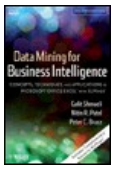

Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner, Second Edition

by Galit Shmueli, Nitin R. Patel and Peter C. Bruce
John Wiley & Sons (US). (c) 2010. Copying Prohibited.

Chapter 8: Naive Bayes

In this chapter we first present the complete, or exact, Bayesian classifier. We see that it can be used either to maximize overall classification accuracy or in the case where we are interested mainly in identifying records belonging to a particular class of interest. We next see how it is impractical in most cases and learn how to modify it (the "naive Bayesian classifier") so that it is generally applicable. The naive Bayesian classifier can be used only with categorical variables.

8.1 Introduction

The naive Bayes method (and, indeed, an entire branch of statistics) is named after the Reverend Thomas Bayes (1702-1761). To understand the naive Bayes classifier, we first look at the complete, or exact, Bayesian classifier. The basic principle is simple. For each record to be classified:

1. Find all the other records just like it (i.e., where the predictor values are the same).
2. Determine what classes they all belong to and which class is more prevalent.
3. Assign that class to the new record.

Alternatively (or in addition), it may be desirable to tweak the method so that it answers the question: What is an estimated probability of belonging to the class of interest? instead of Which class is the most probable? Obtaining class probabilities allows using a sliding cutoff to classify a record as belonging to class *i*, evenif *i* is not the most probable class for that record. This approach is useful when there is a specific class of interest that we are interested in identifying, and we are willing to "overidentify" records as belonging to this class. (See Chapter 5 for more details on the use of cutoffs for classification and on asymmetric misclassification costs)

Cutoff Probability Method

1. Establish a cutoff probability for the class of interest above which we consider that a record belongs to that class.
2. Find all the training records just like the new record (i.e., where the predictor values are the same).
3. Determine the probability that those records belong to the class of interest.
4. If that probability is above the cutoff probability, assign the new record to the class of interest.

Conditional Probability Both procedures incorporate the concept of *conditional probability*, or the probability of event *A* given that event *B* has occurred [denoted *P*(A|B)]. In this case, we will be looking at the probability of the record belonging to class *i* given that its predictor values take on the values x_1, x_2, …, *Xp*. In general, for a response with *m* classes C_1, C_2, …, C_m, and the predictors x_1, x_2, …, x_p, we want to compute

(8.1) $$P(C_i|x_1, \ldots, x_p).$$

To classify a record, we compute its probability of belonging to each of the classes in this way, then classify the record to the class that has the highest probability or use the cutoff probability to decide whether it should be assigned to the class of interest.

From this definition, we see that the Bayesian classifier works only with categorical predictors. If we use a set of numerical predictors, then it is highly unlikely that multiple records will have identical values on these numerical predictors. Therefore, numerical predictors must be binned and converted to categorical predictors. *The Bayesian classifier is the only classification or prediction method presented in this book that is especially suited for (and limited to) categorical predictor variables*. Consider Example 1.

Example 1: Predicting Fraudulent Financial Reporting

An accounting firm has many large companies as customers. Each customer submits an annual financial report to the firm, which is then audited by the accounting firm. For simplicity, we will designate the outcome of the audit as "fraudulent" or "truthful," referring to the accounting firm's assessment of the customer's financial report. The accounting firm has a strong incentive to be accurate in identifying fraudulent reports—if it passes a fraudulent report as truthful, it would be in legal trouble.

The accounting firm notes that, in addition to all the financial records, it also has information on whether or not the customer has had prior legal trouble (criminal or civil charges of any nature filed against it). This information has not been used in previous audits, but the accounting firm is wondering whether it could be used in the future to identify reports that merit more intensive review. Specifically, it wants to know whether having had prior legal trouble is predictive of fraudulent reporting.

In this case, each customer is a record, and the response of interest, *Y* = {fraudulent, truthful}, has two classes into which a company can be classified: C_1 = fraudulent and C_2 = truthful. The predictor variable—"prior legal trouble"—has two values: 0 (no prior legal trouble) and 1 (prior legal trouble).

The accounting firm has data on 1500 companies that it has investigated in the past. For each company, it has information on whether the financial report was judged fraudulent or truthful and whether the company had prior legal trouble. After partitioning the data into a training set (1000 firms) and a validation set (500 firms), the counts in the training set are shown in Table 8.1.

Table 8.1: PIVOT TABLE FOR FINANCIAL REPORTING EXAMPLE

	Prior Legal (X= 1)	No Prior Legal (X = 0)	Total
Fraudulent (C1)	50	50	100
Truthful (C_2)	180	720	900
Total	230	770	1000

8.2 Applying the Full (Exact) Bayesian Classifier

Now consider the financial report from a new company, which we wish to classify as fraudulent or truthful by using these data. To do this, we compute the probabilities, as above, of belonging to each of the two classes.

If the new company had had prior legal trouble, the probability of belonging to the fraudulent class would be *P*(fraudulent | prior legal) = 50/230 (there were 230 companies with prior legal trouble in the training set, and 50 of them had fraudulent financial reports). The probability of belonging to the other class, truthful, is, of course, the remainder = 180/230.

Using the "Assign to the Most Probable Class" Method If a company had prior legal trouble, we assign it to the "not-fraudulent" class. Similar calculations for the truthful case are left as an exercise to the reader. In this example, using the assign to the most probable class method, all records are assigned to the not-fraudulent class. This is the same result as the naive rule of "assign all records to the majority class."

Using the Cutoff Probability Method In this example, we are more interested in identifying the fraudulent reports—those are the ones that can land the auditor in jail. We recognize that, in order to identify the fraudulent reports, some truthful reports will be misidentified as fraudulent, and the overall classification accuracy may decline. Our approach is, therefore, to establish a cutoff value for the probability of being fraudulent, and classify all records above that value as fraudulent. The technical formula for the calculation of this probability that a record belongs to class C_1 is as follows:

(8.2) $$P(C_i|x_1,\ldots,x_p) = \frac{P(x_1,\ldots,x_p|C_i)P(C_i)}{P(x_1,\ldots,x_p|C_1)P(C_1)+\cdots+P(x_1,\ldots,x_p|C_m)P(C_m)}.$$

In this example (where frauds are more rare), if the cutoff were established at 0.20, we would classify a prior legal trouble record as fraudulent because *P*(fraudulent | prior legal) = 50/230 = 0.22. The user can treat this cutoff as a "slider" to be adjusted to optimize performance, like other parameters in any classification model.

Practical Difficulty with the Complete (Exact) Bayes Procedure

The approach outlined above amounts to finding all the records in the sample that are exactly like the new record to be classified in the sense that the predictor values are all the same. This was easy in the small examples presented above, where there was just one predictor.

When the number of predictors gets larger (even to a modest number like 20), many of the records to be classified will be without exact matches. This can be understood in the context of a model to predict voting on the basis of demographic variables. Even a sizable sample may not contain even a single match for a new record who is a male Hispanic with high

income from the U.S. Midwest who voted in the last election, did not vote in the prior election, has three daughters and one son, and is divorced. And this is just eight variables, a small number for most data mining exercises. The addition of just a single new variable with five equally frequent categories reduces the probability of a match by a factor of 5.

Solution: Naive Bayes

In the naive Bayes solution, we no longer restrict the probability calculation to those records that match the record to be classified. Instead we use the entire dataset.

Returning to our original basic classification procedure outlined at the beginning of the chapter, we recall that this procedure for classifying a new record was:

1. Find all the other records just like it (i.e., where the predictor values are the same).
2. Determine what classes they all belong to and which class is more prevalent.
3. Assign that class to the new record.

The naive Bayes modification (for the basic classification procedure) is as follows:

1. For class 1, find the individual probabilities that each predictor value in the record to be classified $(x_1, \ldots, x_p)$ occurs in class 1.
2. Multiply these probabilities times each other, then times the proportion of records belonging to class 1.
3. Repeat steps 1 and 2 for all the classes.
4. Estimate a probability for class *i* by taking the value calculated in step 2 for class *i* and dividing it by the sum of such values for all classes.
5. Assign the record to the class with the highest probability value for this set of predictor values.

The naive Bayes formula to calculate the probability that a record with a given set of predictor values $x_1, \ldots, x_p$ belongs to class 1 (Q) among *m* classes is as follows:

(8.3)
$$P_{nb}(C_1|x_1, \ldots x_p) = \frac{P(C_1)\left[P(x_1|C_1)P(x_2|C_1)\cdots P(x_p|C_1)\right]}{P(C_1)\left[P(x_1|C_1)P(x_2|C_1)\cdots P(x_p|C_1)\right] + \cdots + P(C_m)\left[P(x_1|C_m)P(x_2|C_m)\cdots P(x_p|C_m)\right]}.$$

This is a somewhat formidable formula; see Example 2 for a simpler numerical version. In probability terms, we have made a simplifying assumption that the exact *conditional probability* (the probability of belonging to a given class, given a set of predictor values) is well approximated by the product of the *unconditional probabilities* that those predictor values occur in the given class, overall, times the probability that a record belongs to that class, divided by the product of the *unconditional probabilities* that those predictor values occur across all classes. If predictor values are independent of one another, this approximation is the same as the exact value. In practice, the procedure works quite well—primarily because what is usually needed is not a probability value for each record that is accurate in absolute terms but just a reasonably accurate *rank ordering*.

Note that if all we are interested in is a rank ordering, and the denominator remains the same for all classes, it is sufficient to concentrate only on the numerator. The disadvantage of this approach is that the probability values it yields, while ordered correctly, are not on the same scale as the exact values that the user would anticipate.

The above procedure is for the basic case where we seek maximum classification accuracy for all classes. In the case of the relatively *rare class of special interest*, the procedure is:

1. Establish a cutoff probability for the class of interest above which we consider that a record belongs to that class.
2. For the class of interest, compute the probability that each predictor value in the record to be classified $(x_1, \ldots, x_p)$ occurs in the training data.
3. Multiply these probabilities times each other, then times the proportion of records belonging to the class of interest.

4. Estimate the probability for the class of interest by taking the value calculated in step 3 for the class of interest and dividing it by the sum of the similar values for all classes.
5. If this value falls above the cutoff, assign the new record to the class of interest, otherwise not.
6. Adjust the cutoff value as needed, as a parameter of the model.

Example 2: Predicting Fraudulent Financial Reports, Two Predictors

Let us expand the financial reports example to two predictors, and, using a small subset of data, compare the complete (exact) Bayes calculations to the naive Bayes calculations.

Consider the 10 customers of the accounting firm listed in Table 8.2. For each company, we have information on whether it had prior legal trouble, whether it is a small or large company, and whether the financial report was found to be fraudulent or truthful. Using this information, we will calculate the conditional probability of fraud, given each of the four possible combinations *{y*, small}, *{y*, large}, *{n*, small}, *{n*, large}.

Table 8.2: INFORMATION ON 10 COMPANIES

Prior Legal Trouble	Company Size	Status
Yes	Small	Truthful
No	Small	Truthful
No	Large	Truthful
No	Large	Truthful
No	Small	Truthful
No	Small	Truthful
Yes	Small	Fraudulent
Yes	Large	Fraudulent
No	Large	Fraudulent
Yes	Large	Fraudulent

Complete (Exact) Bayes Calculations The four probabilities are computed as:

P (fraudulent |PriorLegal = *y*, Size = small) = 1/2 = 0.5.

P (fraudulent |PriorLegal = *y*, Size = large) = 2/2 = 1.

P (fraudulent |PriorLegal = *n*, Size = small) = 0/3 = 0. *P*(fraudulent|PriorLegal = *n*, Size = large) = 1/3 = 0.33.

Naive Bayes Calculations Now we compute the naive Bayes probabilities. For the conditional probability of fraudulent behaviors given {Prior Legal = y, Size = small}, the numerator is a multiplication of the proportion of {Prior Legal = *y}* instances among the fraudulent companies, times the proportion of { Size = small} instances among the fraudulent companies, times the proportion of fraudulent companies: (3/4)(1/4)(4/10) = 0.075. However, to get the actual probabilities, we must also compute the numerator for the conditional probability of truth given {Prior Legal = y, Size = small}: (1/6)(4/6)(6/10) = 0.067. The denominator is then the sum of these two conditional probabilities (0.075 + 0.067 = 0.14). The conditional probability of fraudulent behaviors given {Prior Legal = *y*, Size = small} is therefore 0.075/0.14 = 0.53. In a similar fashion, we compute all four conditional probabilities:

$$P_{\text{nb}}(\text{fraudulent}|\text{PriorLegal} = y, \text{Size} = \text{small}) = \frac{(3/4)(1/4)(4/10)}{(3/4)(1/4)(4/10) + (1/6)(4/6)(6/10)} = 0.53.$$

$$P_{\text{nb}}(\text{fraudulent}|\text{PriorLegal} = y, \text{Size} = \text{large}) = 0.87.$$

$$P_{\text{nb}}(\text{fraudulent}|\text{PriorLegal} = n, \text{Size} = \text{small}) = 0.07.$$

$$P_{\text{nb}}(\text{fraudulent}|\text{PriorLegal} = n, \text{Size} = \text{large}) = 0.31.$$

Note how close these naive Bayes probabilities are to the exact Bayes probabilities. Although they are not equal, both would lead to exactly the same classification for a cutoff of 0.5 (and many other values). It is often the case that the rank ordering of probabilities is even closer to the exact Bayes method than are the probabilities themselves, and for classification purposes it is the rank orderings that matter.

We now consider a larger numerical example, where information on flights is used to predict flight delays.

Example 3: Predicting Delayed Flights

Predicting flight delays can be useful to a variety of organizations: airport authorities, airlines, and aviation authorities. At times, joint task forces have been formed to address the problem. Ifsuch an organization were to provide ongoing real-time assistance with flight delays, it would benefit from some advance notice about flights that are likely to be delayed.

In this simplified illustration, we look at six predictors (see Table 8.3). The outcome of interest is whether or not the flight is delayed *(delayed* means arrived more than 15 minutes late). Our data consist of all flights from the Washington, D.C., area into the New York City area during January 2004. The percentage of delayed flights among these 2346 flights is 18%. The data were obtained from the Bureau of Transportation Statistics (available on the Web at www.transtats.bts.gov). The goal is to accurately predict whether or not a new flight (not in this dataset), will be delayed.

Table 8.3: DESCRIPTION OF VARIABLES FOR FLIGHT DELAY EXAMPLE

Day of Week	Coded as: 1 = Monday, 2 = Tuesday, …, 7 = Sunday
Departure Time	Broken down into 18 intervals between 6:00 AM and 10:00 PM
Origin	Three airport codes: DCA (Reagan National), IAD (Dulles), BWI (Baltimore-Washington Int'l)
Destination	Three airport codes: JFK (Kennedy), LGA (LaGuardia), EWR (Newark)
Carrier	Eight airline codes: CO (Continental), DH (Atlantic Coast),
	DL (Delta), MQ (American Eagle), OH (Comair),
	RU (Continental Express), UA (United), and US (USAirways)
Weather	Coded as 1 if there was a weather-related delay

A record is a particular flight. The response is whether the flight was delayed, and thus it has two classes (1 = *Delayed* and 0 = *On time)*. In addition, information is collected on the predictors listed in Table 8.3.

The data were first partitioned into training and validation sets (with a 60%: 40% ratio), and then a naive Bayes classifier was applied to the training set.

The top table in Figure 8.1 shows the ratios of delayed flights and on-time flights in the training set (called Prior Class Probabilities). The bottom table shows the conditional probabilities for each class, as a function of the predictor values. Note that the conditional probabilities in the output can be computed simply by using pivot tables in Excel, looking at the percentage of records in a cell relative to the entire class. This is illustrated in

Table 8.4: PIVOT TABLE OF DELAYED AND ON-TIME FLIGHTS BY DESTINATION AIRPORT (ROWS)

	Delayed %	On Time %	Total %
EWR	38.67	28.36	30.36

JFK	18.75	17.65	17.87
LGA	42.58	53.99	51.78
Total	100.00	100.00	100.00

Prior Class Probabilities

According to relative occurrences in training data

Class	Prob.	
1	0.193792581	<-- Success Class
0	0.806207419	

Conditional Probabilities

	Classes-->			
Input Variables	1		0	
	Value	Prob	Value	Prob
CARRIER	CO	0.06640625	CO	0.038497653
	DH	0.33984375	DH	0.243192488
	DL	0.109375	DL	0.2
	MQ	0.1796875	MQ	0.112676056
	OH	0.01171875	OH	0.017840376
	RU	0.21484375	RU	0.170892019
	UA	0.0078125	UA	0.016901408
	US	0.0703125	US	0.2
DAY_OF_WEEK	1	0.203125	1	0.128638498
	2	0.16015625	2	0.139906103
	3	0.12890625	3	0.152112676
	4	0.12890625	4	0.159624413
	5	0.1640625	5	0.181220657
	6	0.0703125	6	0.131455399
	7	0.14453125	7	0.107042254
DEP_TIME_BLK	0600-0659	0.03515625	0600-0659	0.061971831
	0700-0759	0.05078125	0700-0759	0.060093897
	0800-0859	0.0546875	0800-0859	0.071361502
	0900-0959	0.0234375	0900-0959	0.053521127
	1000-1059	0.01953125	1000-1059	0.057276995
	1100-1159	0.01953125	1100-1159	0.038497653
	1200-1259	0.0546875	1200-1259	0.062910798
	1300-1359	0.05078125	1300-1359	0.068544601
	1400-1459	0.15234375	1400-1459	0.110798122
	1500-1559	0.08203125	1500-1559	0.064788732
	1600-1659	0.07421875	1600-1659	0.078873239
	1700-1759	0.15625	1700-1759	0.094835681
	1800-1859	0.03125	1800-1859	0.043192488
	1900-1959	0.08984375	1900-1959	0.040375587
	2000-2059	0.01953125	2000-2059	0.030985915
	2100-2159	0.0859375	2100-2159	0.061971831
DEST	EWR	0.38671875	EWR	0.283568075
	JFK	0.1875	JFK	0.176525822
	LGA	0.42578125	LGA	0.539906103
ORIGIN	BWI	0.09375	BWI	0.068544601
	DCA	0.484375	DCA	0.635680751
	IAD	0.421875	IAD	0.295774648
Weather	0	0.92578125	0	1
	1	0.07421875	1	0

FIGURE 8.1: OUTPUT FROM NAIVE BAYES CLASSIFIER APPLIED TO FLIGHT DELAYS (TRAINING) DATA

Table 8.2, which displays the percent of delayed (or on-time) flights by destination airport as a percentage of the total delayed (or on-time) flights.

Note that in this example there are no predictor values that were not represented in the training data except for on-time flights (Class = 0) when the weather was bad (Weather = 1). When the weather was bad, all flights in the training set were delayed.

To classify a new flight, we compute the probability that it will be delayed and the probability that it will be on time. Recall that since both will have the same denominator, we can just compare the numerators. Each numerator is computed by multiplying all the conditional probabilities of the relevant predictor values and, finally, multiplying by the proportion of that class [in this case P(delayed) = 0.19]. For example, to classify a Delta flight from DCA to LGA between 10 and 11 AM on a Sunday with good weather, we compute the numerators:

$\hat{P}$ (delayed | Carrier = DL, Day of Week = 7, DepartureTime = 1000 — 1059,

Destination = LGA, Origin = DCA, Weather = 0)

α(0.11)(0.14)(0.020)(0.43)(0.48)(0.93)(0.19) = 0.000011

$\hat{P}$ (ontime | Carrier = DL, Day of Week = 7, DepartureTime = 1000 — 1059,

Destination = LGA, Origin = DCA, Weather = 0)

α(0.2)(0.11)(0.06)(0.54)(0.64)(1)(0.81) = 0.00034

The symbol αmeans "is proportional to," reflecting the fact that this calculation deals only with the numerator in the naive Bayes formula (8.3).

It is, therefore, more likely that the flight will be on time. Note that a record with such a combination of predictors does not exist in the training set, and therefore we use the naive Bayes rather than the exact Bayes.

To compute the actual probability, we divide each of the numerators by their sum:

$$\hat{P}(\text{delayed} \mid \text{Carrier} = \text{DL}, \text{DayofWeek} = 7, \text{DepartureTime} = 1000 - 1059, \text{Destination} = \text{LGA}, \text{Origin} = \text{DCA}, \text{Weather} = 0) = \frac{0.000011}{0.000011 + 0.00034} = 0.03$$

$$\hat{P}(\text{ontime} \mid \text{Carrier} = \text{DL}, \text{DayofWeek} = 7, \text{DepartureTime} = 1000 - 1059, \text{Destination} = \text{LGA}, \text{Origin} = \text{DCA}, \text{Weather} = 0) = \frac{0.00034}{0.000011 + 0.00034} = 0.97$$

Of course, we rely on software to compute these probabilities for any records of interest (in the training set, the validation set, or for scoring new data). Figure 8.2 shows the estimated probabilities and classifications for a sample of flights in the validation set.

XLMiner : Naive Bayes - Classification of Validation Data

Data range	['Flight Delays.xls']'Data_Partition1'!C1340: H2219	Back to Navigator

Cut off Prob. Val. for Success (Updatable)	0.5	(Updating the value here will NOT update value in summary report)

RowId.	Predicted Class	Actual Class	Prob. for 1 (success)	CARRIERF	DAY_OF WEEK	DEP_ TI ME _BLK	DEST	ORIGIN	Weather
2	0	0	0.160552079	DH	4	1600- 1659	JF K	DCA	0
3	0	0	0.197147877	DH	4	1200- 1259	LGA	IAD	0
7	0	0	0.248536067	DH	4	1200- 1259	JF K	IAD	0
8	0	0	0.263631618	DH	4	1600- 1659	JF K	IAD	0
11	0	0	0.281467602	DH	4	2100- 2159	LGA	IAD	0
13	0	0	0.025209812	DL	4	0900- 0959	LGA	DCA	0
14	0	0	0.048830719	DL	4	1200- 1259	LGA	DCA	0
15	0	0	0.07510312	DL	4	1400- 1459	LGA	DCA	0
16	0	0	0.088673655	DL	4	1700- 1759	LGA	DCA	0
22	0	0	0.113149152	MQ	4	1300- 1359	LGA	DCA	0
24	0	0	0.179013723	MQ	4	1500- 1559	LGA	DCA	0
25	0	0	0.277045255	MQ	4	1900- 1959	LGA	DCA	0
28	0	0	0.018897189	US	4	1100- 1159	LGA	DCA	0
33	0	0	0.366861013	RU	4	1400- 1459	EW R	BWI	0
34	0	0	0.409792076	RU	4	1700- 1759	EW R	BWI	0
40	0	0	0.44593539	DH	4	1700- 1759	EW R	IAD	0
42	0	0	0.403842858	DH	4	2100- 2159	EW R	IAD	0
46	0	0	0.343153176	RU	4	1900- 1959	EW R	DCA	0
47	0	0	0.244033602	RU	4	1400- 1459	EW R	DCA	0
50	0	0	0.18094682	RU	4	1600- 1659	EW R	DCA	0
57	0	1	0.097462126	DH	5	1000- 1059	LGA	IAD	0

FIGURE 8.2: ESTIMATED PROBABILITY OF DELAY FOR A SAMPLE OFTHE VALIDATION SET

Finally, to evaluate the performance of the naive Bayes classifier for our data, we use the classification matrix, lift charts, and all the measures that were described in Chapter 5. For our example, the classification matrices for the training and validation sets are shown in Figure 8.3. We see that the overall error level is around 18% for both the training and validation data. In comparison, a naive rule that would classify all 880 flights in the validation set as on time would have missed the 172 delayed flights, resulting in a 20% error level. In other words, the naive Bayes is only slightly less accurate. However, examining the lift chart (Figure 8.4) shows the strength of the naive Bayes in capturing the delayed flights well.

Training Data scoring - Summary Report

Cut off Prob.Val. for Success (Updatable)	0.5

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	43	213
0	35	1030

Error Report			
Class	# Cases	# Errors	% Error
1	256	213	83.20
0	1065	35	3.29
Overall	1321	248	18.77

Validation Data scoring - Summary Report

Cut off Prob.Val. for Success (Updatable)	0.5

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	30	142
0	15	693

Error Report			
Class	# Cases	# Errors	% Error
1	172	142	82.56
0	708	15	2.12
Overall	880	157	17.84

FIGURE 8.3: CLASSIFICATION MATRICES FOR FLIGHT DELAYS USING A NAIVE BAYES CLASSIFIER

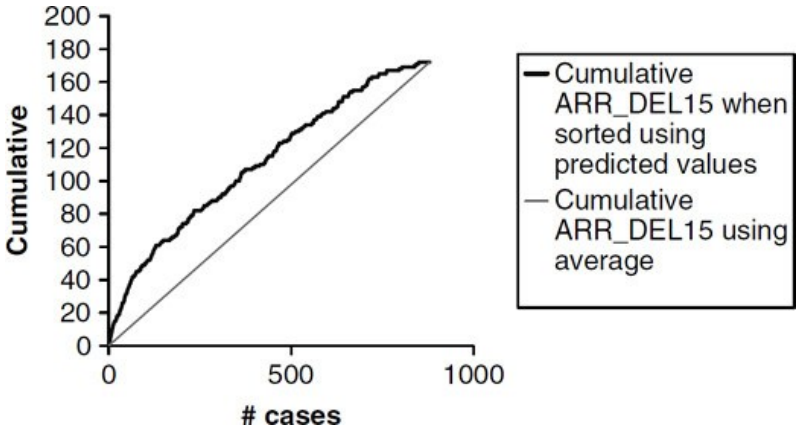

FIGURE 8.4: LIFT CHART OF NAIVE BAYES CLASSIFIER APPLIED TO FLIGHT DELAY DATA

8.3 Advantages and Shortcomings of the Naive Bayes Classifier

The naive Bayes classifier's beauty is in its simplicity, computational efficiency, and good classification performance. In fact, it often outperforms more sophisticated classifiers even when the underlying assumption of (conditionally) independent predictors is far from true. This advantage is especially pronounced when the number of predictors is very large.

Three issues should be kept in mind, however. First, the naive Bayes classifier requires a very large number of records to obtain good results. Second, where a predictor category is not present in the training data, naive Bayes assumes that a new record with that category of the predictor has zero probability. This can be a problem if this rare predictor value is important. For example, consider the target variable *bought high-value life insurance* and the predictor category *own yacht*. If the training data have no records with *owns yacht* = 1, for any new records where *owns yacht* = 1, naive Bayes will assign a probability of 0 to the target variable *bought high-value life insurance*. With no training records with *owns yacht* = 1, of course, no data mining technique will be able to incorporate this potentially important variable into the classification model—it will be ignored. With naive Bayes, however, the absence of this predictor actively "out votes" any other information in the record to assigna0tothetarget value (when, in this case, it has a relatively good chance of being a

1). The presence of a large training set (and judicious binning of continuous variables, ifrequired) helps mitigate this effect.

Finally, good performance is obtained when the goal is *classification* or *ranking* of records. However, when the goal is to estimate the *probability of class membership*, this method provides very biased results. For this reason the naive Bayes method is rarely used in credit scoring (Larsen, 2005).

SPAM FILTERING

Filtering spam is perhaps the most widely familiar application of data mining. Spam filtering, which is based in large part on natural language vocabulary, is a natural fit for a naive Bayesian classifier, which uses exclusively categorical variables. Most spam filters are based on this method, which works as follows:

1. Humans review a large number of e-mails, classify them as "spam" or "not spam," and from these select an equal (also large) number of spam e-mails and non-spam emails. This is the training data.
2. These e-mails will contain thousands of words; for each word compute the frequency with which it occurs in the spam class and the frequency with which it occurs in the non-spam class. Convert these frequencies into estimated probabilities (i.e., if the word "free" occurs in 500 out of 1000 spam e-mails, and only 100 out of 1000 non-spam e-mails, the probability that a spam email will contain the word "free" is 0.5, and the probability that a non-spam e-mail will contain the word "free" is 0.1).
3. If the only word in a new message that needs to be classified as spam or not spam is "free," we would classify the message as spam since the Bayesian posterior probability is 0.5/(0.5 + 01) or 5/6 that, given the appearance of "free," the message is spam.
4. Of course, we will have many more words to consider. For each such word, the probabilities described in step 2 are calculated and multiplied together, and formula (8.3) applied to determine the naive Bayes probability of belonging to the classes. In the simple version, class membership (spam or not spam) is determined by the higher probability.
5. In a more flexible interpretation, the "spam" probability is treated as a score for which the operator can establish (and change) a cutoff threshold—anything above that level is classified as spam.
6. Users have the option of building a personalized training database by classifying incoming messages as spam or non-spam, and adding them to the training database. One person's spam may be another person's substance.

It is clear that, even with the "naive" simplification, this is an enormous computational burden. Spam filters now typically operate at two levels—at servers (intercepting some spam that never makes it to your computer) and on individual computers (where you have the option of reviewing it). Spammers have also found ways to "poison" the vocabulary-based Bayesian approach, by including sequences of randomly selected irrelevant words. Since these words are randomly selected, they are unlikely to be systematically more prevalent in spam than in non-spam, and they dilute the effect of key spam terms such as "Viagra" and "free." For this reason, sophisticated spam classifiers also include variables based on elements otherthan vocabulary, such as the number of links in the message, the vocabulary in the subject line, determination of whether the "From:" e-mail address is the real originator (antispoofing), use of HTML and images, and origination at a dynamic or static IP address (the latter are more expensive and cannot be set up quickly).

Problems

8.1 **Personal Loan Acceptance**. The file UniversalBank.xls contains data on 5000 customers of Universal Bank. The data include customer demographic information (age, income, etc.), the customer's relationship with the bank (mortgage, securities account, etc.), and the customer response to the last personal loan campaign (Personal Loan). Among these 5000 customers, only 480 (= 9.6%) accepted the personal loan that was offered to them in the earlier campaign. In this exercise we focus on two predictors: Online (whether or not the customer is an active user of online banking services) and Credit Card (abbreviated CC below) (does the customer hold a credit card issued by the bank), and the outcome Personal Loan (abbreviated Loan below).

Partition the data into training (60%) and validation (40%) sets.

a. Create a pivot table for the training data with Online as a column variable, CC as a row variable, and Loan as a secondary row variable. The values inside the cells should convey the count (how many records are in that cell).

b. Consider the task of classifying a customer that owns a bank credit card and is actively using online banking services. Looking at the pivot table, what is the probability that this customer will accept the loan offer? [This is the probability of loan acceptance (Loan = 1) conditional on having a bank credit card (CC = 1) and being an active user of online banking services (Online = 1).]

c. Create two separate pivot tables for the training data. One will have Loan (rows) as a function of Online (columns) and the other will have Loan (rows) as a function of CC.

d. Compute the following quantities *[P (A|B)* means "the probability of *A* given B"]:
   i. *P*(CC = 1|Loan = 1) (the proportion of credit card holders among the loan
   ii. *P*(Online = 1 | Loan = 1)
   iii. *P*(Loan = 1) (the proportion of loan acceptors)
   iv. *P* (CC = 1|Loan = 0)
   v. *P* (Online = 1|Loan = 0)
   vi. *P*(Loan = 0)

e. Use the quantities computed above to compute the naive Bayes probability *P*(Loan = 1|CC = 1, Online = 1).

f. Compare this value with the one obtained from the crossed pivot table in (b). Which is a more accurate estimate?

g. In XLMiner, run naive Bayes on the data. Examine the "Conditional probabilities" table, and find the entry that corresponds to *P* (Loan = 1|CC = 1, Online = 1). Compare this to the number you obtained in (e).

8.2 **Automobile Accidents**. The file Accidents.xls contains information on 42, 183 actual automobile accidents in 2001 in the United States that involved one of three levels of injury: NO INJURY, INJURY, or FATALITY For each accident, additional information is recorded, such as day of week, weather conditions, and road type. A firm might be interested in developing a system for quickly classifying the severity of an accident based on initial reports and associated data in the system (some of which rely on GPS-assisted reporting).

Our goal here is to predict whether an accident just reported will involve an injury (MAX_SEV_IR = 1 or 2) or will not (MAX_SEV_IR = 0). For this purpose, create a dummy variable called INJURY that takes the value "yes" if MAX_SEV_IR = 1 or 2, and otherwise "no"

a. Using the information in this dataset, if an accident has just been reported and no further information is available, what should the prediction be? (INJURY = Yes or No?) Why?

b. Select the first 12 records in the dataset and look only at the response (INJURY) and the two predictors WEATHERR and TRAF_CON_R.
   i. Create a pivot table that examines INJURY as a function of the 2 predictors for these 12 records. Use all 3 variables in the pivot table as rows/columns, and use counts for the cells.
   ii. Compute the exact Bayes conditional probabilities of an injury (INJURY = Yes) given the six possible combinations of the predictors.
   iii. Classify the 12 accidents using these probabilities and a cutoff of 0.5.
   iv. Compute manually the naive Bayes conditional probability of an injury given WEATHERR = 1 and TRAF_CON_R = 1.
   v. Run a naive Bayes classifier on the 12 records and 2 predictors using XLMiner. Check *detailed report* to obtain probabilities and classifications for all 12 records. Compare this to the exact Bayes classification. Are the resulting classifications equivalent? Is the ranking (= ordering) of observations equivalent?

c. Let us now return to the entire dataset. Partition the data into training/validation sets (use XLMiner's "automatic" option for partitioning percentages).
   i. Assuming that no information or initial reports about the accident itselfare available at the time of prediction (only location characteristics, weather conditions, etc.), which predictors can we include in the analysis? (Use the Data_Codes sheet.)
   ii. Run a naive Bayes classifier on the complete training set with the relevant predictors (and INJURY as the

response). Note that all predictors are categorical. Show the classification matrix.

iii. What is the overall error for the validation set?

iv. What is the percent improvement relative to the naive rule (using the validation set)?

v. Examine the conditional probabilities output. Why do we get a probability of zero for P(INJURY = No | SPDLIM = 5)?