Naïve Bayes Classifier

Bayes Classifier

- A probabilistic framework for solving classification problems

- Bayes theorem:

$$P(A,C) = P(A)P(C \mid A) = P(A \mid C)P(C)$$

Bayes Classifier

Conditional Probabilities:

$$P(C \mid A) = \frac{P(A,C)}{P(A)}$$

$$P(A \mid C) = \frac{P(A,C)}{P(C)}$$

$$P(A,C) = P(A)P(C \mid A) = P(A \mid C)P(C)$$

Bayes Classifier

We have multiple attributes $(A_1, A_2, \ldots, A_n)$

Goal is to predict class C

Specifically, we want to find the value of C that maximizes $P(C \mid A_1, A_2, \ldots, A_n)$

Can we estimate $P(C \mid A_1, A_2, \ldots, A_n)$ directly from data?

Bayes Classifier

- Approach:
  - *compute the posterior probability* $P(C \mid A_1, A_2, \ldots, A_n)$ *for all values of C using the Bayes theorem*

$$P(C \mid A_1 A_2 \ldots A_n) = \frac{P(A_1 A_2 \ldots A_n \mid C) P(C)}{P(A_1 A_2 \ldots A_n)}$$

  - *Choose value of C that maximizes* $P(C \mid A_1, A_2, \ldots, A_n)$

  - *Equivalent to choosing value of C that maximizes*
    $P(A_1, A_2, \ldots, A_n \mid C)\, P(C)$

- How to estimate $P(A_1, A_2, \ldots, A_n \mid C)$?

Naïve Bayes Classifier

- Assume independence among attributes A_i when class is given:
  - $P(A_1, A_2, \ldots, A_n | C) = P(A_1 | C_1)\, P(A_2 | C_2) \ldots P(A_n | C_j)$
  - *Can estimate* $P(A_i | C_j)$ *for all* A_i *and* C_j
  - *the new pattern is classified to* C_j *if* $P(C_j)\, \Pi\, P(A_i | C_j)$ *is maximum*

Example 1

PlayTennis: training examples

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Algorithm

- Algorithm: Discrete-Valued Features

-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 feature value x_{jk} of each feature 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 $\mathbf{S}$;

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

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

Look up tables to assign the label ***c**** to **X′** if

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

Learning Phase :

Outlook	Play=*Yes*	Play=*No*
Sunny	2/9	3/5
Overcast	4/9	0/5
Rain	3/9	2/5

Temperature	Play=*Yes*	Play=*No*
Hot	2/9	2/5
Mild	4/9	2/5
Cool	3/9	1/5

Humidity	Play=*Yes*	Play=No
High	3/9	4/5
Normal	6/9	1/5

Wind	Play=*Yes*	Play=*No*
Strong	3/9	3/5
Weak	6/9	2/5

P(Play=Yes) = 9/14

P(Play=No) = 5/14

- **Test Phase :**

-Given a new instance, predict its label

x′=(Outlook=Sunny, Temperature=Cool, Humidity=High, Wind=Strong)

-Look up tables achieved in the learning phrase

P(Outlook=Sunny|Play=Yes) = 2/9	P(Outlook=Sunny|Play=No) = 3/5
P(Temperature=Cool|Play=Yes) = 3/9	P(Temperature=Cool|Play==No) = 1/5
P(Huminity=High|Play=Yes) = 3/9	P(Huminity=High|Play=No) = 4/5
P(Wind=Strong|Play=Yes) = 3/9	P(Wind=Strong|Play=No) = 3/5
P(Play=Yes) = 9/14	P(Play=No) = 5/14

-Decision making with the MAP rule:

P(Yes|x′): [P(Sunny|Yes) P(Cool|Yes) P(High|Yes) P(Strong|Yes)] P(Play=Yes) = 0.0053

P(No|x′): [P(Sunny|No) P(Cool|No) P(High|No) P(Strong|No)] P(Play=No) = 0.0206

Given the fact **P(Yes|x′) < P(No|x′)** , we label **x′** to be "No".

Example 2

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

Give Birth	Can Fly	Live in Water	Have Legs	Class
yes	no	yes	no	?

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

Give Birth	Can Fly	Live in Water	Have Legs	Class
yes	no	yes	no	?

A: attributes

M: mammals

N: non-mammals

$$P(A|M) = \frac{6}{7} \times \frac{6}{7} \times \frac{2}{7} \times \frac{2}{7} = 0.06$$

$$P(A|N) = \frac{1}{13} \times \frac{10}{13} \times \frac{3}{13} \times \frac{4}{13} = 0.0042$$

$$P(A|M)P(M) = 0.06 \times \frac{7}{20} = 0.021$$

$$P(A|N)P(N) = 0.004 \times \frac{13}{20} = 0.0027$$

P(A|M)P(M) > P(A|N)P(N)

Class: Mammals

Application

- Text classification
- Spam filtering
- Hybrid recommender system
  - *Recommender system apply machine learning and data mining techniques for filtering unseen information and can predict whether a user would like a given resource*
- Online application
  - *Simple emotion modeling*

Naïve Bayes (Summary)

- 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)*