

Naïve Bayes Classifiers

Anna Monreale
Computer Science Department

Introduction to Data Mining, 2nd Edition
Chapter 5.3

Motivation

- Relationship between attributes and class labels may not be deterministic
- Reasons:
 - Noise in the data
 - Confounding factors affecting the classification and not in the data
- Bayesian Classifier exploit the **Bayes Theorem** that combines prior knowledge on the class labels with knowledge derivable from data

Bayes Classifier

- A probabilistic framework for solving classification problems.
- Let P be a probability function that assigns a number between 0 and 1 to events.
- $X = x$ an events is happening
- $P(X = x)$ is the probability that events $X = x$.
- Joint Probability $P(X = x, Y = y)$
- Conditional Probability $P(Y = y | X = x)$
- Relationship: $P(X,Y) = P(Y|X) P(X) = P(X|Y) P(Y)$
- Bayes Theorem: $P(Y|X) = P(X|Y)P(Y) / P(X)$
- Another Useful Property: $P(X =x) = P(X=x, Y=0) + P(X=x, Y=1)$

Bayes Theorem

- Consider a football game. Team 0 wins 65% of the time, Team 1 the remaining 35%. Among the game won by Team 1, 75% of them are won playing at home. Among the games won by Team 0, 30% of them are won at Team 1's field.
- **If Team 1 is hosting the next match, which team will most likely win?**
- Team 0 wins: $P(Y = 0) = 0.65$
- Team 1 wins: $P(Y = 1) = 0.35$
- Team 1 hosted the match, Team 1 wins: $P(X = 1 | Y = 1) = 0.75$
- Team 1 hosted the match Team 0 wins: $P(X = 1 | Y = 0) = 0.30$
- Objective $P(Y = 1 | X = 1)$

Bayes Theorem

Team 1 hosted the match, Team 1 wins

Team 1 wins

Team 1 hosted the match

$$P(Y = 1 | X = 1) = P(X = 1 | Y = 1)P(Y = 1) / P(X = 1) =$$

$$= 0.75 \times 0.35 / (P(X = 1, Y = 1) + P(X = 1, Y = 0))$$

$$= 0.75 \times 0.35 / (P(X = 1 | Y = 1)P(Y=1) + P(X = 1 | Y = 0)P(Y=0))$$

$$= 0.75 \times 0.35 / (0.75 \times 0.35 + 0.30 \times 0.65)$$

$$= 0.5738$$

- Therefore Team 1 has a better chance to win the match

Bayes Theorem for Classification

- X denotes the attribute sets, $X = \{X_1, X_2, \dots, X_d\}$
- Y denotes the class variable
- We treat the relationship probabilistically using $P(Y|X)$

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

Diagram illustrating the components of Bayes' Theorem:

- $P(Y|X)$ is labeled as Posterior Probability.
- $P(X|Y)$ is labeled as Likelihood.
- $P(Y)$ is labeled as Prior Probability.
- $P(X)$ is labeled as Evidence.

Bayes Theorem for Classification

- Learn the posterior $P(Y | X)$ for every combination of X and Y .
- By knowing these probabilities, a test record X' can be classified by finding the class Y' that maximizes the posterior probability $P(Y' | X')$.
- This is equivalent of choosing the value of Y' that maximizes $P(X' | Y')P(Y')$.
- How to estimate it?

Naïve Bayes Classifier

- It estimates the class-conditional probability by **assuming that the attributes are conditionally independent** given the class label y .
- The conditional independence is stated as:

$$P(X|Y = y) = \prod_{i=1}^d P(X_i|Y = y)$$

where each attribute set $X = \{X_1, X_2, \dots, X_d\}$

Conditional Independence

- Given three variables Y, X_1, X_2 we can say that Y is conditionally independent from X_1 given X_2 if the following condition holds:

$$P(Y | X_1, X_2) = P(Y | X_2)$$

- With the conditional independence assumption, instead of computing the class-conditional probability for every combination of X we only need to estimate the conditional probability of each X_i given Y .
- Thus, to classify a record the naive Bayes classifier computes the posterior for each class Y and takes the maximum class as result

$$P(Y|X) = P(Y) \prod_{i=1}^d P(X_i|Y = y) / P(X)$$

How to estimate ?

How to Estimate Probability From Data

- Class $P(Y) = N_y / N$
- N_y number of records with outcome y
- N number of records
- Categorical attributes
$$P(X = x \mid Y = y) = N_{xy} / N_y$$
- N_{xy} records with value x and outcome y
- $P(\text{Evade} = \text{Yes}) = 3/10$
- $P(\text{Marital Status} = \text{Single} \mid \text{Yes}) = 2/3$

<i>Tid</i>	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	Single	90K	Yes

How to Estimate Probability From Data

Continuous attributes

- Discretize the range into bins
 - Continuous vs nominal
 - Estimation: count records with class y and falling in the range
- Probability density estimation:
 - Assume attribute follows a normal distribution
 - 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 $P(X|y)$

How to Estimate Probability From Data

- Normal distribution

$$P(X_i = x_i \mid Y = y) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} e^{-\frac{(x_i - \mu_{ij})^2}{2\sigma_{ij}^2}}$$

- μ_{ij} can be estimated as the mean of X_i for the records that belongs to class y_j .
- Similarly, σ_{ij} as the standard deviation.
- $P(\text{Income} = 120 \mid \text{No}) = 0.0072$
 - mean = 110
 - std dev = 54.54

<i>Tid</i>	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	Single	90K	Yes

M-estimate of Conditional Probability

- If one of the conditional probability is zero, then the entire expression becomes zero.
- For example, given $X = \{\text{Refund} = \text{Yes}, \text{Divorced}, \text{Income} = 120\text{k}\}$, if $P(\text{Divorced} | \text{No})$ is zero instead of $1/7$, then
 - $P(X | \text{No}) = 3/7 \times 0 \times 0.00072 = 0$
 - $P(X | \text{Yes}) = 0 \times 1/3 \times 10^{-9} = 0$
- M-estimate $P(X|Y) = \frac{N_{xy}+mp}{N_y+m}$ (if $P(X|Y) = \frac{N_{xy}+1}{N_y+|Y|}$ is Laplacian estimation)
- m is a parameter, p is a user-specified parameter (e.g. probability of observing x_i among records with class y_j).
- In the example with $m = 3$ and $p = 1/m = 1/3$ (i.e., Laplacian estimation) we have

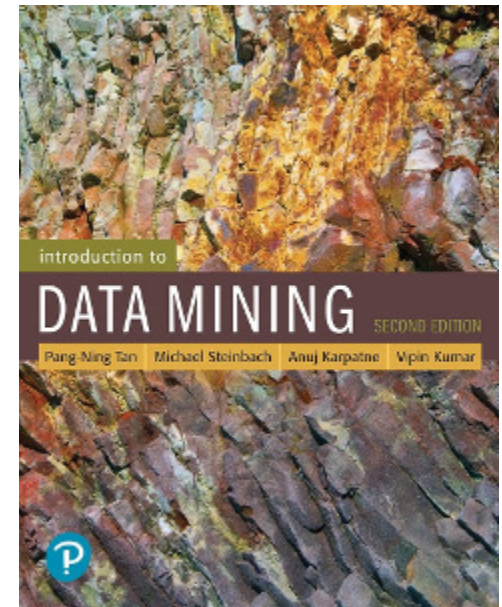
$$P(\text{Married} | \text{Yes}) = (0+3 \times 1/3)/(3+3) = 1/6$$

Naïve Bayes Classifier

- 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, not treated in this course)

References

- Bayesian Classifiers. Chapter 5.3. Introduction to Data Mining.



EXERCISE - NBC

Play-tennis example. estimating $P(x_i | C)$

Outlook	Temperature	Humidity	Windy	Class
sunny	hot	high	false	N
sunny	hot	high	true	N
overcast	hot	high	false	P
rain	mild	high	false	P
rain	cool	normal	false	P
rain	cool	normal	true	N
overcast	cool	normal	true	P
sunny	mild	high	false	N
sunny	cool	normal	false	P
rain	mild	normal	false	P
sunny	mild	normal	true	P
overcast	mild	high	true	P
overcast	hot	normal	false	P
rain	mild	high	true	N

$P(p) = 9/14$
$P(n) = 5/14$

outlook	
$P(\text{sunny} p) =$	$P(\text{sunny} n) =$
$P(\text{overcast} p) =$	$P(\text{overcast} n) =$
$P(\text{rain} p) =$	$P(\text{rain} n) =$
temperature	
$P(\text{hot} p) =$	$P(\text{hot} n) =$
$P(\text{mild} p) =$	$P(\text{mild} n) =$
$P(\text{cool} p) =$	$P(\text{cool} n) =$
humidity	
$P(\text{high} p) =$	$P(\text{high} n) =$
$P(\text{normal} p) =$	$P(\text{normal} n) =$
windy	
$P(\text{true} p) =$	$P(\text{true} n) =$
$P(\text{false} p) =$	$P(\text{false} n) =$

Play-tennis example. estimating $P(x_i | C)$

Outlook	Temperature	Humidity	Windy	Class
sunny	hot	high	false	N
sunny	hot	high	true	N
overcast	hot	high	false	P
rain	mild	high	false	P
rain	cool	normal	false	P
rain	cool	normal	true	N
overcast	cool	normal	true	P
sunny	mild	high	false	N
sunny	cool	normal	false	P
rain	mild	normal	false	P
sunny	mild	normal	true	P
overcast	mild	high	true	P
overcast	hot	normal	false	P
rain	mild	high	true	N

$P(p) = 9/14$
$P(n) = 5/14$

outlook	
$P(\text{sunny} p) = 2/9$	$P(\text{sunny} n) = 3/5$
$P(\text{overcast} p) = 4/9$	$P(\text{overcast} n) = 0$
$P(\text{rain} p) = 3/9$	$P(\text{rain} n) = 2/5$
temperature	
$P(\text{hot} p) = 2/9$	$P(\text{hot} n) = 2/5$
$P(\text{mild} p) = 4/9$	$P(\text{mild} n) = 2/5$
$P(\text{cool} p) = 3/9$	$P(\text{cool} n) = 1/5$
humidity	
$P(\text{high} p) = 3/9$	$P(\text{high} n) = 4/5$
$P(\text{normal} p) = 6/9$	$P(\text{normal} n) = 1/5$
windy	
$P(\text{true} p) = 3/9$	$P(\text{true} n) = 3/5$
$P(\text{false} p) = 6/9$	$P(\text{false} n) = 2/5$

Play-tennis example. estimating $P(x_i | C)$

$P(p) = 9/14$
$P(n) = 5/14$

Outlook	Temperature	Humidity	Windy	Class
rain	hot	high	false	?

outlook	
$P(\text{sunny} p) = 2/9$	$P(\text{sunny} n) = 3/5$
$P(\text{overcast} p) = 4/9$	$P(\text{overcast} n) = 0$
$P(\text{rain} p) = 3/9$	$P(\text{rain} n) = 2/5$
temperature	
$P(\text{hot} p) = 2/9$	$P(\text{hot} n) = 2/5$
$P(\text{mild} p) = 4/9$	$P(\text{mild} n) = 2/5$
$P(\text{cool} p) = 3/9$	$P(\text{cool} n) = 1/5$
humidity	
$P(\text{high} p) = 3/9$	$P(\text{high} n) = 4/5$
$P(\text{normal} p) = 6/9$	$P(\text{normal} n) = 1/5$
windy	
$P(\text{true} p) = 3/9$	$P(\text{true} n) = 3/5$
$P(\text{false} p) = 6/9$	$P(\text{false} n) = 2/5$

$$P(X|p) \cdot P(p) =$$

$$P(X|n) \cdot P(n) =$$

Play-tennis example. estimating $P(x_i | C)$

$P(p) = 9/14$
$P(n) = 5/14$

Outlook	Temperature	Humidity	Windy	Class
rain	hot	high	false	N

outlook	
$P(\text{sunny} p) = 2/9$	$P(\text{sunny} n) = 3/5$
$P(\text{overcast} p) = 4/9$	$P(\text{overcast} n) = 0$
$P(\text{rain} p) = 3/9$	$P(\text{rain} n) = 2/5$
temperature	
$P(\text{hot} p) = 2/9$	$P(\text{hot} n) = 2/5$
$P(\text{mild} p) = 4/9$	$P(\text{mild} n) = 2/5$
$P(\text{cool} p) = 3/9$	$P(\text{cool} n) = 1/5$
humidity	
$P(\text{high} p) = 3/9$	$P(\text{high} n) = 4/5$
$P(\text{normal} p) = 6/9$	$P(\text{normal} n) = 1/5$
windy	
$P(\text{true} p) = 3/9$	$P(\text{true} n) = 3/5$
$P(\text{false} p) = 6/9$	$P(\text{false} n) = 2/5$

$$P(X|p) \cdot P(p) = P(\text{rain}|p) \cdot P(\text{hot}|p) \cdot P(\text{high}|p) \cdot P(\text{false}|p) \cdot P(p) = 3/9 \cdot 2/9 \cdot 3/9 \cdot 6/9 \cdot 9/14 = 0.010582$$

$$P(X|n) \cdot P(n) = P(\text{rain}|n) \cdot P(\text{hot}|n) \cdot P(\text{high}|n) \cdot P(\text{false}|n) \cdot P(n) = 2/5 \cdot 2/5 \cdot 4/5 \cdot 2/5 \cdot 5/14 = 0.018286$$

a) Naive Bayes (3 points)

Given the training set below, build a Naive Bayes classification model (i.e. the corresponding table of probabilities) using (i) the normal formula and (ii) using Laplace formula. What are the main effects of Laplace on the models?

A	B	class
no	green	N
no	red	Y
yes	green	N
no	red	N
no	red	Y
no	green	Y
yes	green	N

Answer:

Normal

		Y	N			Y	N
			3		4	0.43	0.57
		A Y	A N			A Y	A N
yes		0	2	yes		0.00	0.50
no		3	2	no		1.00	0.50
		B Y	B N			B Y	B N
green		1	3	green		0.33	0.75
red		2	1	red		0.67	0.25

Laplace

		Y	N			Y	N
			3		4	0.43	0.57
		A Y	A N			A Y	A N
yes		0	2	yes		0.20	0.50
no		3	2	no		0.80	0.50
		B Y	B N			B Y	B N
green		1	3	green		0.40	0.67
red		2	1	red		0.60	0.33