Data Mining

Classification III – Naïve Bayes (Part B)

Dr. Jason T.L. Wang, Professor

Department of Computer Science

New Jersey Institute of Technology

Example

- The training dataset on the following page has 10 students' performance records.
- We want to predict whether a new student will pass or fail the course, given his/her performance record.

Training Data Set

Student	Assignment	Project	Exam	Label
1	Good	A	High	Pass
2	Good	B	High	Pass
3	Bad	B	Low	Fail
4	Bad	C	High	Fail
5	Good	C	Low	Fail
6	Good	C	High	Pass
7	Bad	B	High	Pass
8	Good	A	Low	Pass
9	Bad	A	Low	Fail
10	Good	B	Low	Pass

Training Phase

P(Pass) = 6/10; P(Fail) = 4/10

P(Good|Pass) = 5/6; P(Bad|Pass) = 1/6

P(High|Fail) = 1/4; P(Low|Fail) = 3/4

• • •

Testing Phase (Making Prediction)

- Testing phase

  Suppose we are given a new instance (a new student's performance record) and we want to predict if this student will pass or fail the course.

  - Given: unlabeled test record X = (Bad, A, High)
  - Want: predict the label of X

Testing Phase (Making Prediction)

- Recall that

  $Y \leftarrow \arg\max P(Yi)\ P(X1|Yi)P(X2|Yi)\ldots P(Xn|Yi)$

  When Yi = Pass,

  $P(Yi)P(X1|Yi)P(X2|Yi)P(X3|Yi)$

  $=P(Pass)P(Bad|Pass)P(A|Pass)P(High|Pass)$

  $=(6/10)(1/6)(2/6)(4/6)$

  $= 0.02222$

Testing Phase (Making Prediction)

When Yi = Fail,

$P(Yi)P(X1|Yi)P(X2|Yi)P(X3|Yi)$

$=P(Fail)P(Bad|Fail)P(A|Fail)P(High|Fail)$

$=(4/10)(3/4)(1/4)(1/4)$

$= 0.01875$

Given X = (Bad, A, High)

$P(Fail|X) < P(Pass|X) \leftarrow 0.01875 < 0.02222$

Testing Phase (Making Prediction)

- Therefore, we predict that the new student with performance record X = (Bad, A, High) will pass the course, based on the Naïve Bayes algorithm.

Some Notes

- We assume that Xi's are conditionally independent given Y (hence the algorithm is called "naïve").
- P(Y) and P(Xi|Y) are all obtained from the training data.
- The size of the training data set matters.

End of
Naïve Bayes Module (Part B)