = spring	4/14 = 0.29	0/2 = 0	0/3 = 0	1/1 = 1
season = summer	6/14 = 0.43	0/2 = 0	0/3 = 0	0/1 = 0
season = autumn	2/14 = 0.14	0/2 = 0	1/3 = 0.33	0/1 = 0
season = winter	**2/14 = 0.14**	**2/2 = 1**	**2/3 = 0.67**	**0/1 = 0**
wind = none	5/14 = 0.36	0/2 = 0	0/3 = 0	0/1 = 0
wind = high	**4/14 = 0.29**	**1/2 = 0.5**	**1/3 = 0.33**	**1/1 = 1**
wind = normal	5/14 = 0.36	1/2 = 0.5	2/3 = 0.67	0/1 = 0
rain = none	5/14 = 0.36	1/2 =0.5	1/3 = 0.33	0/1 = 0
rain = slight	8/14 = 0.57	0/2 = 0	0/3 = 0	0/1 = 0
rain = heavy	**1/14 = 0.07**	**1/2 = 0.5**	**2/3 = 0.67**	**1/1 = 1**
Prior Probability	**14/20 = 0.70**	**2/20 = 0.10**	**3/20 = 0.15**	**1/20 = 0.05**

Naive Bayes Algorithm

$\underline{\text{class} = \text{on time}}$

$$0.70 \times 0.64 \times 0.14 \times 0.29 \times 0.07 = 0.0013$$

$\underline{\text{class} = \text{late}}$

$$0.10 \times 0.50 \times 1.00 \times 0.50 \times 0.50 = 0.0125$$

$\underline{\text{class} = \text{very late}}$

$$0.15 \times 1.00 \times 0.67 \times 0.33 \times 0.67 = 0.0222$$

$\underline{\text{class} = \text{cancelled}}$

$$0.05 \times 0.00 \times 0.00 \times 1.00 \times 1.00 = 0.0000$$

Solution

$$\underline{\text{class} = \text{very late}}$$

$$0.15 \times 1.00 \times 0.67 \times 0.33 \times 0.67 = 0.0222$$

Nearest Neighbour Classification

Basic k-Nearest Neighbour Classification Algorithm

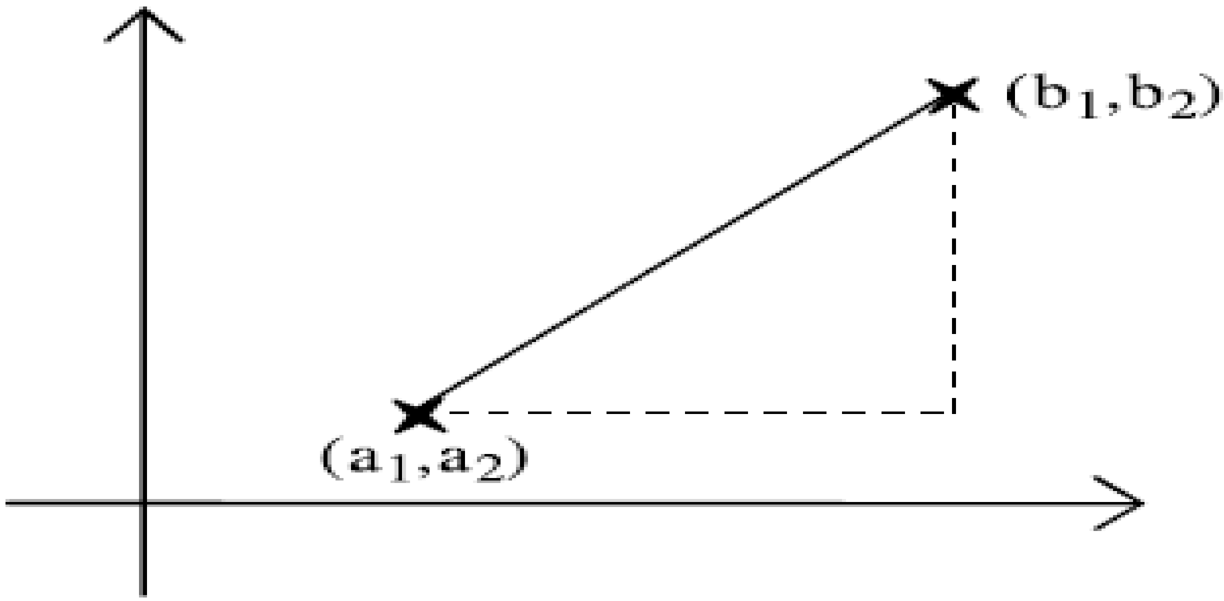
– Find the k training instances that are closest to the unseen instance.

– Take the most commonly occurring classification for these k instances.

Useful with numerical attributs

Distance Measures

- Euclidean Distance



$$\sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \ldots + (a_n - b_n)^2}$$

Normalisation

In general if the lowest value of attribute A is min and the highest value is max, we convert each value of A, say a, to $(a - min)/(max - min)$.

Adjustement of the Euclidean distance

$$\sqrt{w_1(a_1 - b_1)^2 + w_2(a_2 - b_2)^2 + \ldots + w_n(a_n - b_n)^2}$$