SEKOLAH TINGGI MANAJEMEN INFORMATIKA DAN TEKNIK KOMPUTER

STMIK STIKOM Bali

STIKOM BALI

DATA MINING
Statistical Modelling

Made Satria Wibawa

Campus of Technology and Creativity

Background: methods to create classifiers

- There are three methods to establish a classifier

  a) Model a classification rule directly

  Examples: k-NN, decision trees, perceptron, SVM

  b) Model the probability of class memberships given input data

  Example: perceptron with the cross-entropy cost

  c) Make a probabilistic model of data within each class

  Examples: ***naive Bayes***, model based classifiers

a) and *b*) are examples of discriminative classification

c) is an example of generative classification

b) and *c*) are both examples of probabilistic classification

Statistical Modelling

Merupakan teknik sederhana yang menggunakan semua atribut untuk membuat kontribusi dalam pengambilan keputusan.

Dalam statistical modelling, atribut dianggap sama pentingnya (*equally important)* dan tidak terikat terhadap atribut lainnya *(independent)* dalam satu kelas.

Namun, hal tersebut tidak realistis dalam keadaan sebenarnya. Di kehidupan nyata atribut tidak sama pentingnya ataupun independen.

Walaupun begitu, skema statistical yang sederhana ini memberikan hasil yang baik dalam aplikasinya.

Bayesian Theorem

Menggambarkan probabilitas dari sebuah kejadian, berdasarkan kondisi yang memiliki hubungan dengan kejadian tersebut.

Sebagai contoh, jika kanker memiliki hubungan dengan usia. Maka dengan menggunakan teorema Bayesian, usia seseorang dapat digunakan untuk menghitung probabilitas seseorang mengidap penyakit kanker.

$$\Pr[H|E] = \frac{\Pr[E|H]\Pr[H]}{\Pr[E]}$$

Naïve Bayes

Naïve Bayes Algorithm (for discrete input attributes) has two phases

– **1. Learning Phase**: Given a training set **S**,

For each target value of c_i $(c_i = c_1, \cdots, c_L)$

$\hat{P}(C = c_i) \leftarrow$ estimate $P(C = c_i)$ with examples in $\mathbf{S}$;

For every attribute value x_{jk} of each attribute X_j $(j = 1, \cdots, n;\ k = 1, \cdots, N_j)$

$\hat{P}(X_j = x_{jk} \mid C = c_i) \leftarrow$ estimate $P(X_j = x_{jk} \mid C = c_i)$ with examples in $\mathbf{S}$;

Learning is easy, just create probability tables.

Output: conditional probability tables; for $X_j, N_j \times L$ elements

– **2. Test Phase**: Given an unknown instance $\mathbf{X}' = (a'_1, \cdots, a'_n)$

Look up tables to assign the label c^* to $\mathbf{X}'$ if

$$[\hat{P}(a'_1 \mid c^*) \cdots \hat{P}(a'_n \mid c^*)]\hat{P}(c^*) > [\hat{P}(a'_1 \mid c) \cdots \hat{P}(a'_n \mid c)]\hat{P}(c), \quad c \neq c^*, c = c_1, \cdots, c_L$$

Classification is easy, just multiply probabilities

Naïve Bayes : MAP

- MAP classification rule
  - **MAP**: **M**aximum **A P**osterior
  - Assign x to c^* if

$$P(C=c^* \mid \mathbf{X}=\mathbf{x}) > P(C=c \mid \mathbf{X}=\mathbf{x}) \quad c \neq c^*,\ c = c_1, \cdots, c_L$$

- **Method of** Generative classification with the MAP rule
  1. Apply Bayesian rule to convert them into posterior probabilities

$$P(C=c_i \mid \mathbf{X}=\mathbf{x}) = \frac{P(\mathbf{X}=\mathbf{x} \mid C=c_i)P(C=c_i)}{P(\mathbf{X}=\mathbf{x})}$$
$$\propto P(\mathbf{X}=\mathbf{x} \mid C=c_i)P(C=c_i)$$
$$\text{for } i = 1, 2, \cdots, L$$

  2. Then apply the MAP rule

Contoh Naïve Bayes

Outlook	Temperature	Humidity	Windy	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

Contoh Naïve Bayes

Outlook	Temperature	Humidity	Windy	Play
Sunny	Cool	High	True	?

$$to\ Play\ or\ Not\ to\ play\ ?$$

Contoh Naïve Bayes

Outlook			Temperature			Humidity			Windy			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								

$$\text{Likehood yes} = \frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{9}{14} = 0.00508$$

$$\text{Likehood no} = \frac{3}{5} \times \frac{1}{5} \times \frac{4}{5} \times \frac{3}{5} \times \frac{5}{14} = 0.02056$$

Contoh Naïve Bayes

$$\Pr[H|E] = \frac{\Pr[E|H]\Pr[H]}{\Pr[E]}$$

$$\text{Probability of } yes = \frac{0.00508}{0.00508 + 0.02056} \times 100\% = 19.81\%$$

$$\text{Probability of } no = \frac{0.02056}{0.00508 + 0.02056} \times 100\% = 80.18\%$$

Naïve Bayes for Numeric Value

Outlook			Temperature		Humidity		Windy			Play	
	yes	*no*	*yes*	*no*	*yes*	*no*		*yes*	*no*	*yes*	*no*
sunny	2	3	83	85	86	85	false	6	2	9	5
overcast	4	0	70	80	96	90	true	3	3		
rainy	3	2	68	65	80	70					
			64	72	65	95					
			69	71	70	91					
			75		80						
			75		70						
			72		90						
			81		75						

Naïve Bayes for Numeric Value

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

$$\sigma = \sqrt{\frac{\sum (x-\mu)^2}{N}}$$

$$x = \frac{\sum x}{N}$$

Naïve Bayes for Numeric Value

Outlook			Temperature			Humidity			Windy			Play	
	yes	*no*		*yes*	*no*		*yes*	*no*		*yes*	*no*	*yes*	*no*
sunny	2	3		83	85		86	85	false	6	2	9	5
overcast	4	0		70	80		96	90	true	3	3		
rainy	3	2		68	65		80	70					
				64	72		65	95					
				69	71		70	91					
				75			80						
				75			70						
				72			90						
				81			75						
sunny	2/9	3/5	*mean*	73	74.6	*mean*	79.1	86.2	false	6/9	2/5	9/14	5/14
overcast	4/9	0/5	*std*	6.2	7.9	*std*	10.2	9.7	true	3/9	3/5		
rainy	3/9	2/5											

Naïve Bayes for Numeric Value

$$f(temperature = 66|yes) = \frac{1}{\sqrt{2\pi} \times \sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Issues in Naïve Bayes

1. Violation of Independence Assumption

2. Zero conditional probability Problem

Issues in Naïve Bayes

First Issue

1. Violation of Independence Assumption

   Events are correlated

   - For many real world tasks, $P(X_1, \cdots, X_n | C) \neq P(X_1 | C) \cdots P(X_n | C)$
   - Nevertheless, naïve Bayes works surprisingly well anyway!

THANK YOU