BITS F464: Machine Learning

19 Naive Bayes Classifier

Dr. Kamlesh Tiwari
Assistant Professor, Department of CSIS,
BITS Pilani, Pilani Campus, Rajasthan-333031 INDIA

March 08, 2021 ONLINE (Campus @ BITS-Pilani Jan-May 2021)

http://ktiwari.in/ml

Probability: Independence

Assume there are
three T-shirts T_r, T_g, T_b
and
three shorts S_r, S_g, S_b
in a shop

What is the probability that someone would purchase (T_r, S_r)

Independence

Selling any T-shirt or shorts

$$Pr(T_r, S_r) = Pr(T_r) \times Pr(S_r) = 1/3 \times 1/3 = 1/9$$

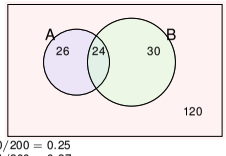
Dependence

Selling T-shirt and shorts in set of same color

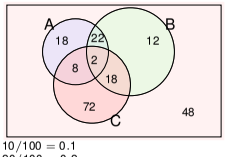
$$Pr(T_r, S_r) = Pr(T_r \ \& \ S_r) = 1/3$$

There is something called conditional independence

Probability: Conditional independence



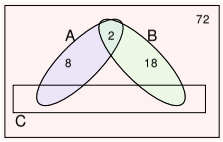
$p(A) = 50/200 = 0.25$
$p(B) = 54/200 = 0.27$
$p(A) \times p(B) = 0.0675$
$p(A \cap B) = 24/200 = 0.12 \neq p(A) \times p(B)$



$p(A|C) = 10/100 = 0.1$
$p(B|C) = 20/100 = 0.2$
$p(A \cap B|C) = 2/100 = 0.02 = p(A|C) \times p(B|C)$

Independence is obtained by C

Independence does not mean conditional independence



- A, and B are independent due to the construction.
- But, in C. Happening of one completely rules out the other.

Bayes Optimal Classifier

Switching the question, from "which is most probable hypothesis?" to "what is the most probable classification of the new instance?"
Is it possible to do better then MAP?

Example: Let posterior probabilities of three hypotheses h_1, h_2, h_3 given the training data are 0.4, 0.3, and 0.3 (obviously h_1 is MAP)

- Let classification of a new instance x is positive by h_1 and negative by h_2 and h_3
- By taking all hypotheses into account, the probability that x is positive is 0.4, and negative is 0.6
  - *Most probable classification is negative and it differs from MAP*

Bayes optimal classification:

$$\underset{v_j \in V}{\text{argmax}} \sum_{h_i \in H} P(v_j|h_i) P(h_i|D)$$

where classification $v_j \in V$ and $P(v_j|D)$ is the correct classification

Bayes Optimal Classifier

$$\boxed{\text{argmax}_{v_j \in V} \sum_{h_i \in H} P(v_j|h_i) P(h_i|D)}$$

$V = \{\oplus, \ominus\}$

$P(h_1\|D)$=0.4	$P(\ominus\|h_1)$=0	$P(\oplus\|h_1)$=1
$P(h_2\|D)$=0.3	$P(\ominus\|h_2)$=1	$P(\oplus\|h_2)$=0
$P(h_3\|D)$=0.3	$P(\ominus\|h_3)$=1	$P(\oplus\|h_3)$=0

Therefore,

$$\sum_{h_i \in H} P(\oplus|h_i) P(h_i|D) = 0.4 \qquad \sum_{h_i \in H} P(\ominus|h_i) P(h_i|D) = 0.6$$

and

$$\text{argmax}_{v_j \in \{\oplus, \ominus\}} \sum_{h_i \in H} P(v_j|h_i) P(h_i|D) = \ominus$$

This type of classifier is called a Bayes optimal classifier, or Bayes optimal learner.

Bayes Optimal Classifier

- **Best Classifier:** No other classification method using the same hypothesis space and same prior knowledge can outperform this method on an average.
- This method maximizes the probability that the new instance is classified correctly, given the available data, hypothesis space, and prior probabilities over the hypotheses
- **Note** that the predictions made by Bayes optimal classifier may not contained in H. (It is possible that there is no hypothesis in H giving same classification as Bayes optimal classifier.)

Issue: Although the Bayes optimal classifier obtains the best performance that can be achieved from the given training data, it can be quite costly to apply[1]

[1] we could have infinitely many hypothesis

GIBBS Algorithm [2]

A less optimal method is the Gibbs algorithm

- For a new instance x
  1. Choose a hypothesis $h \in H$ at random, according to the posterior probability distribution over H
  2. Use h to predict the classification of the instance x

Importance

Under certain conditions the expected misclassification error for the Gibbs algorithm is at most twice the expected error of the Bayes optimal classifier

[2] Opper, Manfred and Haussler, David, "Generalization performance of Bayes optimal classification algorithm for learning a perceptron", In Physical Review Letters, 66(20), pp-2677, APS-1991

Naive Bayes Classifier

Bayes classifier is a highly practical Bayesian learning method

- In some domains, its performance found to be comparable to neural network and decision tree
- The Bayesian approach to classify a new instance is to assign the most probable target value describing the instance

$$v_{MAP} = \text{argmax}_{v_j \in V} P(v_j | a_1, a_2, ..., a_n)$$

- We can use Bayes theorem to rewrite this expression as

$$\begin{aligned} v_{MAP} &= \underset{v_j \in V}{\text{argmax}} \frac{P(a_1, a_2, ..., a_n | v_j)\ P(v_j)}{P(a_1, a_2, ..., a_n)} \\ &= \underset{v_j \in V}{\text{argmax}} P(a_1, a_2, ..., a_n | v_j)\ P(v_j) \end{aligned} \quad (1)$$

Naive Bayes has assumption is that the attribute values are conditionally independent given the target value

Naive Bayes Classifier

Attribute values are conditionally independent given the target value

- Under this assumption,
- Given a target value, the probability of observing the conjunction $< a_1, a_2, ..., a_n >$ is just the product of the probabilities.

$$P(a_1, a_2, ..., a_n | v_j) = \Pi_i P(a_i | v_j)$$

Naive Bayes classifier

is the one which

$$\underset{v_j \in V}{\text{argmax}} P(v_j)\ \Pi_i P(a_i | v_j)$$

Caution: "conditionally independent" assumption may NOT be true

Example: Naive Bayes Classification

Given the data

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

Determine classification for $< Rainy, Hot, High, Strong >$

Example: Naive Bayes Classification

Day	Outlook	Temperature	Humidity	Wind	Play
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D6	Rainy	Cool	Normal	Strong	No
D8	Sunny	Mild	High	Weak	No
D14	Rainy	Mild	High	Strong	No

Day	Outlook	Temperature	Humidity	Wind	Play
D3	Overcast	Hot	High	Weak	Yes
D4	Rainy	Mild	High	Weak	Yes
D5	Rainy	Cool	Normal	Weak	Yes
D7	Overcast	Cool	Normal	Strong	Yes
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rainy	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes

Outlook

P(Yes) = 9/14

P(No) = 5/14

	Yes	No
Sunny	2/9	3/5
Overcast	4/9	0/5
Rainy	3/9	2/5

Example: Naive Bayes Classification

- P(Yes) = 9/14 P(No) = 5/14

Outlook

	Yes	No
Sunny	2/9	3/5
Overcast	4/9	0/5
Rainy	3/9	2/5

Humidity

	Yes	No
High	3/9	4/5
Low	6/9	1/5

Wind

	Yes	No
Strong	3/9	3/5
Weak	6/9	2/5

Temperature

	Yes	No
Hot	2/9	2/5
Mild	4/9	2/5
Cool	3/9	1/5

Example: Naive Bayes Classification

For $x =< Rainy, Hot, High, Strong >$

P(Yes)

- $P(Yes) \times P(x|Yes)$
- $P(Yes) \times [P(Rainy|Yes) \times P(Hot|Yes) \times P(High|Yes) \times P(Strong|Yes)]$
- $9/14 \times [3/9 \times 2/9 \times 3/9 \times 3/9]$
- 0.005291...

P(No)

- $P(No) \times P(x|No)$
- $P(No) \times [P(Rainy|No) \times P(Hot|No) \times P(High|No) \times P(Strong|No)]$
- $5/14 \times [2/5 \times 2/5 \times 4/5 \times 3/5]$
- 0.027428...

So the classification of the new data item x is No

Laplace smoothing

Consider data = [apple,mango,mango,apple,banana,apple,apple,mango,apple,mango]

- Here there are three class in the data {mango,apple,banana}
- So p(apple)=5/10, p(mango)=4/, p(banana)=1/10
- Can we add some hypothetical points in data? say $k = 1$ we are adding one instance of all the items in the data. then the probabilities are p(apple)=6/13, p(mango)=5/13, p(banana)=2/13
- This method is called Laplace smoothing
- Very useful for the cases when number of classes is large and data size is small. Due to this we may miss some of the classes appearing in the data. See the example below

Outlook	Yes	No
Sunny	2/9	3/5
Overcast	4/9	0/5
Rainy	3/9	2/5

Outlook	Yes	No
Sunny	3/12	4/8
Overcast	5/12	1/8
Rainy	4/12	3/8

Thank You!

Thank you very much for your attention! (Reference[3])

Queries ?

[3] [1] Book - *Machine Learning*, ch-6, Tom M. Mitchell.