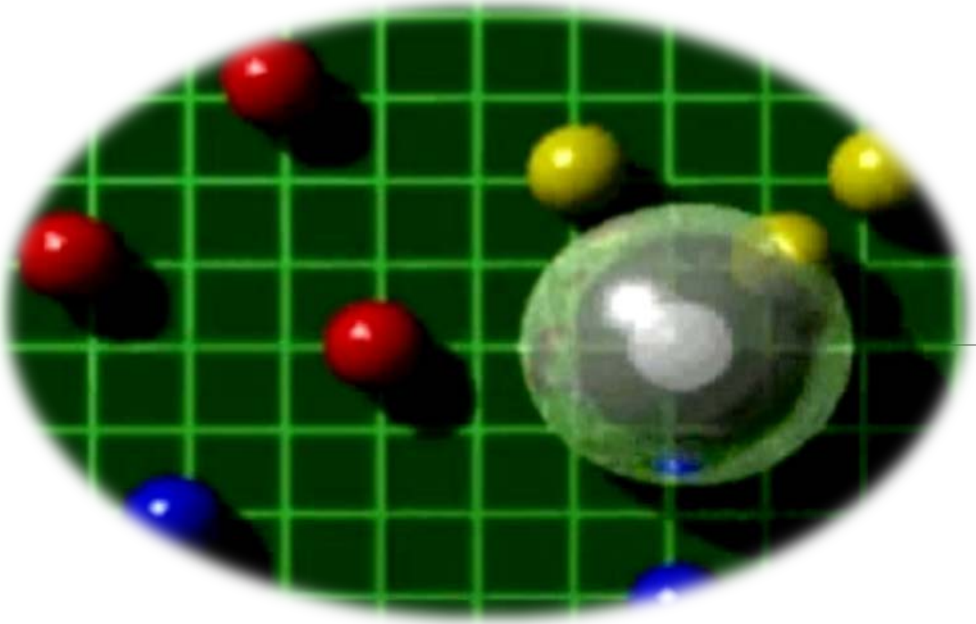
STATISTICAL METHODS IN DATA MINING

DR. ALPER VAHAPLAR

Previously on Course…

- ✓Exploring Data,
- ✓Summary Statistics,
- ✓Data Visualization,
- ✓Measure of Similarity,
- ✓Hierarchical Clustering,
- ✓K-means Clustering
- ✓Density Based Clustering
- ✓Grid Based Clustering
- ✓Model Based clustering

Today



...

- ✓Training on k-means
- ✓Classification
- ✓K-nearest neighborhood
- ✓Bayesian Classification

Supervised – Unsupervised Learning

- ✓ **Supervised learning**
  - is a ***machine learning*** technique for creating a function from training data.
  - The ***training data*** consist input objects (typically vectors), and desired outputs.
  - The output can be a continuous value (called ***regression***), or can predict a class label of the input object (called ***classification***).
  - The task of the supervised learner is to predict the value of the function for any valid input object after having seen a number of training examples.
  - The learner has to generalize from the presented data to unseen situations in a "reasonable" way
- ✓ **Unsupervised learning**
  - is a method of ***machine learning*** where a model is fit to observations.
  - It is distinguished from ***supervised learning*** by the fact that there is ***no a priori*** output.
  - In unsupervised learning, a data set of input objects is gathered. Unsupervised learning then typically treats input objects as a set of ***random variables***. A joint density model is then built for the data set.

Classification

- ✓ The task of assigning *previously unseen* objects to one of several *predefined categories.*
- ✓ Finding a model for class attribute as a function of other attributes.
- ✓ Predicts categorical labels (unlike estimation or prediction).
- ✓ Is a 2-step process:

1. **Model construction**
   - Each tuple/sample is assumed to belong to a predefined class, as determined by the class label attribute,
   - The set of tuples used for model construction is training set,
   - The model is represented as classification rules, trees, or mathematical formulae.
2. **Model usage** (Classifying future or unknown objects)
   - Estimate accuracy rate of the model on a test set,
   - If the accuracy is acceptable, use the model to classify data tuples whose class labels are not known.

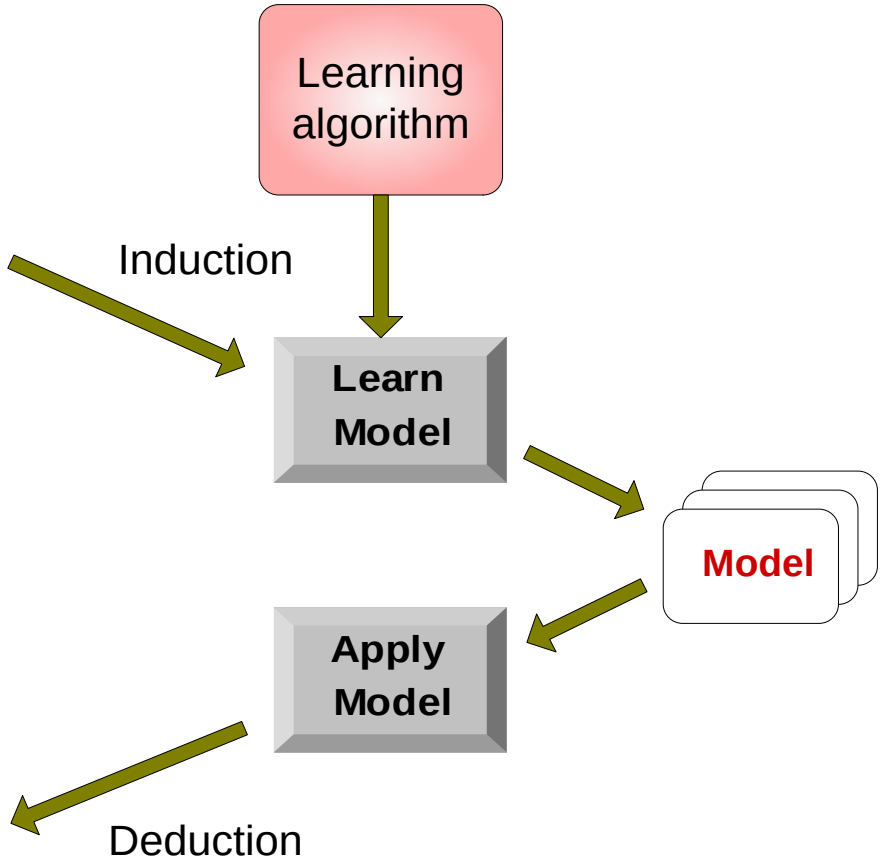
Illustrating Classification Task

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

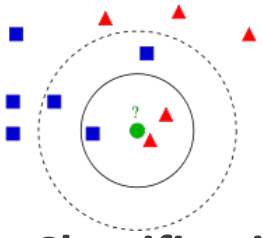
Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set



Classification Techniques

1. **K-Nearest Neighbor**

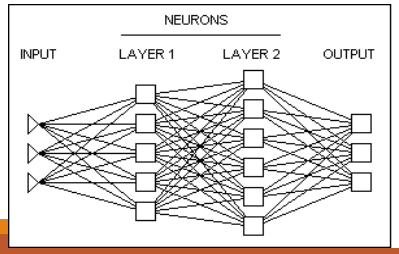


2. **Bayesian Classification**

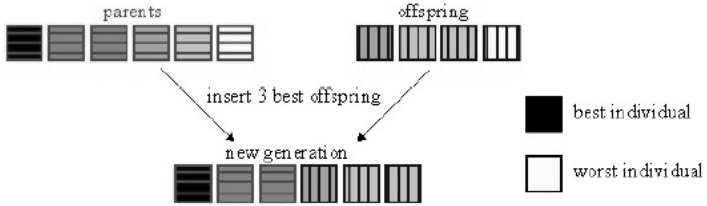
$$c = \max_{c_j} \frac{p(c_j)}{p(d)} \prod_{i=1}^{n} p(a_i \mid c_j)$$

3. **Decision Trees**

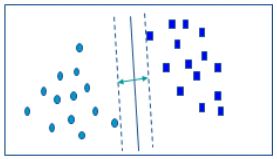
4. **Neural Networks**



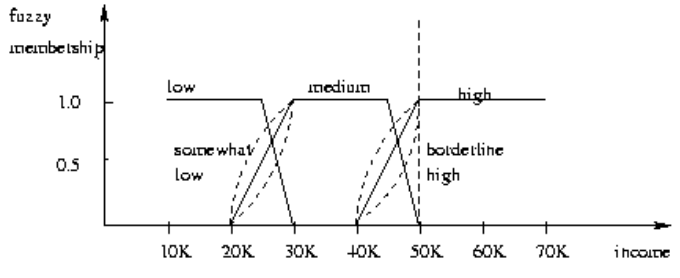
5. **Genetic Algorithms**



6. **Support Vector Machines (SVM)**

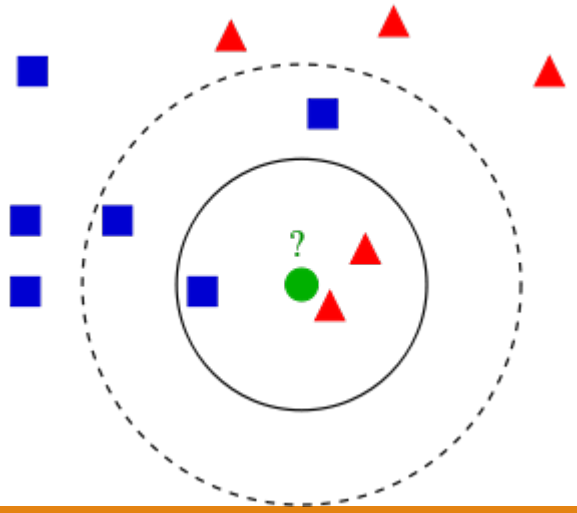


7. **Fuzzy Set Approaches**

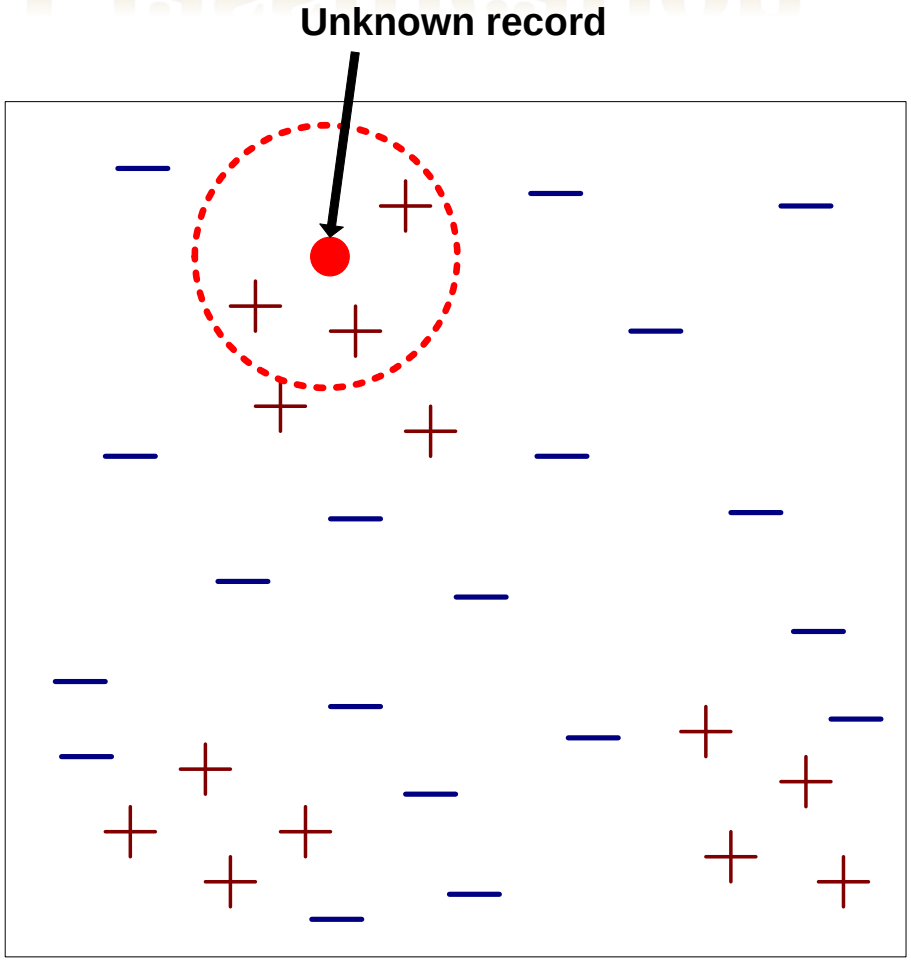


k-nearest Neighborhood Classification

- ✓ Is an example of instance based learning,
- ✓ Lazy learner – not an eager learner ,
- ✓ Training data set is stored,
- ✓ Classification for a new unclassified record is found by comparing it to ***k*** most similar records in the training set.
- ✓ If *k* = 1, then the object is simply assigned to the class of its nearest neighbor.

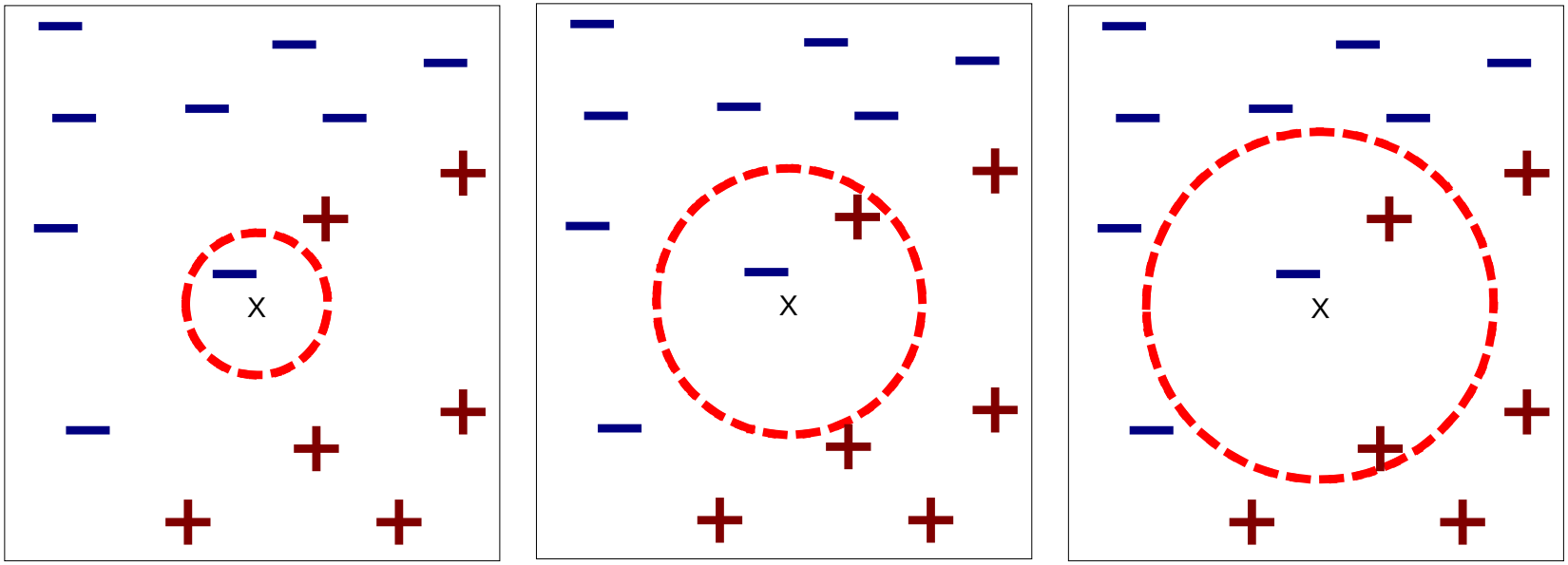


k-nearest Neighborhood Classification



- Requires three things
  - 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)

k-nearest Neighborhood Classification

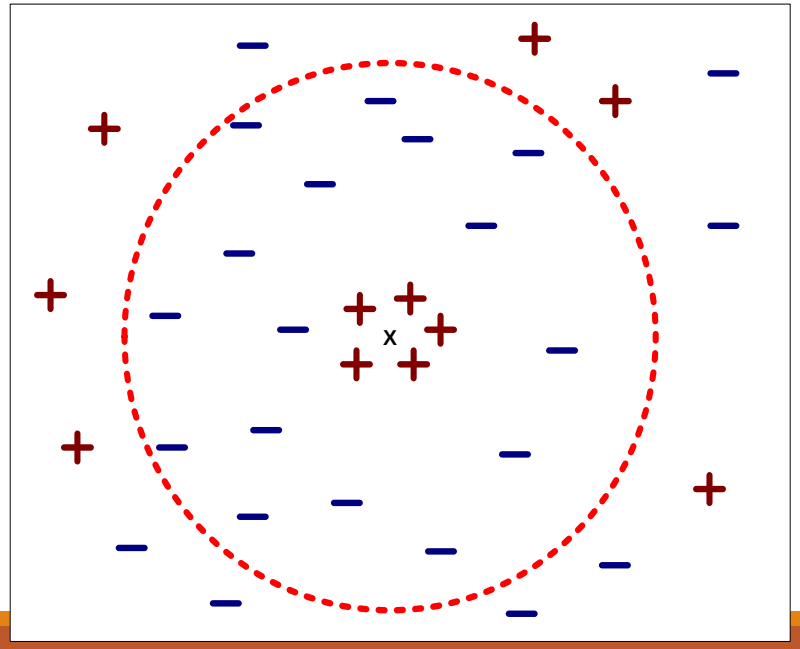


(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

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

k-nearest Neighborhood Classification

- ✓ Choosing the value of k:
  - If k is too small, sensitive to outliers or noise.
  - If k is too large, locally interesting behaviour will be overlooked.



k-nearest Neighborhood Classification

- ✓ Advantages:
  - No model is built,
  - Building model is cheap,
  - Simple technique, easily implemented,
  - Well suited for records with multiple class labels,
  - Can sometimes be the best method
- ✓ Disadvantages:
  - Hard to decide *k*,
  - Requires computation of a distance for all new records.

K-NN Example

- ✓ Using iris data find the class of the following for k=2, k=3, and k=4.

sepal-length	sepal-width	petal-length	petal-width	class
5.1	2.6	5.5	1.1	???

k-NN for Estimation and Prediction

- ✓ k-NN may be used for estimation and prediction as well as for *continuous* valued target variables.
- ✓ Locally Weighted Averaging

$$\hat{y}_{new} = \frac{\sum_i w_i y_i}{\sum_i w_i}$$

$$w_i = 1/\mathrm{d}(\mathrm{new}, x_i)^2$$

k-NN for Estimation and Prediction

Record	Age	Na/K	BP	Age_{MMN}	Na/K_{MMN}	Distance
New	17	12.5	?	0.05	0.25	—
A	16.8	12.4	120	0.0467	0.2471	0.009305
B	17.2	10.5	122	0.0533	0.1912	0.17643
C	19.5	13.5	130	0.0917	0.2794	0.09756

$$\hat{y}_{\text{new}} = \frac{\sum_i w_i y_i}{\sum_i w_i} = \frac{\frac{120}{0.009305^2} + \frac{122}{0.17643^2} + \frac{130}{0.09756^2}}{\frac{1}{0.009305^2} + \frac{1}{0.17643^2} + \frac{1}{0.09756^2}} = 120.0954.$$

k-NN for Estimation and Prediction

	sepal-length	sepal-width	petal-length	petal-width
	????	2.3	3.3	1.2
1	4.9	2.4	3.3	1
2	5.1	2.5	3	1.1
3	5	2	3.5	1
4	5.5	2.4	3.7	1

Bayesian Classifiers

- ✓ A probabilistic framework for solving classification problems
- ✓ Conditional Probability:

$$P(C \mid A) = \frac{P(A,C)}{P(A)} \qquad P(A \mid C) = \frac{P(A,C)}{P(C)}$$

- ✓ Bayes theorem:

$$P(C \mid A) = \frac{P(A \mid C)P(C)}{P(A)}$$

Example of Bayes Theorem

- ✓ Given:
  - A doctor knows that meningitis causes stiff neck 50% of the time
  - Prior probability of any patient having meningitis is 1/50,000
  - Prior probability of any patient having stiff neck is 1/20

- ✓ If a patient has stiff neck, what's the probability he/she has meningitis?

$$P(M \mid S) = \frac{P(S \mid M)P(M)}{P(S)} = \frac{0.5 \times 1/50000}{1/20} = 0.0002$$

Bayesian Classifiers

- ✓ Consider each attribute and class label as random variables
- ✓ Given a record with attributes (A_1, A_2,…,A_n)
  - Goal is to predict class C
  - Specifically, we want to find the value of C that maximizes P(C| A_1, A_2,…,A_n)

$$P(C|A_1, A_2, A_3 \dots, A_n) = \frac{P(A_1, A_2, A_3 \dots, A_n|C)P(C)}{P(A_1, A_2, A_3 \dots, A_n)}$$

- ✓ Equivalent to choosing value of C that maximizes

$$P(A_1, A_2, \dots, A_n|C)\, P(C)$$

Bayesian Classifiers

- ✓ **X = (age <= 30 , income = medium, student = yes, credit_rating = fair)**
- ✓ **P(C|X) = ?**

$P(C_1)$: P(buys_computer = "yes") = 9/14 = **0.643**
$P(C_2)$: P(buys_computer = "no") = 5/14= **0.357**

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

- ✓ Compute $P(X|C_i)$ for each class

P(age = "<=30" | buys_computer = "yes") = 2/9 = **0.222**
P(age = "<= 30" | buys_computer = "no") = 3/5 = **0.6**
P(income = "medium" | buys_computer = "yes") = 4/9 = **0.444**
P(income = "medium" | buys_computer = "no") = 2/5 = **0.4**
P(student = "yes" | buys_computer = "yes) = 6/9 = **0.667**
P(student = "yes" | buys_computer = "no") = 1/5 = **0.2**
P(credit_rating = "fair" | buys_computer = "yes") = 6/9 = **0.667**
P(credit_rating = "fair" | buys_computer = "no") = 2/5 = **0.4**

$P(X|C_1)$: P(X|buys_computer = "yes") = 0.222 x 0.444 x 0.667 x 0.667 = **0.044**
$P(X|C_2)$: P(X|buys_computer = "no") = 0.6 x 0.4 x 0.2 x 0.4 = **0.019**

$P(X|C_i)*P(C_i)$: P(X|buys_computer = "yes") * P(buys_computer = "yes") = **0.028**
P(X|buys_computer = "no") * P(buys_computer = "no") = **0.007**

Therefore, X belongs to class ("buys_computer = yes")

Bayesian Classifiers

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

A: attributes

M: mammals (7)

N: non-mammals (13)

$$P(A \mid M) = \frac{6}{7} \times \frac{6}{7} \times \frac{2}{7} \times \frac{2}{7} = 0.06$$

$$P(A \mid N) = \frac{1}{13} \times \frac{10}{13} \times \frac{3}{13} \times \frac{4}{13} = 0.0042$$

$$P(A \mid M)P(M) = 0.06 \times \frac{7}{20} = 0.021$$

$$P(A \mid N)P(N) = 0.004 \times \frac{13}{20} = 0.0027$$

Give Birth	Can Fly	Live in Water	Have Legs	Class
yes	no	yes	no	?

P(A|M)P(M) > P(A|N)P(N)

=> Mammals