IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

NPTEL

NAÏVE BAYES PART-2

LECTURE 32

DR. GAURAV DIXIT

DEPARTMENT OF MANAGEMENT STUDIES

NAÏVE BAYES

- Bayes Model for classification
  - Predictors should also be categorical
  - Numerical variables will have to be converted into categorical variables through binning
- Open Excel
- Complete or Exact Bayes Limitations
  - For a model even with small no. of predictors, many new observations to be classified might not get exact matches
  - Probability of a match might reduce significantly on adding just one variable to the set of predictors

NAÏVE BAYES

- Instead of Complete or Exact Bayes, switch to Naïve Bayes
  - In Naïve Bayes, all the records are used instead of relying on just the matching records
- Naïve Bayes Modification
  - For class i of outcome variable, compute the probabilities (P_1, P_2, …, P_p) of belonging to class i for each predictor's value (x_1, x_2, …, x_p) taken by the new observation to be classified
  - Compute $P_1 \times P_2 \times \ldots \times P_p \times P(C_i)$
  - Execute previous two steps for all the classes

NAÏVE BAYES

- Naïve Bayes Modification
  - To compute the probability of the new observation belonging to class i, divide the value computed in step 2 by the summation of values computed in step 2 for all the classes
  - Execute previous step for all the classes
  - Classify the new observation to the class with the highest probability value

NAÏVE BAYES

- Naïve Bayes formula

$$P(C_i|x_1, x_2, \ldots, x_p) = \frac{[P(x_1|C_i)P(x_2|Ci) \ldots P(xp|Ci)]P(Ci)}{[P(x_1|C_1)P(x_2|C_1) \ldots P(xp|C_1)]P(C_1) + \ldots + [P(x_1|Cm)P(x_2|Cm) \ldots P(xp|Cm)]P(Cm)}$$

  - Naïve Bayes formula is directly derived from the exact Bayes formula after making following assumption:
  - Predictors' values $\{x_1, x_2, \ldots, xp\}$ occur independent of each other for a given class

$$P(x_1, x_2, \ldots, xp|Ci) \equiv P(x_1|Ci)P(x_2|Ci) \ \ldots P(xp|Ci)$$

NAÏVE BAYES

- Naïve Bayes formula
  - For classification, naïve Bayes formula works quite well
  - Since we don't require probabilities values to be accurate in absolute term, rather just a reasonably accurate rank ordering of these values
  - For the same reason, we should use the numerator only and drop the denominator which is common for all the classes
- Steps when we have a class of interest
  - User specified cut off value for the class of interest

NAÏVE BAYES

- Steps when we have a class of interest
  - Compute the probabilities (P_1, P_2, …, P_p) of belonging to class of interest for each predictor’s value (x_1, x_2, …, x_p) taken by the new observation to be classified
  - Compute $P_1 \times P_2 \times \ldots \times P_p \times P(\text{Class of interest})$
  - Execute previous two steps for all the classes
  - To compute the probability of the new observation belonging to class of interest, divide the value computed in step 2 by the summation of values computed in step 2 for all the classes

Key References

- Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data by EMC Education Services (2015)
- Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner by Shmueli, G., Patel, N. R., & Bruce, P. C. (2010)

Thanks…