Naïve Bayes Classifier

한양대학교 ERICA
소프트웨어융합대학
COLLEGE OF COMPUTING
인공지능학과
Department of
Artificial Intelligence

Modified from Prof. Debasis Samanta's slides

Review of Probability Theory

- Unconditional independence

$$P(V_1, V_2, \ldots V_k) = \prod_{i=1}^{k} P(V_i \mid V_{i-1}, \ldots, V_1) = \prod_{i=1}^{k} P(V_i)$$

  - e.g., *P* (A=a, B=b)
    - P(a,b)=P(ab)=P(a)P(b)

- Conditional independence

$$P(V_1, V_2, \ldots, V_k \mid V) = \prod_{i=1}^{k} P(V_i \mid V_{i-1}, \ldots V_1, V) = \prod_{i=1}^{k} P(V_i \mid V)$$

  - e.g., P(A=a,B=b|C=c)
    - P(ab|c)=P(a|c)P(b|c)

ZeroR

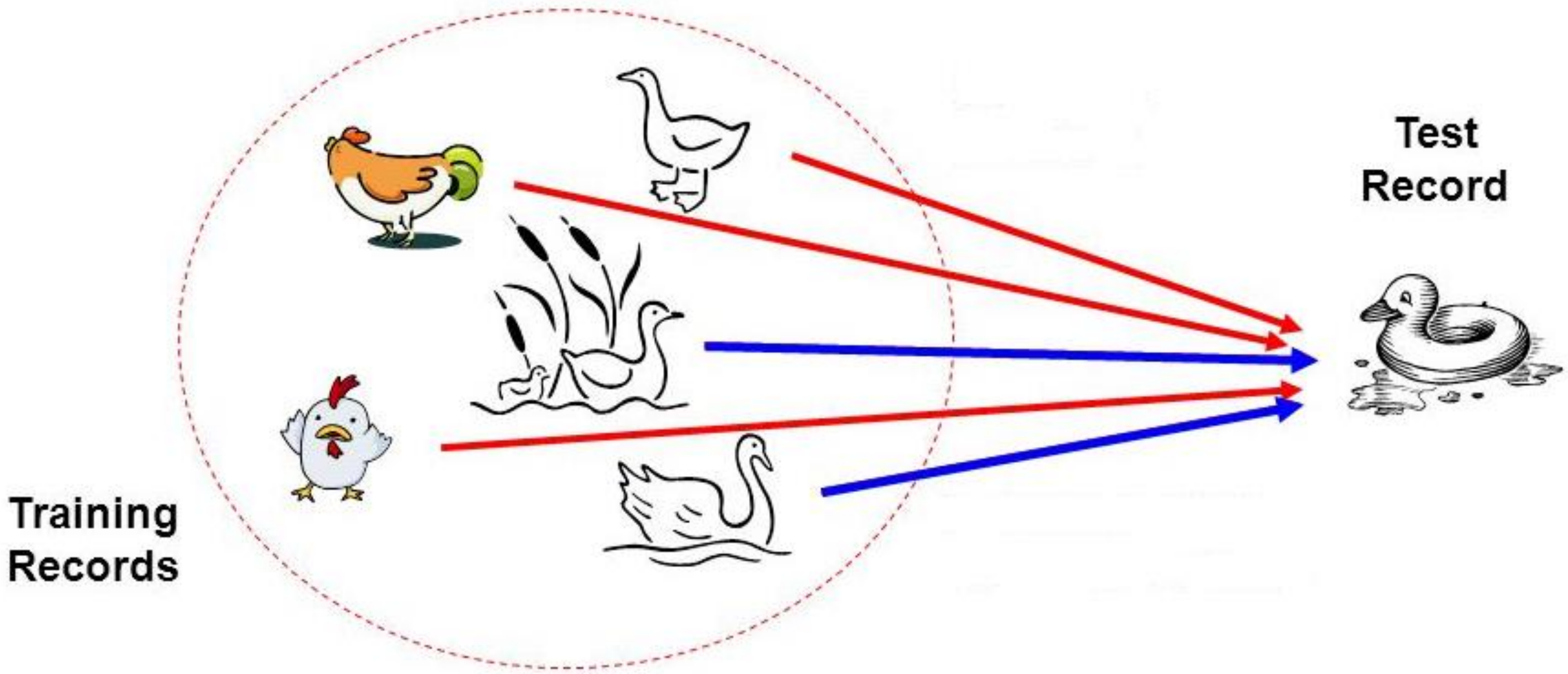
- Just answer the majority class of train data all the time
  - Play=No: 5/14 tuples
  - Play=Yes: 9/14 tuples
- Answer Play=Yes for every test data

Outlook	Temp	Humidity	Wind	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

Ian H. Witten's slide

Bayesian Classifier

- Principle
  - If it walks like a duck, quacks like a duck, then it is probably a duck



OneR: One attribute does all the work

Outlook	Temp	Humidity	Wind	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

Attribute	Rules	Errors	Total errors
Outlook	Sunny → No	2/5	4/14
	Overcast → Yes	0/4	
	Rainy → Yes	2/5	
Temp	Hot → No*	2/4	5/14
	Mild → Yes	2/6	
	Cool → Yes	1/4	
Humidity	High → No	3/7	4/14
	Normal → Yes	1/7	
Wind	False → Yes	2/8	5/14
	True → No*	3/6	

* indicates a tie

Ian H. Witten's slide

“Naïve Bayes” method

- Opposite strategy: use all the attributes
  - OneR: One attribute does all the work
- Two assumptions: Attributes are
  - equally important a priori
  - statistically independent (given the class value)
    - i.e., knowing the value of one attribute says nothing about the value of another (if the class is known)
- Independence assumption is never correct!
- But … often works well in practice

Ian H. Witten’s slide

Bayesian Classifiers

- A statistical classifier
  - Performs probabilistic prediction, i.e., predicts class membership probabilities
    - Output $p(C_1), p(C_2) \ldots p(C_k)$

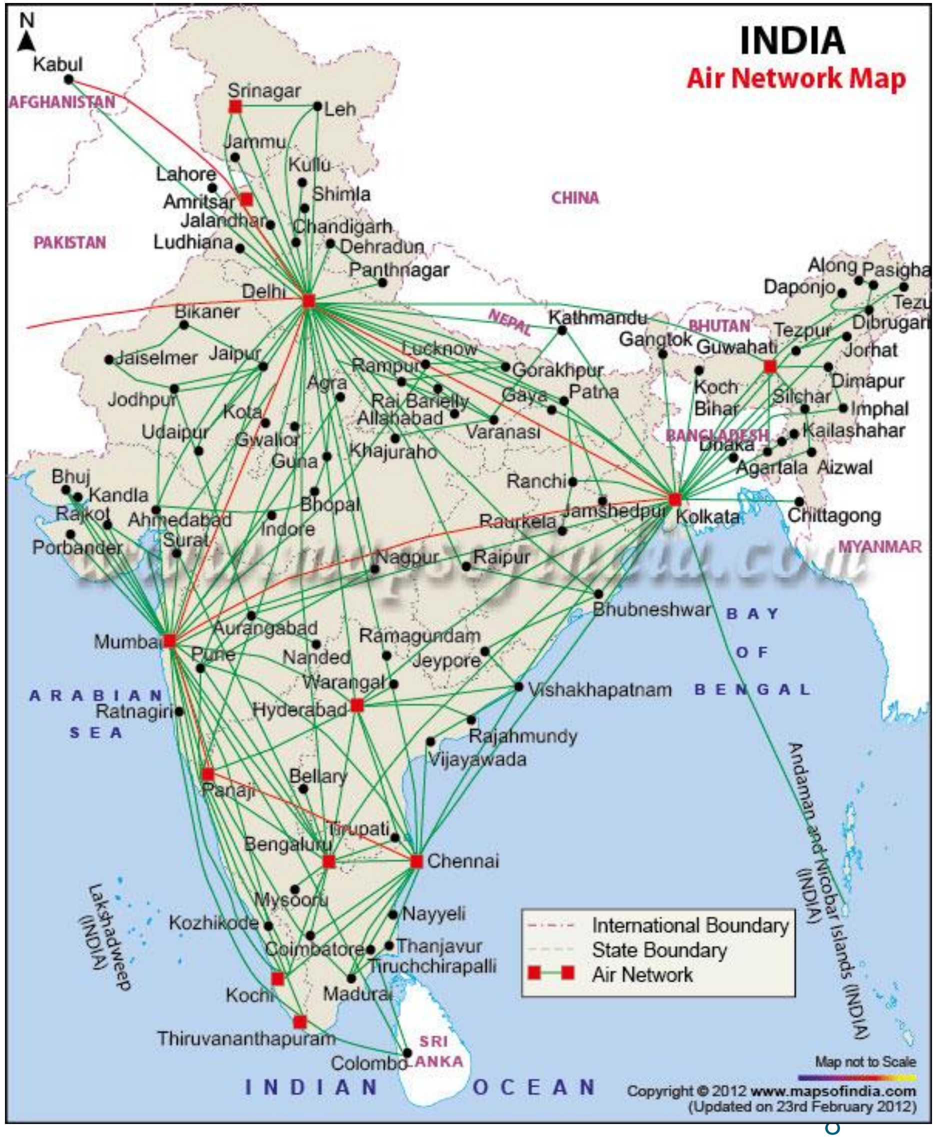
- Foundation
  - Based on Bayes' Theorem.

- Assumptions
  1. The classes are mutually exclusive and exhaustive.
  2. The attributes are independent given the class.

- Called "Naïve" classifier because of these assumptions.
  - Empirically proven to be useful.
  - Scales very well.

Example: Bayesian Classification

- **Example 8.2:** Air Traffic Data
  - Let us consider a set observation recorded in a database
    - Regarding the arrival of airplanes in the routes from any airport to New Delhi under certain conditions.



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
Holiday	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

Cond. to next slide…

Air-Traffic Data

Cond. from previous slide…

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

- In this database, there are four attributes

$$A = [\text{ Day, Season, Fog, Rain}]$$

  with 20 tuples.
- The categories of classes are:

$$C = [\text{On Time, Late, Very Late, Cancelled}]$$

- Given this is the knowledge of data and classes, we are 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.

Bayesian Classifier

- In many applications, the relationship between the attributes set and the class va riable is non-deterministic.
  - In other words, a test cannot be classified to a class label with certainty.
  - In such a situation, the classification can be achieved probabilistically.

- The Bayesian classifier is an approach for modelling probabilistic relationships be tween the attribute set and the class variable.

- More precisely, Bayesian classifier use Bayes' Theorem of Probability for classificat ion.

- Before going to discuss the Bayesian classifier, we should have a quick look at th e Bayes' Theorem.

Bayes' Theorem

What you know?	What you want to know?
$P(E\|F)$	$P(F\|E)$
$P(Test\ result\|Disease)$	$P(Disease\|Test\ result)$
$P(Power\|Fault)$	$P(Fault\|Power)$
$P(Weather\|Delay)$	$P(Delay\|Weather)$

Bayes Theorem

Want $P(F|E)$, Know $P(E|F)$

- For any events E and F where P(E)>0 and P(F)>0

$$\underbrace{P(F|E)}_{\text{Posterior}} = \frac{\overbrace{P(E|F)}^{\text{Likelihood}}\overbrace{P(F)}^{\text{Prior}}}{P(E)}$$

- Proof)

$$P(F|E) = \frac{P(EF)}{P(E)}$$ Conditional probability

$$= \frac{P(E|F)P(F)}{P(E)}$$ Chain rule

Naïve Bayesian Classifier

$$\boldsymbol{X} = (x_1, x_2, \dots, x_k) \xrightarrow{\text{Naïve Bayesian Classifier}} y \in \{C_1, C_2, \dots, C_m\}$$

- Classification is to derive the **maximum posteriori**, i.e., the maximal $P(C_i|\boldsymbol{X})$
- This can be derived from **Bayes' theorem**

$$P(C_i|\mathbf{X}) = \frac{P(\mathbf{X}|C_i)P(C_i)}{P(\mathbf{X})}$$

- Since P(X) is constant for all classes, only needs to be maximized

$$P(C_i|\mathbf{X}) \propto P(\mathbf{X}|C_i)P(C_i)$$

- Due to the independent assumption,

$$P(C_i|\mathbf{X}) \propto P(\mathbf{X}|C_i)P(C_i) = \left(\prod_{j=1}^{k} P\left(x_j|C_i\right)\right)P(C_i)$$

Naïve Bayesian Classifier

- Given $\boldsymbol{X}$
- Output C_i which has the maximum $P(\boldsymbol{X}|C_i)P(C_i) = \left(\prod_{j=1}^{k} P\left(x_j|C_i\right)\right)P(C_i)$
- Example)

Week Day	Winter	High	Heavy	???

$$P([WeekDay, Winter, High, Heavy]|On\ time)$$

$$= P(WeekDay|On\ time)P(Winter|On\ time)P(High|On\ time)P(On\ time)$$

Naïve Bayesian Classifier

$$P(\boldsymbol{X}|C_i)P(C_i) = \left(\prod_{j=1}^{k} P(x_j|C_i)\right) P(C_i)$$

- **Example:** With reference to the Air Traffic Dataset mentioned earlier, let us tabulate all the posterior and prior probabilities as shown below.

		Class			
Attribute		On Time	Late	Very Late	Cancelled
Day	Weekday	9/14 = 0.64	½ = 0.5	3/3 = 1	0/1 = 0
	Saturday	2/14 = 0.14	½ = 0.5	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	0/1 = 0
	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

Naïve Bayesian Classifier

$$P(\boldsymbol{X}|C_i)P(C_i) = \left(\prod_{j=1}^{k} P(x_j|C_i)\right)P(C_i)$$

	Attribute	Class			
		On Time	Late	Very Late	Cancelled
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	5/14 = 0.36	1/2 = 0.5	1/3 = 0.33	0/1 = 0
	Slight	8/14 = 0.57	0/2 = 0	0/3 = 0	0/1 = 0
	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 Bayesian Classifier

$$P(\boldsymbol{X}|C_i)P(C_i) = \left(\prod_{j=1}^{k} P(x_j|C_i)\right)P(C_i)$$

Instance:

Week Day	**Winter**	**High**	**Heavy**	**???**

Case1: Class = On Time : 0.70 × 0.64 × 0.14 × 0.29 × 0.07 = 0.0013

Case2: Class = Late : 0.10 × 0.50 × 1.0 × 0.50 × 0.50 = 0.0125

Case3: Class = Very Late : 0.15 × 1.0 × 0.67 × 0.33 × 0.67 = 0.0222

Case4: Class = Cancelled : 0.05 × 0.0 × 0.0 × 1.0 × 1.0 = 0.0000

Case3 is the strongest; Hence correct classification is **Very Late**

Example: Play Tennis

Outlook	Temp	Humidity	Wind	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

Ian H. Witten's slide

Example: Play tennis

Outlook	Temp.	Humidity	Wind	Play
Sunny	Cool	High	True	?

Evidence E

$$\Pr[yes \mid E] = \Pr[Outlook = Sunny \mid yes]$$
$$\times \Pr[Temperature = Cool \mid yes]$$
$$\times \Pr[Humidity = High \mid yes]$$
$$\times \Pr[Windy = True \mid yes]$$
$$\times \frac{\Pr[yes]}{\Pr[E]}$$

Probability of class "yes"

$$= \frac{\frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{9}{14}}{\Pr[E]}$$

Outlook			Temperature			Humidity			Wind			Play	
	Yes	*No*		*Yes*	*No*		*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

Ian H. Witten's slide

Example: Play tennis

Outlook			Temperature			Humidity			Wind			Play	
	Yes	*No*		*Yes*	*No*		*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

A new day:

Outlook	Temp.	Humidity	Wind	Play
Sunny	Cool	High	True	?

$$\Pr[H \mid E] = \frac{\Pr[E_1 \mid H]\Pr[E_2 \mid H] \ldots \Pr[E_n \mid H]\Pr[H]}{\Pr[E]}$$

Likelihood of the two classes

For “yes” = 2/9 × 3/9 × 3/9 × 3/9 × 9/14 = 0.0053

For “no” = 3/5 × 1/ × 4/5 × 3/5 × 5/14 = 0.0206

Conversion into a probability by normalization:

P(“yes”) = 0.0053 / (0.0053 + 0.0206) = 0.205

P(“no”) = 0.0206 / (0.0053 + 0.0206) = 0.795

Bayesian Classification: Why?

- A statistical classifier: performs probabilistic prediction, i.e., predicts class membership probabilities
- Foundation: Based on Bayes' Theorem.
- Performance: A simple Bayesian classifier, naïve Bayesian classifier, has comparable performance with decision tree and selected neural network classifiers
- Incremental: Each training example can incrementally increase/decrease the probability that a hypothesis is correct — prior knowledge can be combined with observed data

Naïve Bayesian Classifier

Pros and Cons

- The Naïve Bayes' approach is a very popular one, which often works well.
- However, it has a number of potential problems
  - It relies on all attributes being categorical.
  - If the data is less, then it estimates poorly.
  - …

Naïve Bayesian Classifiers :Continuous Attributes

- In real life situation, all attributes are not necessarily be categorical, In fact, there is a mix of both categorical and continuous attributes.
- In the following, we discuss the schemes to deal with continuous attributes in Bay esian classifier.

1. Discretize each continuous attribute and then replace the continuous values with its corresponding discrete intervals.

$$24.3℃ \rightarrow [20℃, 25℃)$$

2. Assume a certain form of probability distribution for the continuous variable

   Gaussian distribution is widely used

$$P(x : \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where, μ and σ^2 denote mean and variance, respectively.

Naïve Bayesian Classifiers :Continuous Attributes

$$P(\boldsymbol{X}|C_i)P(C_i) = \left(\prod_{j=1}^{k} P(x_j|C_i)\right)P(C_i)$$

- For each class C_i and attribute j,
  - $P(x_j|\mathrm{C}_i) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} e^{-\frac{(x_j-\mu_{ij})^2}{2\sigma_{ij}{}^2}}$
    - μ_{ij}: Sample mean
    - σ_{ij}^2: Sample variance

Naïve Bayesian Classifiers :Additive smoothing

Naïve Bayes

$$P(\boldsymbol{X}|C_i)P(C_i) = \left(\prod_{j=1}^{k} P(x_j|C_i)\right)P(C_i)$$

Approach to overcome the limitations in Naïve Bayesian Classification

$$P(x_j|C_i) = \frac{n_{c_i}}{n}$$

n = total number of instances from class C_i

n_{c_i} = number of training examples from class C_i that take the value $x_j = a_j$

- If the training data size is too small..
  - $P(x_j|C_i) \rightarrow 0$ for some ij
- Additive smoothing

$$P(x_j|C_i) = \frac{n_{c_i} + \alpha}{n + m_j\alpha}$$

$m_j: \# \ of \ possible \ values \ of \ attribute \ i$