BAYES CLASSIFIER

(www.aplysit.com | www.ivan.siregar.biz)

APLYSIT – IT SOLUTION CENTER

Jl. Ir. H. Djuanda 109
Bandung

Ivan Michael Siregar

ivan.siregar@gmail.com

Data Mining

2010

Bayesian Method

- Our focus this lecture
- Learning and classification methods based on probability theory.
- Bayes theorem plays a critical role in probabilistic learning and classification.
- Uses *prior* probability of each category given no information about an item.
- Categorization produces a *posterior* probability distribution over the possible categories given a description of an item.

Bayes Theorem

- P(A) : probability of A
- P(A|B) : probability of A given B
- P(A ∩ B) : probability of A and B together

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

- P(A ∩ B) = P(A|B) x P(B)

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

We can predict P(A|B) if P(B|A), P(A), and P(B) are given.

Guys, just go to example on page 13 for quick understanding!!!

Basic Probability Formulas

- Product rule

$$P(A \wedge B) = P(A \mid B)P(B) = P(B \mid A)P(A)$$

- Sum rule

$$P(A \vee B) = P(A) + P(B) - P(A \wedge B)$$

- Bayes theorem

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$

- Theorem of total probability, if event *Ai* is mutually exclusive and probability sum to 1

$$P(B) = \sum_{i=1}^{n} P(B \mid A_i)P(A_i)$$

Bayes Theorem

- Given a hypothesis *h* and data *D* which bears on the hypothesis:

- *P(h)*: independent probability of *h*: *prior probability*
- *P(D)*: independent probability of *D*
- *P(D|h)*: conditional probability of *D* given h: *likelihood*
- *P(h|D)*: conditional probability of *h* given *D*: *posterior probability*

Does Patient Have Cancer or Not

- A patient takes a lab test and the result comes back positive. It is known that the test returns a correct positive result in only 99% of the cases and a correct negative result in only 95% of the cases. Furthermore, only 0.03 of the entire population has this disease.

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?

Maximum A Posterior

- Based on Bayes Theorem, we can compute the *Maximum A Posterior* (MAP) hypothesis for the data
- We are interested in the best hypothesis for some space H given observed training data D.

$$h_{MAP} \equiv \underset{h \in H}{\operatorname{argmax}} \; P(h \mid D)$$

$$= \underset{h \in H}{\operatorname{argmax}} \; \frac{P(D \mid h) P(h)}{P(D)}$$

$$= \underset{h \in H}{\operatorname{argmax}} \; P(D \mid h) P(h)$$

***H*: set of all hypothesis.**

Note that we can drop *P(D)* as the probability of the data is constant (and independent of the hypothesis).

Maximum Likehood

- Now assume that all hypotheses are equally probable a priori, i.e., *P(hi) = P(hj)* for all *hi, hj* belong to *H*.
- This is called assuming a *uniform prior*. It simplifies computing the posterior:

$$h_{ML} = \arg\max_{h \in H} P(D \mid h)$$

- This hypothesis is called the *maximum likelihood hypothesis.*

Desirable Properties of Bayes Classifier

- *Incrementality:* with each training example, the prior and the likelihood can be updated dynamically: flexible and robust to errors.
- *Combines prior knowledge and observed data:* prior probability of a hypothesis multiplied with probability of the hypothesis given the training data
- *Probabilistic hypothesis:* outputs not only a classification, but a probability distribution over all classes

Bayes Classifier

Assumption: training set consists of instances of different classes described *cj* as conjunctions of attributes values

Task: Classify a new instance *d* based on a tuple of attribute values into one of the classes *cj* $\in$ *C*

Key idea: assign the most probable class using Bayes Theorem.

$$c_{MAP} = \underset{c_j \in C}{\operatorname{argmax}} P(c_j \mid x_1, x_2, \ldots, x_n)$$

$$= \underset{c_j \in C}{\operatorname{argmax}} \frac{P(x_1, x_2, \ldots, x_n \mid c_j) P(c_j)}{P(x_1, x_2, \ldots, x_n)}$$

$$= \underset{c_j \in C}{\operatorname{argmax}} P(x_1, x_2, \ldots, x_n \mid c_j) P(c_j)$$

Parameter Estimation

- $P(c_j)$
  - Can be estimated from the frequency of classes in the training examples.
- $P(x_1,x_2,\ldots,x_n|c_j)$
  - $O(|X|^n \bullet |C|)$ parameters
  - Could only be estimated if a very, very large number of training examples was available.
- Independence Assumption: attribute values are conditionally independent given the target value: ***naïve*** *Bayes*.

$$P(x_1,x_2,\ldots,x_n \mid c_j) = \prod_i P(x_i \mid c_j)$$

$$c_{NB} = \operatorname*{arg\,max}_{c_j \in C} P(c_j)\prod_i P(x_i \mid c_j)$$

Properties

- Estimating $P(x_i \mid c_j)$ instead of $P(x_1, x_2, \ldots, x_n \mid c_j)$greatly reduces the number of parameters (and the data sparseness).
- The learning step in Naïve Bayes consists of estimating$P(x_i \mid c_j)$ and $P(c_j)$ based on the frequencies in the training data
- An unseen instance is classified by computing the class that maximizes the posterior
- When conditioned independence is satisfied, Naïve Bayes corresponds to MAP classification.

Example: Play Tennis

Data set

Outlook = categorical	Temperature = categorical	Humidity = binary	Windy = binary	Play = CLASS
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**

Predict class label for

X=(outlook=sunny, Temperature=cool, Humadity=high, Windy=true)

Example: Play Tennis

Recap

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								

Probaility of play=yes given X is:

$$P(yes|X)=\frac{P(X_1|yes)P(X_2|yes)P(X_3|yes)P(X_4|yes)P(yes)}{P(X)}$$

Example: Play Tennis

Compare between P(yes|X) and P(no|X)

$$P(yes|X)=\frac{\frac{2}{9}\,\frac{3}{9}\,\frac{3}{9}\,\frac{3}{9}\,\frac{9}{14}}{P(X)}=\boxed{\frac{0.0053}{P(X)}}$$

$$P(no|X)=\frac{\frac{3}{5}\,\frac{1}{5}\,\frac{4}{5}\,\frac{3}{5}\,\frac{5}{14}}{P(X)}=\boxed{\frac{0.0206}{P(X)}}$$

Because value of ***P(yes|X)*** is **greater** than ***P(no|X)***, then test record of X = (Outlook = Sunny, Temperature = Cool, Humidity = High, Windy = true) will be classified as *class* label Play tennis = No.

References

1. Neapolitan, Richard, Bayesian Network, 2006