Computer Science 477

Naïve Bayes and Nearest Neighbor Classification

Lecture 4

Classification

- Dividing up objects so that each is assigned to one of a number of mutually exhaustive and exclusive categories.
- To devise a scheme for classifying new instances we use a training set of existing (past) labeled instances.
- By abstracting known classification of existing instance, develop a predictive mechanism
- First method: classic Bayesian probabilities

Probability

(Kolmogorov's axioms,
first published in German 1933)

- All probabilities are between 0 and 1. For any proposition *a*, $0 \leq P(a) \leq 1$
- $P(true)=1, \quad P(false)=0$

The probability of disjunction is given by

$$P(a \vee b) = P(a) + P(b) - P(a \wedge b)$$

- Product rule

$$P(a \wedge b) = P(a \mid b)P(b)$$

$$P(a \wedge b) = P(b \mid a)P(a)$$

Theorem of total probability

- If events $A_1, \ldots, A_n$ are mutually exclusive with

$$\sum_{i=1}^{n} P(A_i) = 1$$

then

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$$

$$P(B) = \sum_{i=1}^{n} P(B, A_i)$$

Bayes's rule

- (The Reverend Thomas Bayes 1702-1761)
- He set down his findings on probability in "Essay Towards Solving a Problem in the Doctrine of Chances" (1763), published posthumously in the *Philosophical Transactions of the Royal Society of London*

$$P(b \mid a) = \frac{P(a \mid b)P(b)}{P(a)}$$

Train History

- Historical database of train performance
- How us probabilities to classify new instance:

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

weekday	winter	high	heavy	????

Pick the Majority

- Choose the most frequent classification.
- Train is on time more than any other classification
  - Correct 70% of the time (historically)
- Does not take advantage of the accumulated information
- But might be as good as you can do.
- Alternative: Use conditional probabilities.
- Example: probability that class = on time given that season = winter.

Conditional Probabilities

- The probability of an event, given the occurrence of some other event is conditional probability
- Written as, e.g.

$$P(\text{class} = \text{ on time} \mid \text{season} = \text{ winter})$$

- Consulting the table:
- $\dfrac{\text{Class = on time}}{\text{season = winter}} = \dfrac{2}{6} = 0.33$
- $P(\text{class} = \text{on time}|\text{season} = \text{winter}) = \dfrac{2}{6} = 0.33$
- $P(\text{late}|\text{season} = \text{winter}) = \dfrac{1}{6} = 0.17$
- $P(\text{class} = \text{very late}|\text{season} = \text{winter}) = \dfrac{3}{6} = 0.5$
- $P(\text{class} = \text{cancelled}|\text{season} = \text{winter}) = \dfrac{0}{6} = 0$
- Note that **very late** is the largest (0.5) so might conclude that most likely classification is **very late**.
  - Different from the calculated prior probability

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

Conditional Probabilities - Naïve Bayes

- For

weekday	winter	high	heavy	????

- Calculate

$$P(\text{class} = \text{on time} \mid \text{day} = \text{weekday and season} = \text{winter and wind} = \text{high and rain} = \text{heavy})$$

- There are only two instances with this combination of attribute values
- The Naïve Bayes algorithm provides a scheme for combining prior probabilities and conditional probabilities in a single formula
- Also uses conditional probabilities, but differently

Naïve Bayes

- Instead, for example, of concluding that the class is very late given that the season is winter

  $$\mathrm{P}(\text{class} = \text{very late} | \text{season} = \text{winter})$$

  calculate the probability that the season is winter given that the class is very late

  $$\mathrm{P}(\text{season} = \text{winter} | \text{class} = \text{very late})$$

- Calculated as the number of times **season=winter** and **class=very late** occur in the same instance, divided by the number of times the class is **very late**
- Similarly, calculate other conditional probabilities, e.g., $\mathrm{P}(\text{rain} = \text{none} | \text{class} = \text{very late})$

Conditional and Prior Probabilities

- Conditional probability $P(day = weekday|class = on\ time)$ – number of instances for which **day=weekday** and **class=on time**, divided by the total number of instances for which the **class=on time**

- Number of instances for which **day=weekday** is 9 and **class=on time**
- Number of instances for which day=weekday is 14
- $\frac{9}{14} = 0.64$
- Prior probability of class=very late divided by the total number of instances, i.e., $\frac{3}{20} = 0.25$

	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

Bayes Theorem

- Now calculate the probabilities of interest
- Posterior probabilities of each possible class occurring for a specified instance, for know values of the attributes.
- Given a set of k mutually exclusive and exhaustive classifications $c_1, c_2, \dots, c_k$, which have prior probabilities $P(c1), P(c_2), \dots, P(c_k)$, respectively, and n attributes $a_1, a_2, \dots, a_n$ which for a given instance have values $v_1, v_2, \dots, v_n$ respectively, the posterior probability of class c_i occurring for the specified instance can be shown to be proportional to

$$P(c_i) \times P(a_1 = v_1 \text{ and } a_2 = v_2 \dots \text{ and } a_n = v_n \mid c_i)$$

- Making the assumption that the attributes are independent, the value of this expression can be calculated using the product

$$P(c_i) \times P(a_1 = v_1 \mid c_i) \times P(a_2 = v_2 \mid c_i) \times \dots \times P(a_n = v_n \mid c_i)$$

- We calculate this product for each value of i from 1 to k and choose the classification that has the largest value.

Naïve Bayes

$$P(c_i) \times P(a_1 = v_1|c_i) \times P(a_2$$

$$= v_2|c_i) \times \ldots \times P(a_n = v_n \mid c_i)$$

- Also written (using Π-notation) as

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

Naïve Bayes

- Given conditions

weekday	winter	high	heavy	????

- What is the probability that the train will be on time?
- On time – 0.70
- Weekday | on time – 0.64
- Winter | on time – 0.14
- High | on time 0.29
- Heavy | on time 0.07
- $0.70 \times 0.64 \times 0.14 \times 0.29 \times 0.07$
  $= 0.0013$

	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

Naïve Bayes

- Given conditions

weekday	winter	high	heavy	????

- What is the probability that the train will be late?
- Late – 0.10
- Weekday | late – 0.50
- Winter | late – 1.00
- High | late 0.50
- Heavy | late 0.50
- $0.10 \times 0.50 \times 1.00 \times 0.50 \times 0.50$

  $= 0.0125$

	class = on time	class = late	class = very late	class = can-celled
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**

Naïve Bayes

- Given conditions

weekday	winter	high	heavy	????

- What is the probability that the train will be cancelled?
- Cancelled – 0.05
- Weekday | cancelled – 0.00
- Winter | cancelled – 0.00
- High | cancelled 1.00
- Heavy | cancelled 1.00
- $0.05 \times 0.00 \times 0.00 \times 1.00 \times 1.00$

  $= 0.0000$

	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

Naïve Bayes

- Given conditions

weekday	winter	high	heavy	????

- What is the probability that the train will be very late?
- Very late – 0.15
- Weekday | very late – 0.10
- Winter | very late – 0.67
- High | very late 0.33
- Heavy | very late 0.67
- $0.15 \times 1.00 \times 0.67 \times 0.33 \times 0.67$

  $= 0.0222$

	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**

Naïve Bayes

- Given conditions

weekday	summer	high	heavy	????

- What is the probability that the train will be very late?
- Very late – 0.15
- Weekday | very late – 0.10
- Summer | very late – 0.00
- High | very late 0.33
- Heavy | very late 0.67
- $0.15 \times 1.00 \times 0.00 \times 0.33 \times 0.67$

  $= 0.00$

	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**

Nearest Neighbor Classification

- Mainly used when all attribute values are continuous
  - Can be modified to deal with categorical attributes.
- Idea: estimate the classification of an unseen instance using the classification of the instance or instances that are *closest* to it
  - Most similar to it

Example

- Suppose a training set with just two instances:

a	b	c	d	e	f	Class
yes	no	no	6.4	8.3	low	negative
yes	yes	yes	18.2	4.7	high	positive

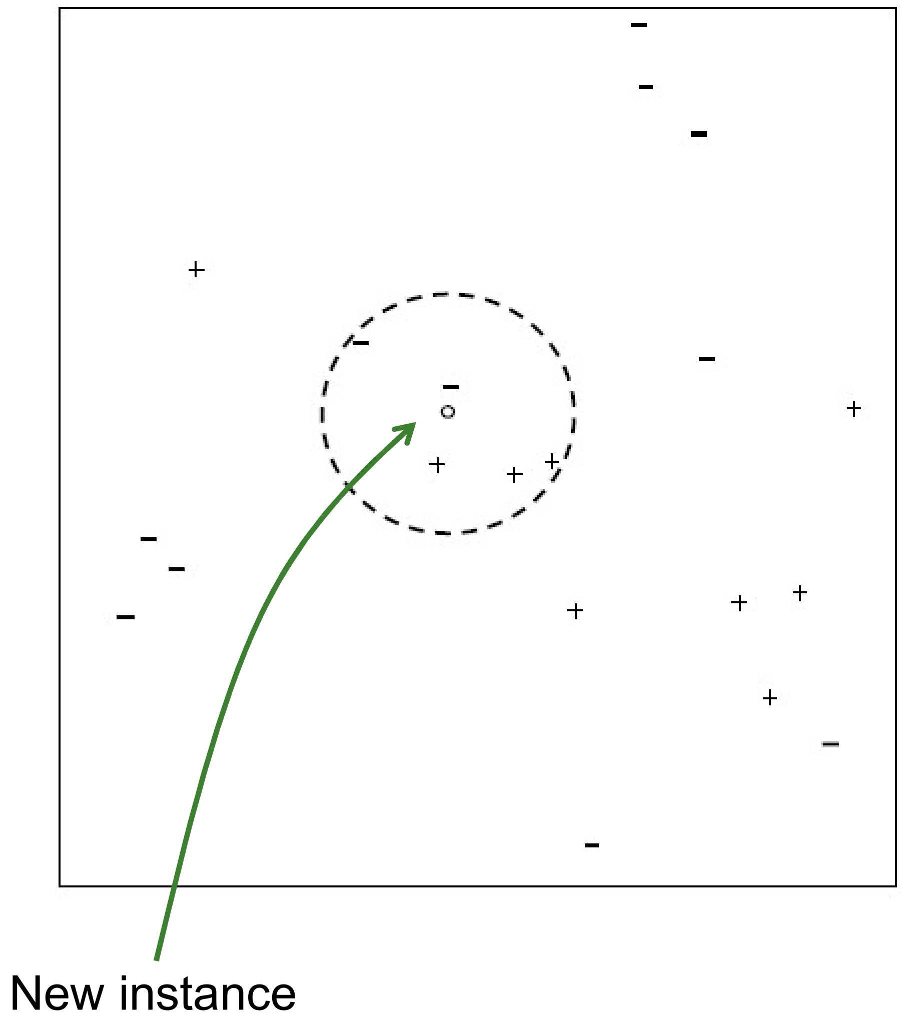
- Presented with new instance:

yes	no	no	6.6	8.0	low	???

- Resembles, intuitively, negative instance
- Hence, classify it as negative.
- General strategy:
  - Find the k training instances that are closest to the unseen instance
  - Take the most commonly occurring classification for these k instances.

Example Training Set

Attribute 1	Attribute 2	Class
0.8	6.3	−
1.4	8.1	−
2.1	7.4	−
2.6	14.3	+
6.8	12.6	−
8.8	9.8	+
9.2	11.6	−
10.8	9.6	+
11.8	9.9	+
12.4	6.5	+
12.8	1.1	−
14.0	19.9	−
14.2	18.5	−
15.6	17.4	−
15.8	12.2	−
16.6	6.7	+
17.4	4.5	+
18.2	6.9	+
19.0	3.4	−
19.6	11.1	+



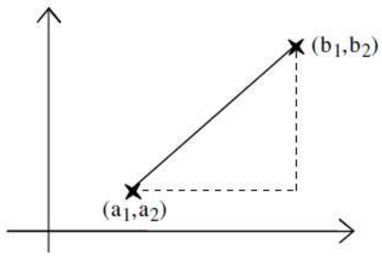
Common Constraint on Distance Measures

- Previous example had two attributes, dimensions
  - Can be Visualized
- Can be extended to *n*-dimensions
- Presuppose *distance measure*
- Usually – not always - impose three requirements:.
  - **dist**(*A*,*A*) = 0.
  - Symmetry condition**:**
    - **dist**(*A*,*B*) = **dist**(*B*,*A*) (the *symmetry condition*).
  - *Triangle inequality:*
    - **dist**(*A*,*B*) ≤ **dist**(*A*,*Z*) + **dist**(*Z*,*B*).

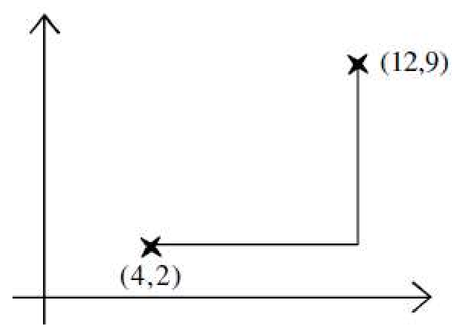
Distance Measures

- *Euclidean distance* between points $(a_1, a_2, \ldots, an)$ and $(b_1, b_2, \ldots, b_n)$ in *n*-dimensional space is

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



- *Manhattan distance*:
- Distance between the points (4, 2) and (12, 9) in Figure 2.9 is (12 − 4) + (9 − 2) = 8 + 7 = 15.



Maximum Dimension Distance

- Largest absolute difference between any pair of corresponding attribute values.
  - Absolute difference is the difference converted to a positive number if it is negative.
- Example:

6.2	−7.1	−5.0	18.3	−3.1	8.9

8.3	12.4	−4.1	19.7	−6.2	12.4

- Maximum Dimension Distance:

$$12.4 - (-7.1) = 19.5$$

Distance Measures

- Euclidean Distance
- Cosine Similarity
- Hamming Distance
- Manhattan Distance
- Chebyshev Distance
- Minkowski Distance
- Jaccard Distance
- Haversine
- Sørensen-Dice Index

Normalization

Mileage (miles)	Number of doors	Age (years)	Number of owners
18,457	2	12	8
26,292	4	3	1

- Millage dominates
  - Millage and Age *not* independent

- Normalize all values

- Lowest value of attribute *A* in training set is *min* and the highest value is *max*, we convert each value of *A*, say *a*, to (*a* − *min*)/(*max* − *min*).

Categorical Attributes

- Weakness of the nearest neighbor approach - no entirely satisfactory way of dealing with categorical attributes.
- One possibility is to say that the difference between any two identical values of the attribute is zero and that the difference between any two different values is 1.
  - Amounts to saying (for a color attribute) red − red = 0, red − blue = 1, blue − green = 1, etc.
- Sometimes there is an ordering (or partial ordering) of the values of an attribute
  - Might have values good, average and bad.
  - Can treat the difference between good and average or between average and bad as 0.5 and the difference between good and bad as 1.
  - May be the best we can do in practice.