# CSC-272: Big Data Spring 2020

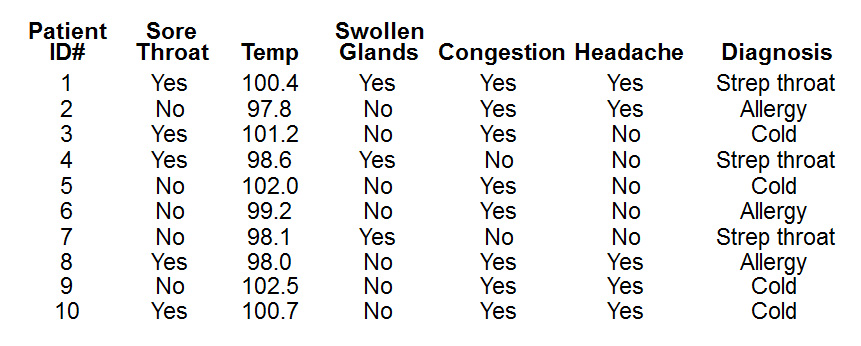
**PROBLEM SET #4 – Due Wednesday, March 18 25 @ 10:30am**

**Name: Boone Tison**

*You may use whatever combination of word processing and hand-written work is most clear.*

***Numerical Naïve Bayes***

Consider the following dataset:



We will begin by picking up where problem set #3 left off. You didn’t think you were going to avoid a numerical Naïve Bayes problem, did you? ☺

1. Complete the chart on the next page. It is completely filled out for all of the nominal attributes. That is – you only have to fill out the Temp section. Refer to Table 4.4 on page 101 of the text for a refresher. You may use a calculator, spreadsheet, or other method to calculate the mean and standard deviation, *but you must calculate it yourself*. You will be expected to perform these calculations again later on without any help.

2. Use this data to compute the Naïve Bayes probabilities for the following new instance. Show all work. Be sure to normalize the answers as shown in lecture. Your answers below should be given in percentages, and they should add up to 100%. Use the probability density function from page 102 of the text. Once again, you may use an automated method to calculate it, but you must do it yourself. You may check with others to make sure you got it right. (Make sure that your probability density calculations are clear in your work.)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Sore Throat** | **Fever** | **Swollen Glands** | **Congestion** | **Headache** | **Diagnosis** |
| NO | 98.6 | NO | NO | YES | ???? |

***Probability of Strep Throat:***

= 1/3 \* .442 \* 1/5 \* 2/3 \* 1/3 \* 3/10 = 0.00196444444

= .02%

***Probability of Allergy:***

= 2/3 \* .738 \* 4/5 \* 1/5 \* 2/3 \* 3/10 = 0.015744

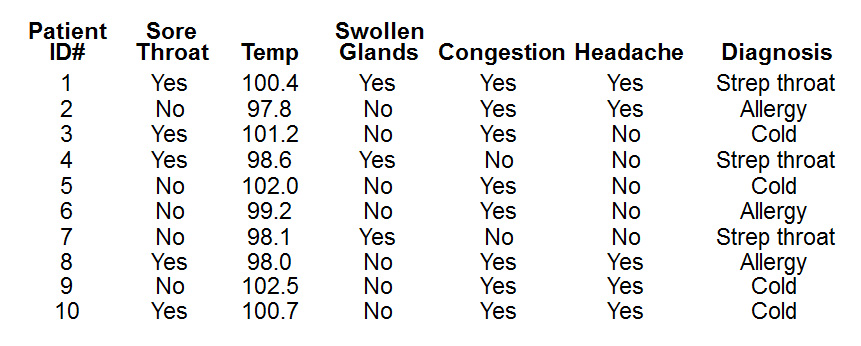
= .21%

***Probability of Cold:***

= 2/4 \*522.062 \* 5/6 \* 1/6 \* 2/4 \* 4/10 = 7.2508611111

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | ***Sore Throat*** | | | ***Temp*** | | | ***Swollen Glands*** | | | ***Congestion*** | | | ***Headache*** | | | ***Diagnosis*** | | |
|  | Strep | Allergy | Cold | Strep | Allergy | Cold | Strep | Allergy | Cold | Strep | Allergy | Cold | Strep | Allergy | Cold | Strep | Allergy | Cold |
| YES | 2 | 1 | 2 | 100.4, 98.6, 98.1 | 97.8,  99.2,  98.0 | 101.2,  102.0,  102.5,  100.7 | 3 | 0 | 0 | 1 | 3 | 4 | 1 | 2 | 2 | 3 | 3 | 4 |
| NO | 1 | 2 | 2 |  |  |  | 0 | 3 | 4 | 2 | 0 | 0 | 2 | 1 | 2 |  |  |  |
| YES | 2/3 | 1/3 | 2/4 | µ=99.0 | µ=98.3 | µ=101.6 | 3/3 | 0/3 | 0/4 | 1/3 | 3/3 | 4/4 | 1/3 | 2/3 | 2/4 | 3/10 | 3/10 | 4/10 |
| NO | 1/3 | 2/3 | 2/4 | σ=0.98 | σ=0.61 | σ=0.69 | 0/3 | 3/3 | 4/4 | 2/3 | 0/3 | 0/4 | 2/3 | 1/3 | 2/4 |  |  |  |
| YES\* |  |  |  |  |  |  | 4/5 | 1/5 | 1/6 |  | 4/5 | 5/6 |  |  |  |  |  |  |
| NO\* |  |  |  |  |  |  | 1/5 | 4/5 | 5/6 |  | 1/5 | 1/6 |  |  |  |  |  |  |

= 99.75%

***Building Decision Trees***

3. Continue to consider the same dataset. Use the algorithm discussed in lecture to construct a valid decision tree for this data. Handle the Temp attribute using discretization with a minimum bucket size of 2. Do not use a binary split. (Based on your experience with the last problem set, you shouldn’t really need that last bit of instruction.) Show all of your work with the format used in lecture. Make sure that your *goodness* calculations are clear. Explain yourself when it would be helpful.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 97.8 | 98.0 | 98.1 | 98.6 | 99.2 | 100.4 | 100.7 | 101.2 | 102.0 | 102.5 |
| Allergy | Allergy | Strep | Strep | Allergy | Strep | Cold | Cold | Cold | Cold |

(-inf – 98.0] = Allergy 2/2 (98.0 – 100.4] = Strep ¾ (100.4 – inf) = Cold 4/4

Sore Throat: Yes -> Strep Throat\* (2/5) Accuracy = 40% Goodness = 20

No -> Cold\* (2/5)

Temperature: <=98.0 -> Allergy (2/2) Accuracy = 90% Goodness = 90

98>x<=100.4 -> Strep (3/4)

>100.4 -> Cold (4/4)

Swollen: Yes -> Strep Throat (3/3) Accuracy = 70% Goodness = 70

No -> Cold (4/7)

Congestion: Yes -> Cold (4/8) Accuracy = 60% Goodness = 60

No -> Strep throat (2/2)

Headache: Yes -> Cold\* (2/5) Accuracy = 40% Goodness = 20

No -> Strep\* (2/5)

Focus now on branch with ¾ - Patient IDs: 7, 4, 6, 1

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Patient ID | Sore Thro | Temp | Swollen | Congestion | Headache | Diagnosis |
| 1 | Yes | 100.4 | Yes | Yes | Yes | Strep |
| 4 | Yes | 98.6 | Yes | No | No | Strep |
| 6 | No | 99.2 | No | Yes | No | Allergy |
| 7 | No | 98.1 | Yes | No | No | Strep |

Sore Throat: Yes -> Strep (2/2) Accuracy = 75% Goodness = 75

No -> Allergy\* (1/2)

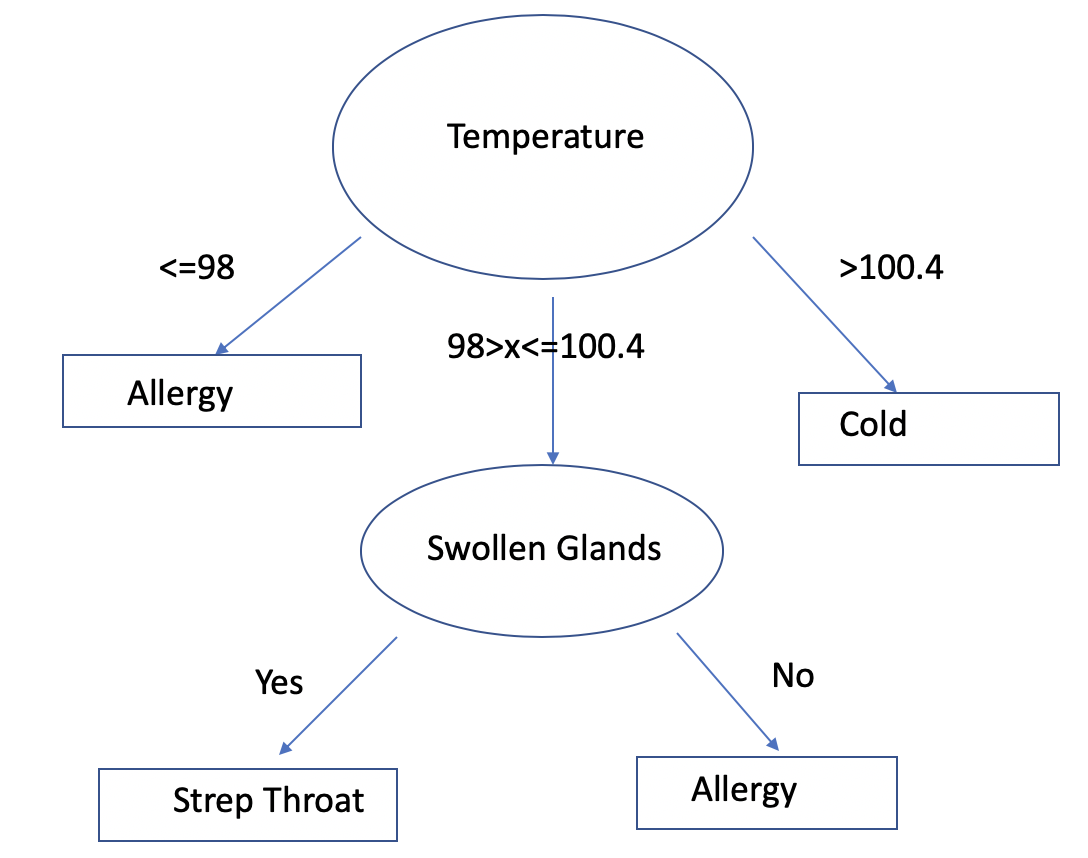
Swollen: Yes -> Strep (3/3) Accuracy = 100% Goodness = 0

No -> Allergy (1/1)

Congestion: Yes -> Allergy\* (1/2) Accuracy = 75% Goodness = 75

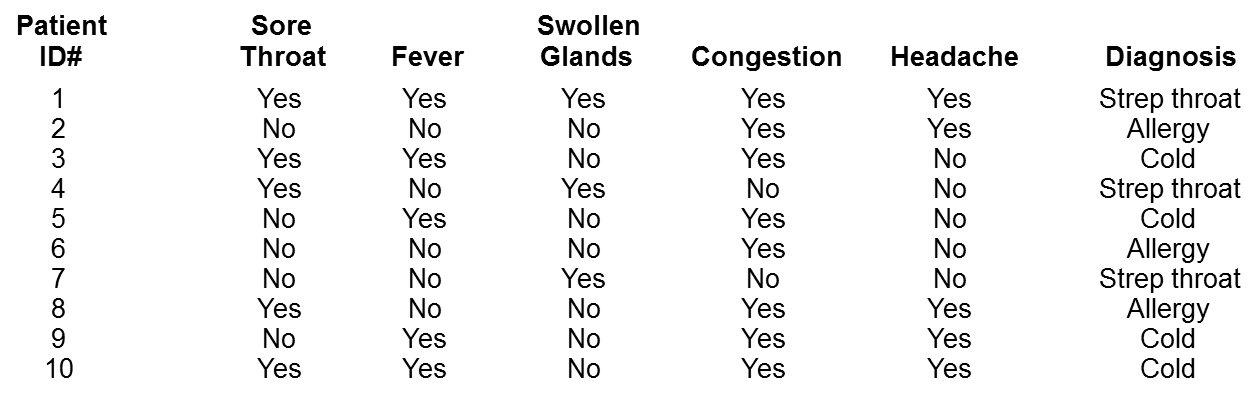
No -> Strep (2/2)

Headache: Yes -> Strep (1/1) Accuracy = 75% Goodness = 75

 No -> Strep (2/3)

***Generating Rule Sets for Classification***

Back to the all nominal dataset again:



4. Use the PRISM covering algorithm to generate as many rules as necessary for the class Cold. Show all of your work. Explain yourself when it would be helpful. You do not have to derive the rules for Strep throat or Allergy, though you should certainly keep those in mind for exam studying purposes.

If?

Then diagnosis = cold

Sore Throat = Yes 2/5

Sore Throat = No 2/5

Fever = Yes 4/5

Fever = No 0/5

Swollen Glands = Yes 0/3

Swollen Glands = No 4/7

Congestion = Yes 4/8

Congestion = No 0/2

Headache = Yes 2/5

Headache = No 2/5

Keep only applicable cases

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Patient ID | Sore Throat | Fever | Swollen | Congestion | Headache | Diagnosis |
| 1 | Yes | Yes | Yes | Yes | Yes | Strep |
| 3 | Yes | Yes | No | Yes | No | Cold |
| 5 | No | Yes | No | Yes | No | Cold |
| 9 | No | Yes | No | Yes | Yes | Cold |
| 10 | Yes | Yes | No | Yes | Yes | Cold |

Sore Throat = Yes 2/3

Sore Throat = No 2/2

Swollen Glands = Yes 0/1

Swollen Glands = No 4/4

Congestion = Yes 4/5

Congestion = No 0/0

Headache = Yes 2/3

Headache = No 2/2

If Fever = Yes and Swollen Glands = No then diagnosis = Cold