Homework 1

ISyE 6420

Spring 2022

Use of unsolicited electronic and printed resources is allowed except the communication that violates Georgia Tech Academic Integrity Rules (e.g., direct communication about solutions with a third party, use of HW-solving sites, and similar). [1]

1. Carpal Tunnel Syndrome Tests. Carpal tunnel syndrome is the most common entrapment neuropathy. The cause of this syndrome is hard to determine, but it can include trauma, repetitive maneuvers, certain diseases, and pregnancy.

Three commonly used tests for carpal tunnel syndrome are Tinel's sign, Phalen's test, and the nerve conduction velocity test. Tinel's sign and Phalen's test are both highly sensitive (0.97 and 0.92, respectively) and specific (0.91 and 0.88, respectively). The sensitivity and specificity of the nerve conduction velocity test are 0.93 and 0.87, respectively.[2]

Assume that the tests are conditionally independent.

Calculate the sensitivity and specificity of a combined test if combining is done

(a) in a serial manner;[3]

(b) in a parallel manner.[4]

(c) Find Positive Predictive Value (PPV) for tests in (a) and (b) if the prevalence of carpal tunnel syndrome in the general population is approximately 50 cases per 1000 subjects.

2. A Simple Naïve Bayes Classifier: 6420 Students going to Beach. Assume that for each instance covariates $x_1, \ldots, x_p$ are given and one of I different classes $\{1, 2, \ldots, I\}$ that the instance belongs to. Bayes classifier assigns the class according to maximum probability

$$\mathbb{P}(\text{Class } i|x_1, \ldots, x_p) \propto \mathbb{P}((x_1, \ldots, x_p)|\text{Class } i) \times \mathbb{P}(\text{Class } i),\ i = 1, \ldots, I,$$

conditionally that probabilities on the right hand side can be assessed/elicited. The symbol $\propto$ stands for the proportionality relation, exact probabilities $\mathbb{P}(\text{Class } i|x_1, \ldots, x_p)$ satisfy

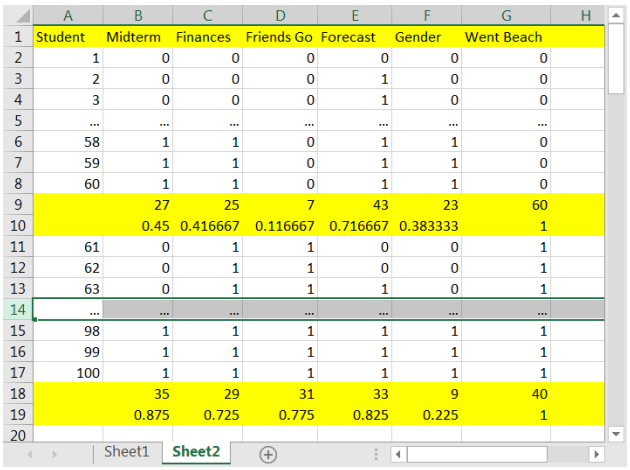
$$\sum_{i=1}^{I} \mathbb{P}(\text{Class } i|x_1, \ldots, x_p) \ = \ 1.$$

---

[2] For definitions of Sensitivity, Specificity, and PPV, consult Chapter 4: Sensitivity, Specificity, and Relatives, in the textbook at `http://statbook.gatech.edu`.

[3] Tests are combined in a serial manner if the combination is declared positive when all tests are positive.

[4] Tests are combined in a parallel manner if the combination is declared positive when at least one test is positive.



	A	B	C	D	E	F	G	H
1	Student	Midterm	Finances	Friends Go	Forecast	Gender	Went Beach	
2	1	0	0	0	0	0	0	
3	2	0	0	0	1	0	0	
4	3	0	0	0	1	0	0	
5	...	...	...	...	...	...	...	
6	58	1	1	0	1	1	0	
7	59	1	1	0	1	1	0	
8	60	1	1	0	1	1	0	
9		27	25	7	43	23	60	
10		0.45	0.416667	0.116667	0.716667	0.383333	1	
11	61	0	1	1	0	0	1	
12	62	0	1	1	0	0	1	
13	63	0	1	1	1	0	1	
14	...	...	...	...	...	...	...	
15	98	1	1	1	1	1	1	
16	99	1	1	1	1	1	1	
17	100	1	1	1	1	1	1	
18		35	29	31	33	9	40	
19		0.875	0.725	0.775	0.825	0.225	1	
20								

Sheet1 Sheet2

The probability $\mathbb{P}((x_1, \ldots, x_p)|\text{Class } i)$ could be quite difficult to assess due to possible interdependencies among the predictors. The naïve Bayes classifier assumes conditional independence, that is, given the Class i, all x_i are independent,

$$\begin{aligned}\mathbb{P}((x_1, \ldots, x_p)|\text{Class } i) &= \mathbb{P}(x_1|\text{Class } i) \times \mathbb{P}(x_2|\text{Class } i) \times \ldots \times \mathbb{P}(x_p|\text{Class } i) \\ &= \prod_{j=1}^{p} \mathbb{P}(x_j|\text{Class } i).\end{aligned}$$

The conditional probabilities $\mathbb{P}(x_j|\text{Class } i)$ are usually easier to assess. If we have a training sample, these probabilities can be taken as relative frequencies of items with covariate x_j among the items in the class i.

Thus, class i is selected for which

$$\mathbb{P}(\text{Class } i) \times \prod_{j=1}^{p} \mathbb{P}(x_j|\text{Class } i), \ i = 1, \ldots, I,$$

is maximum.

To illustrate very simple (covariates take values true/false, i.e., 1/0) naïve Bayes classifier, assume the following scenario:

We are interested in predicting whether a student from Bayesian class ISyE6420 will go to the beach (Class No (0)/Class Yes (1)) during Spring Break. To train our classifier, pretend that we have information from Spring 2021 enrollment in ISyE6420 and assume that students from the two enrolments are homogeneous, that is, the information from last year is relevant to this year students.

The imaginary data for 100 students are available, as in file `naive.csv|txt`. The following is recorded: Satisfied with ISyE6420 Midterm results (0 no/1 yes), Personal finances good (0 no/1 yes), Friends joined (0 no/1 yes), Weather forecast good (0 no/1 yes), Gender (0 male/1 female), Went to Beach (0 no/1 yes).

The conditional probabilities $\mathbb{P}(x_j|\text{Class } i)$ can be estimated by the relative frequencies of items with covariate x_j among the items in the class i (yellow in the image). For example,

Jane is happy with her 6420 Midterm results, financially is doing well, however, her friends will not go to the beach and the weather forecast does not look good. Then, for example $\mathbb{P}(\text{Financially doing well}|\text{Went to Beach}) = 29/40 = 0.725$, etc.

```
> pbpropto = 0.875*0.725*(1-0.775)*(1-0.825)*0.225 * 0.4 %0.002248066406250
> pnbpropto = 0.45 * 0.416667 * (1-0.116667) * (1-0.716667) * 0.383333* 0.6
                                                        %0.010793211644998
> pbeach = pbpropto/(pbpropto + pnbpropto) %0.172380835483743
> pnbeach = pnbpropto/(pbpropto + pnbpropto) %0.827619164516257
```

Thus, after normalizing the products, the naïve Bayes assigns the probability of 0.17238 of class 'Going to Beach' to Jane.[5]

Classify the following two students:

(a) Michael did poorly on his Midterm, he already owes some money, his friends will go to the beach, and weather forecast looks fine.

(b) Melissa did well on the Midterm, her finances look good, the weather prognosis looks good, but her friends will not go to the beach.

3. Multiple Choice Exam. A student answers a multiple choice examination with two questions that have four possible answers each. Suppose that the probability that the student knows the answer to a question is 0.80 and the probability that the student guesses is 0.20. If the student guesses, the probability of guessing the correct answer is 0.25. The questions are independent, that is, knowing the answer on one question is not influenced by the other question.

(a) What is the probability that the both questions will be answered correctly?

(b) If answered correctly, what is the probability that the student really knew the correct answer to both questions?

(c) How would you generalize the above from 2 to n questions, that is, what are answers to (a) and (b) if the test has n independent questions? What happens to probabilities in (a) and (b) if $n \to \infty$.

[5] What happens if any of the counts is 0, which may happen in the case of small sample size and large number of predictors/features? In this case the product is zero, leading to (unlikely) conclusion that the probability of the corresponding class label is 0. This is a case of overfitting, in statistical learning literature known as "Black Swan Paradox." In such cases one has to be Bayesian Naive Bayes, and smooth the counts by imposing priors on the class labels and binary predictors, usually Dirichlet and beta. But about priors, later.