CS 4104
APPLIED MACHINE LEARNING

Dr. Hashim Yasin

National University of Computer and Emerging Sciences,

Faisalabad, Pakistan.

PROBABILITIES

Probability

- A probability is the **real-valued function** defined on the sample space Ω that satisfy the following **properties:**
  - For any event $E \subseteq \Omega, 0 \leq P(E) \leq 1$
  - $P(\Omega) = 1$
  - For any set of disjoint events $E_1, E_2, \ldots E_k \in \Omega$

$$P\left(\bigcup_{i=1}^{k} E_i\right) = \sum_{i=1}^{k} P(E_i)$$

Probability

- A random experiment has uncertain outcome.
- All possible outcomes of a random experiment are the **sample space**.
- An **event** is the subset of these outcomes.
- Consider,
  - there are *n* elementary events (an atomic **event** or sample point) associated with a random experiment and *m* of *n* are favorable to an event *A*,
  - then the probability of occurrence of *A* is

$$P(A) = \frac{m}{n}$$

Probability

Mutually Exclusive Events:

- Two events are mutually exclusive, *if the occurrence of one excludes the occurrence of the other*.
- Example:
  - Tossing a coin (two events)
  - Tossing a dice cube (Six events)

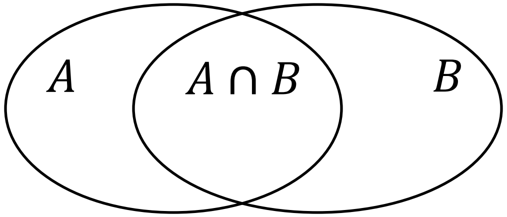
Independent Events:

- Two events are independent *if occurrences of one does not alter the occurrence of other*.
- Example:
  - Tossing both coin and dice cube together.

Joint Probability

- If $P(A)$ and $P(B)$ are the probabilities of two events, then

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$



- If A and B are mutually exclusive, then

$$P(A \cap B) = 0$$

- If A and B are independent events, then

$$P(A \cap B) = P(A, B) = P(A) \times P(B)$$

- Thus, for mutually exclusive events

$$P(A \cup B) = P(A) + P(B)$$

Conditional Probability

- If events are dependent, then their probability is expressed by *conditional probability*.

- The probability that A occurs given B is denoted by $P(A|B)$.

$$P(A|B) = \frac{P(A,B)}{P(B)}$$

Conditional Probability

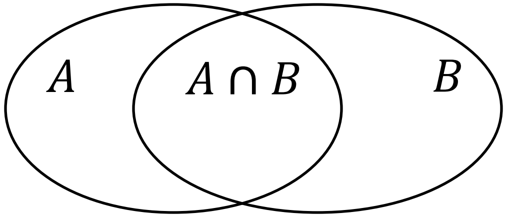
$$P(A|B) = \frac{P(A,B)}{P(B)}$$

- Suppose, A and B are two events associated with a random experiment.
- The probability of A under the condition that B has already occurred and $P(B) \neq 0$, is given by

$$P(A|B) = \frac{\text{Number of events in } B \text{ which are favourable to } A}{\text{Number of events in } B}$$

$$= \frac{\text{Number of events favourable to } A \cap B}{\text{Number of events favourable to } B}$$

$$= \frac{P(A \cap B)}{P(B)}$$



Conditional Probability … Example

X

	circle	square
red	0.20	0.02
blue	0.02	0.01

Y

	circle	square
red	0.05	0.30
blue	0.20	0.20

$$P(red) = 0.20 + 0.02 + 0.05 + 0.3 = 0.57$$

$$P(red \cap circle) = 0.20 + 0.05 = 0.25$$

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$P(\mathrm{X} \mid red \cap circle) = \frac{P(\mathrm{X} \cap red \cap circle)}{P(red \cap circle)} = \frac{0.20}{0.25} = 0.80$$

Conditional Probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$P(A \cap B) = P(A).P(B|A), \qquad if\ P(A) \neq 0$$

$$P(A \cap B) = P(B).P(A|B), \qquad if\ P(B) \neq 0$$

- For three events A, B and C

$$P(A \cap B \cap C) = P(A).P(B).P(C|A \cap B)$$

Conditional Probability

- if events are **mutually exclusive**

$$P(A|B) = 0$$

- if A and B are **independent**

$$P(A|B) = P(A)$$
$$P(B|A) = P(B)$$

- Moreover,

$$P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$$

$$P(A \cap B) = P(B \cap A)$$

Conditional Probability

- **Generalization of Conditional Probability:**

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B \cap A)}{P(B)}$$

$$P(A \cap B) = P(B|A) \times P(A) = P(A|B) \times P(B)$$

$$\boxed{P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}}$$

Bayes Theorem

BAYES THEOREM

Bayes Theorem

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

- Our target is to **determine the best hypothesis** (most probable hypothesis) from some hypothesis space H, given the observed training data D.
- Bayes theorem combines prior knowledge with observed data as,

$$\boxed{P(h|D) = \frac{P(D|h)P(h)}{P(D)}}$$

  - $P(h)$ = prior probability of hypothesis h
  - $P(D)$ = prior probability of training data D
  - $P(h|D)$ = probability of h given D
  - $P(D|h)$ = probability of D given h

Bayes Theorem

- In many learning scenarios, the learner is interested in finding the **most probable hypothesis** $h \in H$ given the observed data D ,or,
  - at least one of the maximally probable if there are several.
- Any such maximally probable hypothesis is called a ***maximum a posteriori* (MAP) hypothesis**.
- We can determine the MAP hypotheses by using Bayes theorem

Bayes Theorem

Find most probable hypothesis given training data

Maximum a posteriori hypothesis h_{MAP}:

$$\begin{aligned} h_{MAP} &= \arg\max_{h \in H} P(h|D) \\ &= \arg\max_{h \in H} \frac{P(D|h)P(h)}{P(D)} \\ &= \arg\max_{h \in H} P(D|h)P(h) \end{aligned}$$

Drop the term $P(D)$ because it is a constant independent of h

Example

Does patient have cancer or not?

A patient takes a lab test and the result comes back positive. The test returns a correct positive result in only 98% of the cases in which the disease is actually present, and a correct negative result in only 97% of the cases in which the disease is not present. Furthermore, 0.008 of the entire population have this cancer.

1. What is the probability that this patient has cancer?
2. What is the probability that he does not have cancer?
3. What is the diagnosis?

Example

Does patient have cancer or not?

A patient takes a lab test and the result comes back positive. The test returns a correct positive result in only 98% of the cases in which the disease is actually present, and a correct negative result in only 97% of the cases in which the disease is not present. Furthermore, 0.008 of the entire population have this cancer.

$$P(cancer) = .008, \qquad P(\neg cancer) = .992$$
$$P(\oplus|cancer) = .98, \qquad P(\ominus|cancer) = .02$$
$$P(\oplus|\neg cancer) = .03, \qquad P(\ominus|\neg cancer) = .97$$

Example

$$P(cancer) = .008, \qquad P(\neg cancer) = .992$$
$$P(\oplus|cancer) = .98, \qquad P(\ominus|cancer) = .02$$
$$P(\oplus|\neg cancer) = .03, \qquad P(\ominus|\neg cancer) = .97$$

The maximum a posteriori hypothesis can be found

$$\boxed{h_{MAP} = \arg\max_{h \in H} P(D|h)P(h)}$$

$$P(\oplus|cancer)P(cancer) = (.98).008 = .0078$$
$$P(\oplus|\neg cancer)P(\neg cancer) = (.03).992 = .0298$$

$$\boxed{\text{Thus, } h_{MAP} = \neg cancer.}$$

NAÏVE BAYES CLASSIFIER

Naïve Bayes Classifier

Assume target function $f : X \rightarrow V$, where each instance x described by attributes $\langle a_1, a_2 \ldots a_n \rangle$.

Most probable value of $f(x)$ is:

$$v_{MAP} = \underset{v_j \in V}{\operatorname{argmax}} P(v_j | a_1, a_2 \ldots a_n)$$

$$\begin{aligned} v_{MAP} &= \underset{v_j \in V}{\operatorname{argmax}} \frac{P(a_1, a_2 \ldots a_n | v_j) P(v_j)}{P(a_1, a_2 \ldots a_n)} \\ &= \underset{v_j \in V}{\operatorname{argmax}} P(a_1, a_2 \ldots a_n | v_j) P(v_j) \end{aligned}$$

Naïve Bayes Classifier

Assumption:

- The naive Bayes classifier is based on the assumption that the attribute values are *conditionally independent* given the target value.

$$P(a_1, a_2 \ldots a_n | v_j) = \prod_i P(a_i | v_j)$$

- Substituting this into

$$v_{MAP} = \underset{v_j \in V}{\operatorname{argmax}} P(a_1, a_2 \ldots a_n | v_j) P(v_j)$$

$$\boxed{v_{NB} = \underset{v_j \in V}{\operatorname{argmax}} P(v_j) \prod_i P(a_i | v_j)}$$

Naïve Bayes Classifier

Naive Bayes Algorithm

Naive_Bayes_Learn($examples$)

For each target value v_j

- $\hat{P}(v_j) \leftarrow$ estimate $P(v_j)$
- For each attribute value a_i of each attribute a
  $\hat{P}(a_i|v_j) \leftarrow$ estimate $P(a_i|v_j)$

Classify_New_Instance(x)

$$v_{NB} = \underset{v_j \in V}{\operatorname{argmax}} \hat{P}(v_j) \prod_{a_i \in x} \hat{P}(a_i|v_j)$$

Example

Days	Season	Fog	Rain	Class
Weekday	Spring	None	None	On Time
Weekday	Winter	None	Slight	On Time
Weekday	Winter	None	None	On Time
Weekday	Winter	High	Slight	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

Air-Traffic Data

Example

Days	Season	Fog	Rain	Class
Saturday	Spring	High	Heavy	Cancelled
Weekday	Summer	High	Slight	On Time
Weekday	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	Heavy	On Time

Air-Traffic Data

Days	Season	Fog	Rain	Class
Weekday	Spring	None	None	On Time
Weekday	Winter	None	Slight	On Time
Weekday	Winter	None	None	On Time
Weekday	Winter	High	Slight	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
Weekday	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	Heavy	On Time

Example

- In this database, there are **four attributes** with 20 tuples.

A = [Day, Season, Fog, Rain]

- The **categories of classes** are:

C = [On Time, Late, Very Late, Cancelled]

- Given this is the knowledge of data and classes, the target is to find most likely classification for any other unseen instance, for example:

Week Day	**Winter**	**High**	**None**	**???**

- Classification technique eventually to map this tuple into an accurate class.

Example

$$v_{NB} = \underset{v_j \in V}{\operatorname{argmax}}\, P(v_j) \prod_i P(a_i|v_j)$$

		Class			
Attribute		On Time(14)	Late(2)	Very Late(3)	Cancelled(1)
Days	**Weekday**	**9/14 = 0.64**	**2/2 = 1**	**3/3 = 1**	**0/1 = 0**
	Saturday	2/14 = 0.14	0/2 = 0	0/3 = 0	1/1 = 1
	Sunday	1/14 = 0.07	0/2 = 0	0/3 = 0	0/1 = 0
	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
	Summer	6/14 = 0.43	0/2 = 0	0/3 = 0	0/1 = 0
	Autumn	2/14 = 0.14	0/2 = 0	1/3= 0.33	0/1 = 0
	Winter	**2/14 = 0.14**	**2/2 = 1**	**2/3 = 0.67**	**0/1 = 0**

Example

$$v_{NB} = \underset{v_j \in V}{\text{argmax}}\, P(v_j) \prod_i P(a_i|v_j)$$

		Class			
Attribute		**On Time(14)**	**Late(2)**	**Very Late(3)**	**Cancelled(1)**
Fog	None	5/14 = 0.36	0/2 = 0	0/3 = 0	0/1 = 0
	High	**4/14 = 0.29**	**1/2 = 0.5**	**1/3 = 0.33**	**1/1 = 1**
	Normal	5/14 = 0.36	1/2 = 0.5	2/3 = 0.67	0/1 = 0
Rain	**None**	**6/14 = 0.43**	**1/2 = 0.5**	**1/3 = 0.33**	**0/1 = 0**
	Slight	6/14 = 0.43	1/2 = 0.5	0/3 = 0	0/1 = 0
	Heavy	2/14 = 0.14	0/2 = 0	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**

Example

$$v_{NB} = \underset{v_j \in V}{\operatorname{argmax}} P(v_j) \prod_i P(a_i|v_j)$$

Instance:

Week Day	Winter	High	None	???

Case 1: Class = On Time:

$0.70 \times 0.64 \times 0.14 \times 0.29 \times 0.43 = 0.0078$

Case 2: Class = Late:

$0.10 \times 1.0 \times 1.0 \times 0.50 \times 0.50 =$ **0.025**

Case 3: Class = Very Late:

$0.15 \times 1.0 \times 0.67 \times 0.33 \times 0.33 = 0.0109$

Case 4: Class = Cancelled:

$0.05 \times 0.0 \times 0.0 \times 1.0 \times 0.0 = 0.0$

Case 2 is the strongest; Hence correct classification is **Late**

Naïve Bayes Classifier

- Highly practical Bayesian learning method
  - In some domains its performance can be comparable to that of neural network and decision tree learning
- **When to use,**
  - Moderate or large training dataset is available
  - Attributes that describe instances are conditionally independent given classification
- **Application**
  - Diagnosis systems (expert systems)
  - Classifying text documents

Reading Material

- **Artificial Intelligence,** A Modern Approach

  Stuart J. Russell and Peter Norvig
  - **Chapter 13.**
- **Machine Learning**

  Tom M. Mitchell
  - **Chapter 6.**