Naïve Bayes

Arun Chauhan,IIIT Dharwad

Medical Tests and Bayes' Theorem

- Let a probability of disease is 1 in 10,000 and the test accuracy of the disease is 99 %
- Let event A is the event you have this disease, and event B is the event that you test positive.
- Data: P(A) = 0.0001 P(~A) = 0.9999

  P(B|A) = 0.99 P(~B|A) = 0.01

  P(B|~A) = 0.01 P(~B|~A) = 0.99
- Given test is positive what is the probability that disease is actually present.

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

$$P(A|B) = \frac{0.99*0.0001}{0.99*0.0001+0.01*0.9999} = 0.0098$$

$$P(\sim A|B) = \frac{0.01*0.9999}{0.99*0.0001+0.01*0.9999} = 0.9902$$

Bayes Algorithm

- Consider a supervised learning in which we want to approximate an unknown target function
- f: X → Y   or equivalently P(Y|X)
- Assume :   Y is Boolean-valued random variable

  $X = \langle X_1, X_2, \ldots\ldots, X_n \rangle$ where X_i is Boolean random variable.
- By applying Bayes rule we can represent
-

$$P(Y = y_i | X = x_k) = \frac{P(X = x_k | Y = y_i) P(Y = y_i)}{\sum_j P(X = x_k | Y = y_j) P(Y = y_j)}$$

Unbiased Learning of Bayes Classifiers is Impractical

- If we want to estimate P(Y) and P(X|Y), how much training data we required?
- For boolean random variable Y, it is observed approx. 100 example are enough to estimate the probability near to true probability.
- But what about P(X|Y)? How many example would be required?
- As, X is vector of *n* boolean variable. We need to estimate set of parameters.

$$\theta_{ij} \equiv P(X = x_i | Y = y_j)$$

  where, index i takes on 2^n possible values.
- Therefore, how many parameters we need to estimate? Ans: $(2^{n+1} - 1)$ Why?
- $(2^n - 1)$ for Y=1 $+ (2^n - 1)$ for Y=0 which equals to $2^{n+1} - 2$
- 1 for P(Y=1)
- Therefore: $2^{n+1} - 2 + 1 = 2^{n+1} - 1$

Naïve Bayes Algorithm

- Is there any way to reduce the sample complexity?
- Naïve Bayes does it by assuming conditional Impendence.
- Here we assume attributes in vector X are conditionally independent given Y.
- Remember that two event A and B are independent if :

$$P(A \cap B) = P(A)P(B), \quad \text{or equivalently, } P(A|B) = P(A)$$

- Similarly, event A and B are conditionally independent if :

$$P(A \cap B|C) = P(A|C)P(B|C) \quad \text{or equivalently, } \quad P(A|B,C) = P(A|C)$$

- Recall that $P(A|B) = \frac{P(A \cap B)}{P(B)}$
- By conditioning on C, we obtain: $P(A|B,C) = \frac{P(A \cap B|C)}{P(B|C)}$
- If A and B are conditionally independent given C, we obtain

$$\begin{aligned} P(A|B,C) &= \frac{P(A \cap B|C)}{P(B|C)} \\ &= \frac{P(A|C)P(B|C)}{P(B|C)} \\ &= P(A|C). \end{aligned}$$

Conditional Independence

- Naïve Bayes algorithm makes assumption that each X_i is conditionally independent of each other given Y, where $\mathbf{X} = \langle X_1 ,\ X_2 ,\ \ldots\ldots, X_n \rangle$
- The value of this assumption is that it dramatically simplifies the representation of P(X|Y), and the problem of estimating it from the training data.
- For example, if $\mathbf{X} = \langle X_1 ,\ X_2 \rangle$, then

$$P(\mathbf{X}|Y) = P(X_1 ,\ X_2 \mid Y) = P(X_1 \mid Y) P(X_2 \mid Y)$$

- In general, when $\mathbf{X} = \langle X_1 ,\ X_2 ,\ \ldots\ldots, X_n \rangle$

$$P(X_1 ,\ X_2 ,\ \ldots\ldots, X_n \mid Y) = \prod_{\{i\,:\,1 \text{ to } n\}} P(X_i \mid Y)$$

- In this case the number of parameters reduced to $2n + 1$ (Vs ($2^{n+1} - 1$))? Why?

Derivation of Naive Bayes Algorithm

- Let Y is any discrete-valued random variable and the attributes of X (X_1, X_2, ……, X_n) be any discrete or real valued attributes.
- Our goal is to train the classifier that will output the probability distribution over the possible values of Y.
- The expression to estimate the probability when Y takes it k_{th} value is as follows:

$$P(Y = y_k | X_1 \dots X_n) = \frac{P(Y = y_k) P(X_1 \dots X_n | Y = y_k)}{\sum_j P(Y = y_j) P(X_1 \dots X_n | Y = y_j)}$$

- Assuming X_i conditionally independent given Y, then:

$$P(Y = y_k | X_1 \dots X_n) = \frac{P(Y = y_k) \prod_i P(X_i | Y = y_k)}{\sum_j P(Y = y_j) \prod_i P(X_i | Y = y_j)}$$

- Hence to classify new X:

$$Y \leftarrow \arg\max_{y_k} \frac{P(Y = y_k) \prod_i P(X_i | Y = y_k)}{\sum_j P(Y = y_j) \prod_i P(X_i | Y = y_j)}$$

Naïve Bayes Example

Outlook	*Temperature*	*Humidity*	*Wind*	*PlayTennis*
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mild	High	Strong	No

P(Playtennis = Yes) =
P(Outlook = Sunny | Tennis = Yes) =
P(Temprature =Cool | Tennis = Yes) =
P(Humidity = High | Tennis = Yes) =
P(Wind = Strong | Tennis = Yes) =

P(Playtennis = No) =
P(Outlook = Sunny | Tennis = No) =
P(Temprature =Cool | Tennis = No) =
P(Humidity = High | Tennis = No) =
P(Wind = Strong | Tennis = No) =

P(yes|Sunny,Cool,High,Strong)=

P(No|Sunny,Cool,High,Strong)=

Naïve Bayes Example

Outlook	*Temperature*	*Humidity*	*Wind*	*PlayTennis*
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mild	High	Strong	No

P(Playtennis = Yes) = 9/14
P(Outlook = Sunny | Tennis = Yes) = 2/9
P(Temprature =Cool | Tennis = Yes) = 3/9
P(Humidity = High | Tennis = Yes) = 3/9
P(Wind = Strong | Tennis = Yes) = 3/9

P(Playtennis = No) = 5/14
P(Outlook = Sunny | Tennis = No) = 3/5
P(Temprature =Cool | Tennis = No) = 1/5
P(Humidity = High | Tennis = No) = 4/5
P(Wind = Strong | Tennis = No) = 3/5

P(yes|Sunny,Cool,High,Strong)=
9/14 * 2/9 * 3/9 * 3/9 * 3/9 = 0.0053

P(No|Sunny,Cool,High,Strong)=
5/14 * 3/5 * 1/5 * 4/5 * 3/5 = 0.206

References

- http://www.cs.cmu.edu/~tom/mlbook/NBayesLogReg.pdf
- Chapter 6. Machine Learning, Tom Mitchell