Classification

Part 3

Dr. Sanjay Ranka
Professor
Computer and Information Science and Engineering
University of Florida, Gainesville

Bayesian Classifiers

- A probabilistic framework to solve the classification problem
- Conditional Probability:

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

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

- Bayes Theorem: $P(C \mid A) = \frac{P(A \mid C)P(C)}{P(A)}$

Bayes Theorem: Example

- 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 set of attributes $(A_1, A_2, \ldots, A_n)$
  - Goal is to predict class C
  - Specifically, we want to find value of C that maximizes $P(C \mid A_1, A_2, \ldots, A_n)$
- Can we estimate $P(C \mid A_1, A_2, \ldots, A_n)$ directly from data?

Bayesian Classifiers

- Approach:
  - Compute the posterior probability P(C | A_1, A_2, … , A_n) for all values of C using the Bayes theorem

$$P(C \mid A_1 A_2 \ldots A_n) = \frac{P(A_1 A_2 \ldots A_n \mid C) P(C)}{P(A_1 A_2 \ldots A_n)}$$

  - Choose value of C that maximizes

$$P(C \mid A_1, A_2, \ldots, A_n)$$

  - Equivalent to choosing value of C that maximizes

$$P(A_1, A_2, \ldots, A_n \mid C)\, P(C)$$

- How to estimate P(A_1, A_2, … , A_n | C)?

Naïve Bayes Classifier

- Assume independence among attributes A_i when class is given:
  - $P(A_1, A_2, \ldots, A_n \mid C_j) = P(A_1 \mid C_j)\, P(A_2 \mid C_j) \ldots P(A_n \mid C_j)$
  - Can estimate $P(A_i \mid C_j)$ for all A_i and C_j
  - New point is classified to C_j if $P(C_j) \prod P(A_i \mid C_j)$ is maximal

How to Estimate Probabilities from Data?

Tid	Refund	Marital Status	Taxable Income	Evade
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	Singl e	90K	Yes

- Class: $P(C) = N_c/N$
  - e.g., $P(No) = 7/10$
- For discrete attributes:

$$P(A_i \mid C_k) = |A_{ik}| / N_c$$

  - where $|A_{ik}|$ is number of instances having attribute A_i and belongs to class C_k
  - Examples:

P(Married | No) = 4/7
P(Refund=Yes | Yes)=0

How to Estimate Probabilities from Data?

- For continuous attributes:
  - Discretize the range into bins
    - one ordinal attribute per bin
    - violates independence assumption
  - Two-way split: $(A < v)$ or $(A > v)$
    - choose only one of the two splits as new attribute
  - Assume attribute obeys certain probability distribution
    - Typically, normal distribution is assumed
    - Use data to estimate parameters of distribution (e.g., mean and standard deviation)
    - Once probability distribution is known, can use it to estimate the conditional probability $P(A_i \mid C)$

How to Estimate Probabilities from Data?

Tid	Refund	Marital Status	Taxable Income	Evade
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	Singl e	90K	Yes

- Normal distribution:

$$P(A_i \mid C_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{-\frac{(A_i - \mu_{ij})^2}{2\sigma_{ij}^2}}$$

  - One for each (A_i, C_j) pair
- For (Income, Class=No):
  - If Class=No
    - sample mean = 110K
    - sample variance = 2975

$$P(Income = 120K \mid No) = \frac{1}{\sqrt{2\pi}(54.54)} e^{-\frac{(120-110)^2}{2(2975)}} = 0.0072$$

Example of Naïve Bayes Classifier

Given a Test instance:

$$X = (\text{Refund} = \text{No}, \text{Married}, \text{Income} = 120\text{K})$$

naive Bayes Classifier:

P(Refund=Yes|No) = 3/7
P(Refund=No|No) = 4/7
P(Refund=Yes|Yes) = 0
P(Refund=No|Yes) = 1
P(Marital Status=Single|No) = 2/7
P(Marital Status=Divorced|No)=1/7
P(Marital Status=Married|No) = 4/7
P(Marital Status=Single|Yes) = 2/7
P(Marital Status=Divorced|Yes)=1/7
P(Marital Status=Married|Yes) = 0

For taxable income:
If class=No: sample mean=110
sample variance=2975
If class=Yes: sample mean=90
sample variance=25

- $P(X \mid \text{Class=No}) = P(\text{Refund=No} \mid \text{Class=No}) \times P(\text{Married} \mid \text{Class=No}) \times P(\text{Income=120K} \mid \text{Class=No})$
  $= 4/7 \times 4/7 \times 0.0072 = 0.0024$
- $P(X \mid \text{Class=Yes}) = P(\text{Refund=No} \mid \text{Class=Yes}) \times P(\text{Married} \mid \text{Class=Yes}) \times P(\text{Income=120K} \mid \text{Class=Yes})$
  $= 1 \times 0 \times 1.2 \times 10^{-9} = 0$

Since $P(X \mid \text{No})P(\text{No}) > P(X \mid \text{Yes})P(\text{Yes})$

Therefore $P(\text{No} \mid X) > P(\text{Yes} \mid X)$

=> Class = No

Naïve Bayes Classifier

- If one of the conditional probability is zero, then the entire expression becomes zero
  - Independence ⇔ multiplication of probabilities
- Laplace correction (also known as m-estimate):

$$\text{Original}: P(A_i \mid C) = \frac{N_{ic}}{N_c}$$

$$\text{Laplace}: P(A_i \mid C) = \frac{N_{ic} + mp}{N_c + m}$$

  - p is the prior probability as specified by the user
  - m is a parameter known as the equivalent sample size

Another Example of Naïve Bayes Classifier

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

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

A: attributes

M: mammals

N: non-mammals

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

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

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

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

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

=> Mammals

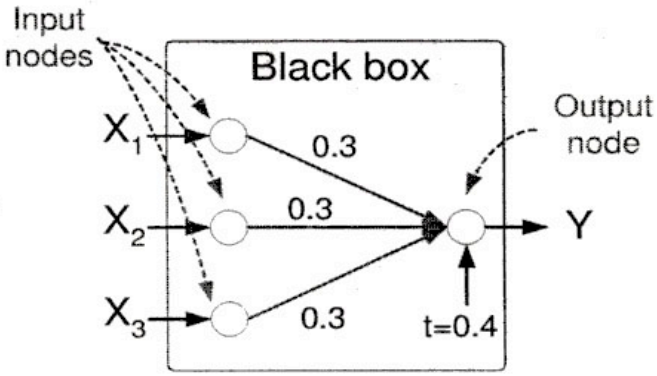
Bayesian Classifiers

- Robust to isolated noise points
- Handle missing values by ignoring the instance during probability estimate calculations
- Robust to irrelevant attributes
- Independence assumption may not hold for some attributes
  - Use other techniques such as Bayesian Belief Networks (BBN)

Artificial Neural Networks

- Classification model can be regarded as a black box that reads input values for X_1, X_2, and X_3, and sends out an output value $f(X_1, X_2, X_3)$ that is consistent with the true output value Y
- Here, $f(X_1, X_2, X_3) = 0.3X_1 + 0.3X_2 + 0.3X_3 - 0.4$

X_1	X_2	X_3	Y
1	0	0	0
1	0	1	1
1	1	0	1
1	1	1	1
0	0	1	0
0	1	0	0
0	1	1	1
0	0	0	0



- Such a black box, which represents its target function using a set of nodes and weighted links, is known as *Artificial Neural Network*

Artificial Neural Networks

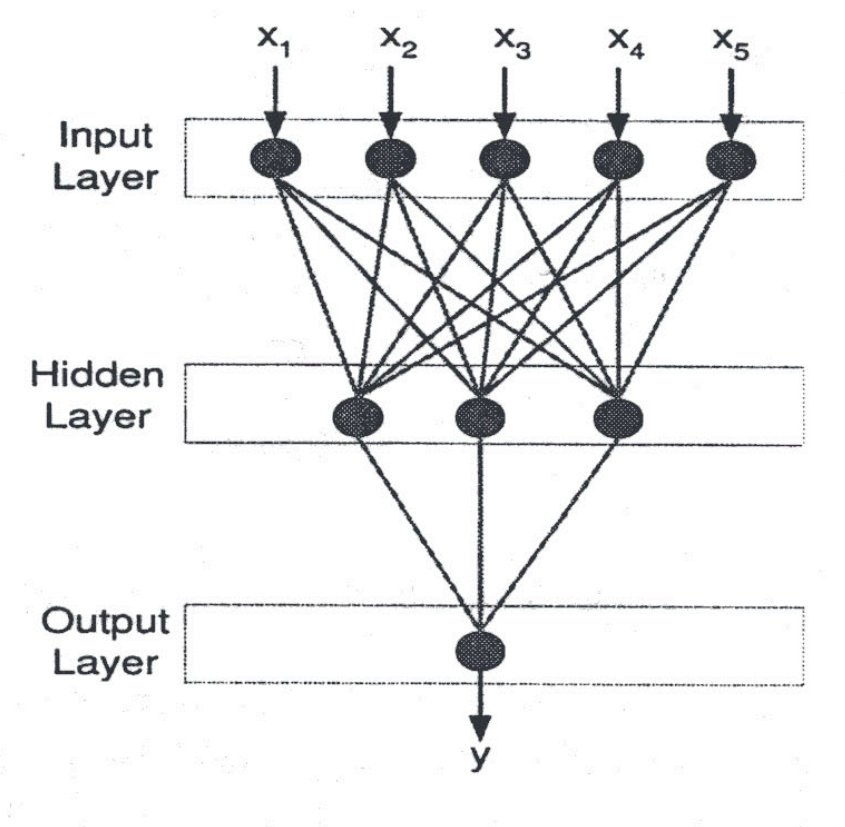
- The study of artificial neural networks was inspired by attempts to model the way human brain works
- Human brain primarily consists of nerve cells called *neurons*, linked together with other neurons via strands of fiber called *axons*
- Neurologists have discovered that the human brain learns by changing the strength of the connection of neurons upon repeated stimulation by the same impulse

Artificial Neural Networks

- Analogous to human brain structure, an ANN is composed of an inter connected assembly of nodes and directed links
- The nodes in an ANN are often called *neurons* or *units*
- Each link is associated with a real valued weight parameter to emulate the connection strength between neurons

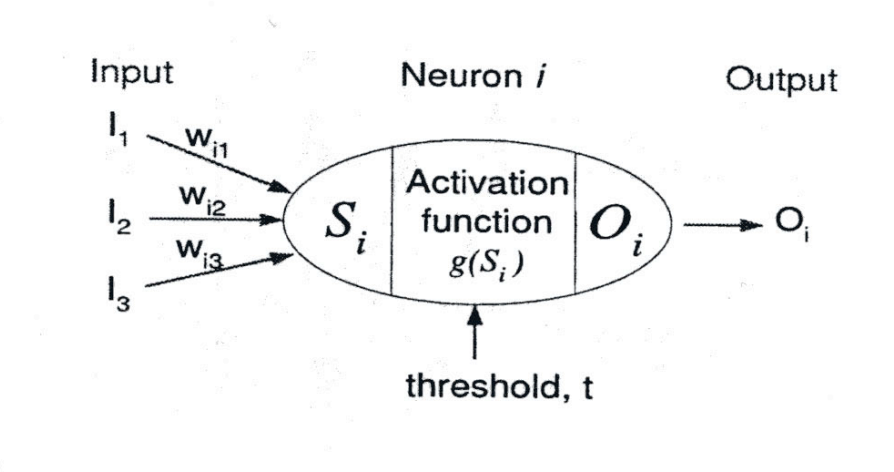
Artificial Neural Networks

- Neurons are organized in layers
- Input layer contains nodes that represent the input variables
- Output layer contains nodes that represent the target (output) variables
- Zero or more hidden layers may reside between the input and output layers

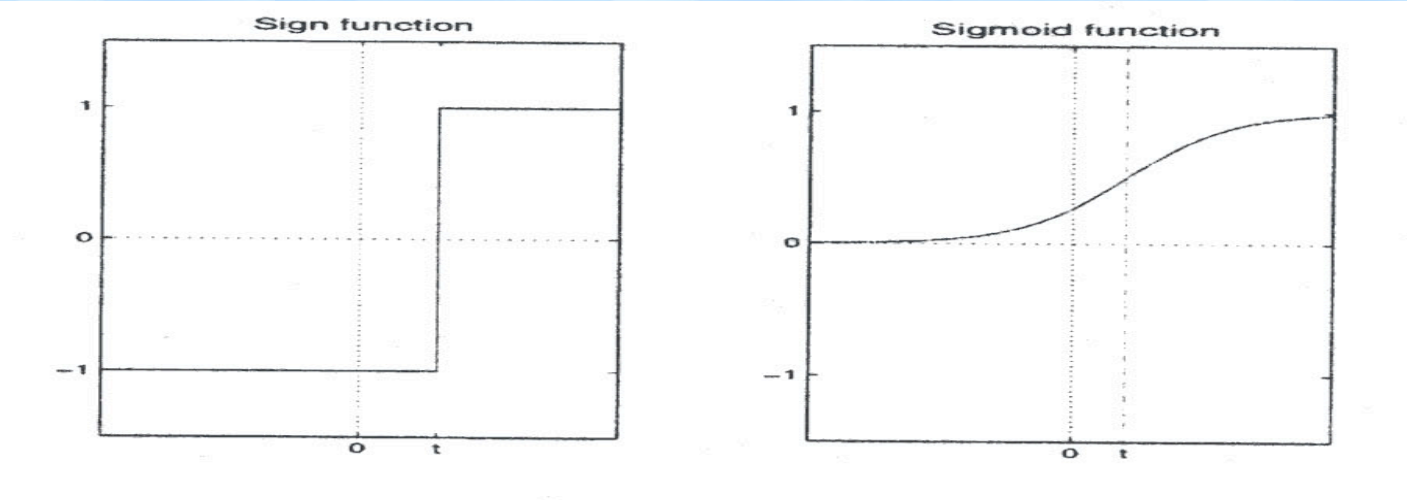


Structure of a Neuron

- Weighted average of the inputs values (*I*) is calculated (*S*)
- Output (*O*) is generated by applying activation function (*g*) and threshold (*t*) to this weighted average



Types of activation functions



- Typical activation functions used for many neural networks applications include the *sign* function and *sigmoid* function.
- Both the functions have been displaced to the right by *t* showing the threshold for the output to become 1

Steps involved in designing ANN

- Determine the number of nodes in input layer
- Determine the number of nodes in output layer
- Select appropriate network topology (number of hidden layers and hidden nodes, feed-forward or recurrent architecture)
- Initialize the weights and thresholds
- Remove training examples with missing values or replace them with their most likely values
- Train the neural network by adjusting the weights of the links until the outputs produced are consistent with class labels of training data (using algorithms like back propagation)

Characteristics of ANN

- Multi layered neural networks are universal approximators i.e. they can be used to approximate any target function
- Neural networks are robust to noise
- Training a neural network is computation intensive. However, classifying an unlabeled instance is very fast