A simple application of

Bayes' formula

CMU, 2006 fall, Tom Mitchell, Eric Xing, midterm, pr. 1.3

Suppose that in answering a question in a multiple choice test, an *examinee* either knows the answer, with probability p, or he guesses with probability $1-p$.

Assume that the probability of answering a *question* correctly is 1 for an examinee who *knows* the answer and $1/m$ for the examinee who *guesses*, where m is the number of *multiple choice alternatives*.

What is the probability that an examinee knew the answer to [such] a question, given that he has correctly answered it?

Answer:

$$\begin{aligned} P(correct|knew) &= \frac{P(knew|correct)\cdot P(knew)}{P(knew|correct)} \\ &= \frac{P(correct|knew)\cdot P(knew)}{P(correct|knew)\cdot P(knew)+P(correct|guessed)\cdot P(guessed)}. \end{aligned}$$

Notice that in the denominator we had to consider the two ways (s)he can get a question correct: by knowing, or by guessing. Plugging in, we get:

$$\frac{p}{p+\frac{1}{m}\cdot(1-p)} = \frac{mp}{mp+1-p}.$$

Baysian Classification

Some exercises

Exemplifying

- the notion of MAP (Maximum A posteriori Probability) hypotheses
- the computation of *expected values* for random variables

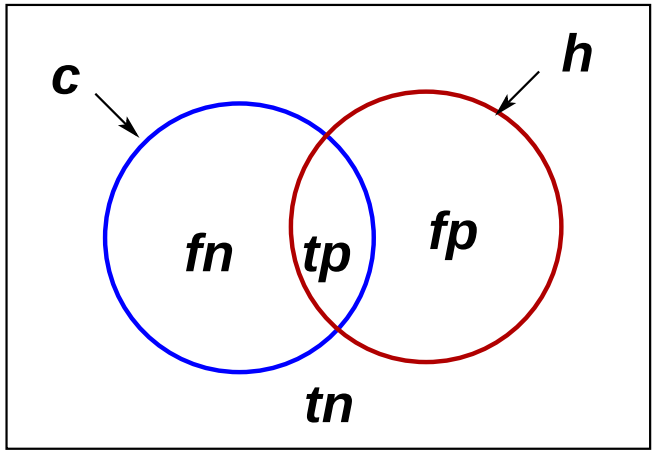
and

- the [use of] *sensitivity* and *specificity* of a test in a real-world application

CMU, 2009 fall, Geoff Gordon, HW1, pr. 2

There is a disease which affects 1 in 500 people. A 100.00 dollar blood test can help reveal whether a person has the disease. A positive outcome indicates that the person may have the disease.

The test has perfect sensitivity (*true positive rate*), i.e., a person who has the disease tests positive 100% of the time. However, the test has 99% specificity (*true negative rate*), i.e., a healthy person tests positive 1% of the time.



sensitivity (or: recall): $\frac{tp}{tp + fn}$

specificity: $\frac{tn}{tn + fp}$

a. A randomly selected individual is tested and the result is positive.

What is the *probability* of the individual having the disease?

b. There is a second more expensive test which costs 10, 000.00 dollars but is exact with 100% *sensitivity* and *specificity*.

If we require all people who test positive with the less expensive test to be tested with the more expensive test, what is the *expected cost* to check whether an individual has the disease?

c. A pharmaceutical company is attempting to decrease the cost of the second (perfect) test.

How much would it have to make the second test cost, so that the first test is no longer needed? That is, at what cost is it cheaper simply to use the perfect test alone, instead of screening with the cheaper test as described in part b?

Answer:

Let's define the following *random variables*:

B: $\begin{cases} \textbf{1/true} & \textbf{for persons affected by that disease,} \\ \textbf{0/false} & \textbf{otherwise;} \end{cases}$

T_1: **the result of the first test: $+$ (in case of disease) or $-$ (otherwise);**

T_2: **the result of the second test: again $+$ or $-$.**

***Known facts*:**

$$P(B) = \frac{1}{500}$$

$$P(T_1 = + \mid B) = 1,\ P(T_1 = + \mid \bar{B}) = \frac{1}{100},$$

$$P(T_2 = + \mid B) = 1,\ P(T_2 = + \mid \bar{B}) = 0$$

a.

$$P(B \mid T_1 = +) \overset{TBayes}{=} \frac{P(T_1 = + \mid B) \cdot P(B)}{P(T_1 = + \mid B) \cdot P(B) + P(T_1 = + \mid \bar{B}) \cdot P(\bar{B})}$$

$$= \frac{1 \cdot \frac{1}{500}}{1 \cdot \frac{1}{500} + \frac{1}{100} \cdot \frac{499}{500}} = \frac{100}{599} \approx 0.1669$$

b.

Let's consider a new random variable:

$$C = \begin{cases} c_1 & \textbf{if the person only takes the first test} \\ c_1 + c_2 & \textbf{if the person takes the two tests} \end{cases}$$

$$\Rightarrow P(C = c_1) = P(T_1 = -) \textbf{ and } P(C = c_1 + c_2) = P(T_1 = +)$$

$$\begin{aligned} \Rightarrow E[C] &= c_1 \cdot (1 - P(T_1 = +)) + (c_1 + c_2) \cdot P(T_1 = +) \\ &= c_1 - c_1 \cdot P(T_1 = +) + c_1 \cdot P(T_1 = +) + c_2 \cdot P(T_1 = +) \\ &= c_1 + c_2 \cdot P(T_1 = +) \\ &= 100 + 10000 \cdot \frac{599}{50000} = 219.8 \approx 220\$ \end{aligned}$$

Note: Here above we used

$$\begin{aligned} P(T_1 = +) &\overset{\textit{total probability form.}}{=} P(T_1 = + \mid B) \cdot P(B) + P(T_1 = + \mid \bar{B}) \cdot P(\bar{B}) \\ &= 1 \cdot \frac{1}{500} + \frac{1}{100} \cdot \frac{499}{500} = \frac{599}{50000} = 0.01198 \end{aligned}$$

c.

$c_n \overset{not.}{=}$ **the new price for the second test** (T_2')

$$\begin{aligned} c_n \leq E[C'] &= c_1 \cdot P(C = c_1) + (c_1 + c_n) \cdot P(C = c_1 + c_n) \\ &= c_1 + c_n \cdot P(T_1 = +) = 100 + c_n \cdot \frac{599}{50000} \end{aligned}$$

$c_n = 100 + c_n \cdot 0.01198 \Rightarrow c_n \approx 101.2125$**.**

Exemplyifying the application of

Naive Bayes and Joint Bayes;

the number of [independent] parameters

to be computed for NB and respectively JB

CMU, 2008 fall, Eric Xing, HW1 pr. 2

Consider a classification problem, the table of observations for which is given nearby. X_1 and X_2 are two binary random variables which are the observed variables. Y is the class label which is observed for the training data given below. We will use the Naive Bayes classifier and the Joint Bayes classifier to classify a new instance after training on the data below, and compare the results.

X_1	X_2	Y	Counts
0	0	0	2
0	0	1	18
1	0	0	4
1	0	1	1
0	1	0	4
0	1	1	1
1	1	0	2
1	1	1	18

a. Construct the Naive Bayes classifier given the data above. Use it to classify the instance $X_1 = 0,\ X_2 = 0$.

b. Construct the Joint Bayes classifier given the data above. Use it to classify the instance $X_1 = 0,\ X_2 = 0$.

c. Compare the number of independent parameters in the two classifiers. Instead of just two observed data variables, if there were n random binary observed variables $\{X_1, \ldots, X_n\}$, what would be the number of parameters required for both classifiers? From this, what can you comment about the rate of growth of the number of parameters for both models as $n \to \infty$?

d. Compute the probabilities $P(Y = 1|X_1 = 0, X_2 = 0)$ for the Naive Bayes classifier $P_{NB}(Y = 1|X_1 = 0, X_2 = 0)$ and for the Joint Bayes classifier $P_{JB}(Y = 1|X_1 = 0, X_2 = 0)$.

e. Why is $P_{NB}(Y = 1|X_1 = 0, X_2 = 0)$ different from $P_{JB}(Y = 1|X_1 = 0, X_2 = 0)$? (*Hint*: Compute $P(X_1, X_2|Y)$.)

f. What happens to the difference between $P_{NB}(Y = 1|X_1 = 0, X_2 = 0)$ and $P_{JB}(Y = 1|X_1 = 0, X_2 = 0)$ if the table entries are changed to the nearby table?

(*Hint*: Will the Naive Bayes assumption be more violated or less violated compared to the previous situation?)

X_1	X_2	Y	Counts
0	0	0	3
0	0	1	9
1	0	0	3
1	0	1	9
0	1	0	3
0	1	1	9
1	1	0	3
1	1	1	9

Naive Bayes and Joint Bayes:
application when a probabilistic distribution
(on input + output variables) is provided

CMU, 2010 spring, T. Mitchell, E. Xing, A. Singh, midterm pr. 2.1

Consider the joint probability distribution over 3 boolean random variables x_1, x_2, and y given in the nearby figure.

x_1	x_2	y	$P(x_1, x_2, y)$
0	0	0	0.15
0	0	1	0.25
0	1	0	0.05
0	1	1	0.08
1	0	0	0.10
1	0	1	0.02
1	1	0	0.20
1	1	1	0.15

a. Express $P(y = 0 \mid x_1, x_2)$ in terms of $P(x_1, x_2, y = 0)$ and $P(x_1, x_2, y = 1)$.

b. Complute the marginal probabilities which will be used by a Naive Bayes classifier. Fill in the following tables.

$P(x_1 \mid y)$	$x_1 = 0$	$x_1 = 1$
$y = 0$		
$y = 1$		

$P(x_2 \mid y)$	$x_2 = 0$	$x_2 = 1$
$y = 0$		
$y = 1$		

y	$P(y)$
$y = 0$	
$y = 1$	

c. Write out an expression for the value of $P(y = 1|x_1 = 1, x_2 = 0)$ predicted by the Naive Bayes classifier.

d. Write out an expression for the value of $P(y = 1|x_1 = 1, x_2 = 0)$ predicted by the Joint Bayes classifier.

e. The expressions you wrote down for parts (c) and (d) should be unequal. Explain why.

Exemplifying

Text classification using the Naive Bayes algorithm

CMU, 2009 spring, Ziv Bar-Joseph, midterm, pr. 2

About 2/3 of your email is spam, so you downloaded an open source spam filter based on word occurrences that uses the Naive Bayes classifier.
Assume you collected the following regular and spam mails to train the classifier, and only three words are informative for this classification, i.e., each email is represented as a 3-dimensional binary vector whose components indicate whether the respective word is contained in the email.

'study'	'free'	'money'	Category	*count*
1	0	0	**Regular**	1
0	0	1	**Regular**	1
1	0	0	**Regular**	1
1	1	0	**Regular**	1
0	1	0	**Spam**	4
0	1	1	**Spam**	4

a. You find that the spam filter uses a prior $P(\text{spam}) = 0.1$. Explain (in one sentence) why this might be sensible.

b. Compute the Naive Bayes parameters, using Maximum Likelihood Estimation (MLE) and applying Laplace's rule ("add-one").

c. Based on the prior and conditional probabilities above, give the model probability $P(\text{spam}|\ s)$ that the sentence

$$s = \text{"money for psychology study"}$$

is spam.

Answer:

a. It is worse for regular emails to be classified as spam than it is for spam email to be classified as regular email.

b. Estimating the Naive Bayes parameters, by MLE and applying Laplace's rule ("add-one"):

$$P(\textbf{study}|\textbf{spam}) = \frac{0+1}{8+2} = \frac{1}{10} \qquad P(\textbf{study}|\textbf{regular}) = \frac{3+1}{4+2} = \frac{2}{3}$$

$$P(\textbf{free}|\textbf{spam}) = \frac{8+1}{8+2} = \frac{9}{10} \qquad P(\textbf{free}|\textbf{regular}) = \frac{1+1}{4+2} = \frac{1}{3}$$

$$P(\textbf{money}|\textbf{spam}) = \frac{4+1}{8+2} = \frac{1}{2} \qquad P(\textbf{money}|\textbf{regular}) = \frac{1+1}{4+2} = \frac{1}{3}$$

c. Classification of the message

$$s = \text{“money for psychology study”},$$

using the a priori probability $P(\text{spam}) = 0.1$:

$$P(\text{spam} \mid s) = P(\text{spam} \mid \text{study}, \neg\text{free}, \text{money})$$

$$\overset{\text{F. Bayes}}{=} \frac{P(\text{study, } \neg\text{free, money} \mid \text{spam}) \cdot P(\text{spam})}{P(\text{study,}\neg\text{free,money} \mid \text{spam})P(\text{spam}) + P(\text{study,}\neg\text{free,money} \mid \text{reg})P(\text{reg})}$$

$$P(\text{study, } \neg\text{free, money}|\text{spam}) \overset{\text{indep. cdt.}}{=} P(\text{study}|\text{spam}) \cdot P(\neg\text{free}|\text{spam}) \cdot P(\text{money}|\text{spam})$$

$$= \frac{1}{10} \cdot \frac{1}{10} \cdot \frac{1}{2} = \frac{1}{200}$$

$$P(\text{study, } \neg\text{free, money}|\text{reg}) \overset{\text{indep. cdt.}}{=} P(\text{study}|\text{reg}) \cdot P(\neg\text{free}|\text{reg}) \cdot P(\text{money}|\text{reg})$$

$$= \frac{2}{3} \cdot \frac{2}{3} \cdot \frac{1}{3} = \frac{4}{27}$$

Therefore,

$$P(\text{spam}|\ s) = \frac{\frac{1}{200} \cdot \frac{1}{10}}{\frac{1}{200} \cdot \frac{1}{10} + \frac{4}{27} \cdot \frac{9}{10}} \approx 0.0037$$

Exemplifying

The computation of the *error rate* for the Naive Bayes algorithm

CMU, 2010 fall, Aarti Singh, HW1, pr. 4.2

Consider a simple learning problem of determining whether Alice and Bob from CA will go to hiking or not $Y : Hike \in \{T, F\}$ given the weather conditions $X_1 : Sunny \in \{T, F\}$ and $X_2 : Windy \in \{T, F\}$ by a Naive Bayes classifier.

Using training data, we estimated the parameters

$$P(Hike) = 0.5$$
$$P(Sunny \mid Hike) = 0.8, \quad P(Sunny \mid \neg Hike) = 0.7$$
$$P(Windy \mid Hike) = 0.4, \quad P(Windy \mid \neg Hike) = 0.5$$

Assume that the true distribution of X_1, X_2, and Y satisfies the Naive Bayes assumption of conditional independence with the above parameters.

a. What is the joint probability that Alice and Bob go to hiking and the weather is sunny and windy, that is $P(Sunny, Windy, Hike)$?

Solution:

$$P(Sunny, Windy, Hike) \overset{cdt.\ indep.}{=} P(Sunny|Hike) \cdot P(Windy|Hike) \cdot P(Hike) = 0.8 \cdot 0.4 \cdot 0.5 = 0.16.$$

b. What is the expected error rate of the Naive Bayes classifier? (Informally, the expected error rate is the probability that an "observation"/instance randomly generated according to the *true* probabilistic distribution of data is incorrectly classified by the Naive Bayes algorithm.)

Solution:

X_1	X_2	Y	$P(X_1, X_2, Y) = P(X_1\|Y) \cdot P(X_2\|Y) \cdot P(Y)$	$Y_{NB}(X_1, X_2)$	$P_{NB}(Y\|X_1, X_2)$
F	F	F	$0.3 \cdot 0.5 \cdot 0.5 = 0.075$	F	0.555556
F	F	T	$0.2 \cdot 0.6 \cdot 0.5 = \mathbf{0.060}$	F	0.444444
F	T	F	$0.3 \cdot 0.5 \cdot 0.5 = 0.075$	F	0.652174
F	T	T	$0.2 \cdot 0.4 \cdot 0.5 = \mathbf{0.040}$	F	0.347826
T	F	F	$0.7 \cdot 0.5 \cdot 0.5 = \mathbf{0.175}$	T	0.421686
T	F	T	$0.8 \cdot 0.6 \cdot 0.5 = 0.240$	T	0.578314
T	T	F	$0.7 \cdot 0.5 \cdot 0.5 = 0.175$	F	0.522388
T	T	T	$0.8 \cdot 0.4 \cdot 0.5 = \mathbf{0.160}$	F	0.477612

***Note*:**
Joint probabilities corresponding to incorrect predictions are shown in bold.

$$
\begin{aligned}
error &\overset{def.}{=} E_P[I_{Y_{NB}(X_1,X_2)\neq Y}] \\
&= \sum_{X_1,X_2,Y} I[Y_{NB}(X_1, X_2) \neq Y] \cdot P(X_1, X_2, Y) \\
&= \mathbf{0.060 + 0.040 + 0.175 + 0.160 = 0.435}
\end{aligned}
$$

***Note*:**
I is the *indicator* function; its value is 1 whenever the associated condition (in our case, $Y_{NB}(X_1, X_2) \neq Y$) is true, and 0 otherwise.

Next, suppose that we gather more information about weather conditions and introduce a new feature denoting whether the weather is X_3 : *Rainy* or not. *Assume* that each day the weather in CA can be either *Rainy* or *Sunny*. That is, it can not be both *Sunny* and *Rainy*. (Similarly, it can not be $\neg$ *Sunny* and $\neg$*Rainy*).

c. In the above new case, are any of the Naive Bayes assumptions violated? Why (not)? What is the joint probability that Alice and Bob go to hiking and the weather is sunny, windy and not rainy, that is $P(Sunny, Windy, \neg Rainy, Hike)$?

Solution:

The conditional independence of variables given the class label assumption of Naive Bayes is violated. Indeed, knowing if the weather is *Sunny* completely determines whether it is *Rainy* or not. Therefore, *Sunny* and *Rainy* are clearly NOT conditionally independent given *Hike*.

$$
\begin{aligned}
& P(Sunny, Windy, \neg Rainy, Hike) \\
&= \underbrace{P(\neg Rainy | Hike, Sunny, Windy)}_{1} \cdot P(Sunny, Windy | Hike) \cdot P(Hike) \\
&\overset{cond.\ indep.}{=} P(Sunny | Hike) \cdot P(Windy | Hike) \cdot P(Hike) \\
&= 0.8 \cdot 0.4 \cdot 0.5 = 0.16.
\end{aligned}
$$

d. What is the expected error rate when the Naive Bayes classifier uses all three attributes? Does the performance of Naive Bayes improve by observing the new attribute Rainy? Explain why.

Solution:

X_1	X_2	X_3	Y	$P(X_1, X_2, Y)$	$P_{NB}(X_1, X_2, X_3, Y) = P(X_3\|Y) \cdot P(X_1\|Y) \cdot P(X_2\|Y) \cdot P(Y)$	$Y_{NB}(X_1, X_2, X_3)$	$P_{NB}(Y\|X_1, X_2, X_3)$
F	F	F	F	0	$0.075 \cdot 0.7 = 0.0525$	F	0.522388
F	F	F	T	**0**	$0.060 \cdot 0.8 = 0.0480$	F	0.477612
F	F	T	F	0.075	$0.075 \cdot 0.3 = 0.0225$	F	0.652174
F	F	T	T	**0.060**	$0.060 \cdot 0.2 = 0.0120$	F	0.347826
F	T	F	F	0	$0.075 \cdot 0.7 = 0.0525$	F	0.621302
F	T	F	T	**0**	$0.040 \cdot 0.8 = 0.0320$	F	0.378698
F	T	T	F	0.075	$0.075 \cdot 0.3 = 0.0225$	F	0.737705
F	T	T	T	**0.040**	$0.040 \cdot 0.2 = 0.0080$	F	0.262295
T	F	F	F	**0.175**	$0.175 \cdot 0.7 = 0.0525$	T	0.389507
T	F	F	T	0.240	$0.240 \cdot 0.8 = 0.1920$	T	0.610493
T	F	T	F	0	$0.175 \cdot 0.3 = 0.0525$	F	0.522388
T	F	T	T	**0**	$0.240 \cdot 0.2 = 0.0480$	F	0.477612
T	T	F	F	**0.175**	$0.175 \cdot 0.7 = 0.0525$	T	0.489022
T	T	F	T	0.160	$0.160 \cdot 0.8 = 0.1280$	T	0.510978
T	T	T	F	0	$0.175 \cdot 0.3 = 0.0225$	F	0.621302
T	T	T	T	**0**	$0.060 \cdot 0.2 = 0.0120$	F	0.378698

The new error rate is:
$0.060 + 0.040 + 0.175 + 0.175 = 0.45 > 0.435$ (see question b).

The Naive Bayes classifier performance drops because the conditional independence assumptions do not hold for the correlated features.

How bad/naive is Naive Bayes?

CMU, 2010 spring, E. Xing, T. Mitchell, A. Singh, midterm, pr. 2.1

Clearly Naive Bayes makes what, in many cases, are overly strong assumptions. But even if those assumptions aren't true, is it possible that Naive Bayes is still pretty good? *Here we will use a simple example to explore the limitations of Naive Bayes.*

Let X_1 and X_2 be i.i.d. Bernoulli(0.5) random variables, and let $Y \in \{1, 2\}$ be some deterministic function of the values of X_1 and X_2.

a. Find a function Y for which the Naive Bayes classifier has a 50% error rate. Given the value of Y, how are X_1 and X_2 correlated?

b. Show that for every function Y, the Naive Bayes classifier will perform no worse than theone above. *Hint*: there are many Y functions, but because of symmetries in the problem you only need to analyze a few of them.

Răspuns

a. Considerăm Y definit conform tabelului alăturat.

X_1	X_2	Y
0	0	1
0	1	2
1	0	2
1	1	1

***Observaţie*: Dacă se consideră valoarea lui Y fixată (fie 1, fie 2), atunci putem să stabilim o regulă astfel încât dacă îl cunoaştem pe X_1 să-l determinăm pe X_2 (şi invers).[a] Altfel spus, X_1 este unic determinat de X_2 (şi invers) dată fiind o valoare fixată a lui Y. Deci condiţia de independenţă condiţională este încălcată. Mai mult, în acest caz avem maximul posibil de „dependenţă" între cele două variabile (în raport cu Y).**

Pe slide-ul următor vom calcula rata erorii înregistrate de algoritmul Bayes Naiv pe datele din tabelul de mai sus.

[a] Pentru $Y = 1$, regula este: X_2 are aceeaşi valoare ca şi X_1. Pentru $Y = 2$, regula este: X_1 şi X_2 au valori complementare.

Bayes Naiv estimează valoarea lui Y astfel:

$$\hat{y} = \underset{y \in \{1,2\}}{\operatorname{argmax}} P(X_1 \mid Y = y) P(X_2 \mid Y = y) P(Y = y)$$

Pentru $X_1 = 0, X_2 = 0$, algoritmul compară următoarele două valori:

$$\begin{aligned} p_1 &= P(X_1 = 0 \mid Y = 1) P(X_2 = 0 \mid Y = 1) P(Y = 1) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8} \\ p_2 &= P(X_1 = 0 \mid Y = 2) P(X_2 = 0 \mid Y = 2) P(Y = 2) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8} \end{aligned}$$

Cum $p_1 = p_2$, algoritmul va alege una dintre ele cu o probabilitate de 0.5. Deoarece valoarea lui Y este 1 din tabel, înseamnă că algoritmul va alege greşit în 50% din cazuri.

Pentru celelalte 3 cazuri, $(X_1 = 0, X_2 = 1)$, $(X_1 = 1, X_2 = 0)$ şi $(X_1 = 1, X_2 = 1)$, se observă uşor că se obţin de asemenea valori egale, iar algoritmul va alege pentru Y una dintre valorile 1 sau 2 cu o probabilitate de 0.5.

Deci pentru această definiţie a lui Y rata erorii este de 50%.

b. Vom calcula rata erorii pentru fiecare dintre cele 3 moduri de definire a lui Y care nu a fost studiat.

***Cazul 1*:**

X_1	X_2	Y
0	0	1
0	1	1
1	0	1
1	1	1

Este similar cu cazul:

Y
2
2
2
2

- **Pentru $X_1 = 0, X_2 = 0$, algoritmul compară:**

$$\begin{aligned} p_1 &= P(X_1 = 0 \mid Y = 1)P(X_2 = 0 \mid Y = 1)P(Y = 1) = \frac{2}{4} \cdot \frac{2}{4} \cdot 1 = \frac{1}{4} \\ p_2 &= P(X_1 = 0 \mid Y = 2)P(X_2 = 0 \mid Y = 2)P(Y = 2) = 0 \cdot 0 \cdot 0 = 0 \end{aligned}$$

  Cum $p_1 > p_2$ algoritmul alege pentru Y valoarea 1, ceea ce este corect.

- **Pentru celelalte 3 cazuri, $(X_1 = 0, X_2 = 1)$, $(X_1 = 1, X_2 = 0)$ şi $(X_1 = 1, X_2 = 1)$, se observă că se obţin aceleaşi valori pentru p_1 şi p_2 ca mai sus, deci algoritmul alege (în mod corect) pentru Y valoarea 1.**

Aşadar, am obţinut că rata erorii este în acest caz 0.

***Cazul 2*:**

X_1	X_2	Y
0	0	1
0	1	1
1	0	1
1	1	2

Cazuri similare:

Y	Y	Y	Y	Y	Y	Y
1	1	2	2	2	2	1
1	2	1	2	2	1	2
2	1	1	2	1	2	2
1	1	1	1	2	2	2

- **Pentru** $X_1 = 0, X_2 = 0$**:**

$$\left.\begin{aligned} p_1 &= P(X_1 = 0 \mid Y = 1)P(X_2 = 0 \mid Y = 1)P(Y = 1) = \frac{2}{3} \cdot \frac{2}{3} \cdot \frac{3}{4} = \frac{1}{3} \\ p_2 &= P(X_1 = 0 \mid Y = 2)P(X_2 = 0 \mid Y = 2)P(Y = 2) = 0 \cdot 0 \cdot \frac{1}{4} = 0 \end{aligned}\right\} \Rightarrow \hat{y} = 1$$

- **Pentru** $X_1 = 0, X_2 = 1$**:**

$$\left.\begin{aligned} p_1 &= P(X_1 = 0 \mid Y = 1)P(X_2 = 1 \mid Y = 1)P(Y = 1) = \frac{2}{3} \cdot \frac{1}{3} \cdot \frac{3}{4} = \frac{1}{6} \\ p_2 &= P(X_1 = 0 \mid Y = 2)P(X_2 = 1 \mid Y = 2)P(Y = 2) = 0 \cdot 1 \cdot \frac{1}{4} = 0 \end{aligned}\right\} \Rightarrow \hat{y} = 1$$

- **Pentru** $X_1 = 1, X_2 = 0$**:**

$$\left.\begin{aligned} p_1 &= P(X_1 = 1 \mid Y = 1)P(X_2 = 0 \mid Y = 1)P(Y = 1) = \frac{1}{3} \cdot \frac{2}{3} \cdot \frac{3}{4} = \frac{1}{6} \\ p_2 &= P(X_1 = 1 \mid Y = 2)P(X_2 = 0 \mid Y = 2)P(Y = 2) = 1 \cdot 0 \cdot \frac{1}{4} = 0 \end{aligned}\right\} \Rightarrow \hat{y} = 1$$

- **Pentru** $X_1 = 1, X_2 = 1$**:**

$$\left.\begin{aligned} p_1 &= P(X_1 = 1 \mid Y = 1)P(X_2 = 1 \mid Y = 1)P(Y = 1) = \frac{1}{3} \cdot \frac{1}{3} \cdot \frac{3}{4} = \frac{1}{12} \\ p_2 &= P(X_1 = 1 \mid Y = 2)P(X_2 = 1 \mid Y = 2)P(Y = 2) = 1 \cdot 1 \cdot \frac{1}{4} = \frac{1}{4} \end{aligned}\right\} \Rightarrow \hat{y} = 2$$

Deci rata erorii este 0 **pentru acestă definiţie a lui** Y**.**

***Cazul 3*:**

X_1	X_2	Y
0	0	1
0	1	2
1	0	1
1	1	2

Cazuri similare:

Y
2
1
2
1

Y
1
1
2
2

Y
2
2
1
1

- **Pentru** $X_1 = 0, X_2 = 0$**:**

$$p_1 = \frac{1}{2} \cdot 1 \cdot \frac{1}{2} = \frac{1}{4},\ p_2 = \frac{1}{2} \cdot 0 \cdot \frac{1}{2} = 0 \Rightarrow p_1 > p_2 \Rightarrow \hat{y} = 1 \ \textbf{(corect)}$$

- **Pentru** $X_1 = 0, X_2 = 1$**:**

$$p_1 = \frac{1}{2} \cdot 0 \cdot \frac{1}{2} = 0,\ p_2 = \frac{1}{2} \cdot 1 \cdot \frac{1}{2} = \frac{1}{4} \Rightarrow p_1 < p_2 \Rightarrow \hat{y} = 2 \ \textbf{(corect)}$$

- **Pentru** $X_1 = 1, X_2 = 0$**:**

$$p_1 = \frac{1}{2} \cdot 1 \cdot \frac{1}{2} = \frac{1}{4}, p_2 = \frac{1}{2} \cdot 0 \cdot \frac{1}{2} = 0 \Rightarrow p_1 > p_2 \Rightarrow \hat{y} = 1 \ \textbf{(corect)}$$

- **Pentru** $X_1 = 1, X_2 = 1$**:**

$$p_1 = \frac{1}{2} \cdot 0 \cdot \frac{1}{2} = 0,\ p_2 = \frac{1}{2} \cdot 1 \cdot \frac{1}{2} = \frac{1}{4} \Rightarrow p_1 < p_2 \Rightarrow \hat{y} = 2 \ \textbf{(corect)}$$

Prin urmare, rata erorii este 0 **şi în acest caz.**

***Cazul 4*: (cel de la punctul a)**

X_1	X_2	Y
0	0	1
0	1	2
1	0	2
1	1	1

Este similar cu cazul:

Y
2
1
1
2

Concluzie: Doar pentru 2 moduri (cazul 4) de definire a lui Y rata erorii este de 50%; pentru celelalte 14 moduri (cazurile 1, 2, 3) rata erorii este 0.

Computing

The *sample complexity* of the Naive Bayes and Joint Bayes Clssifiers

CMU, 2010 spring, Eric Xing, Tom Mitchell, Aarti Singh, HW2, pr. 1.1

A big reason we use the Naive Bayes classifier is that it requires less training data than the Joint Bayes Classifier. This exercise should give you a "feeling" for how great the disparity really is.

Imagine that each *instance* is an independent "*observation*" of the multivariate random variable $\bar{X} = X_1, ..., X_d$, where the X_i are i.i.d. and Bernoulli of parameter $p = 0.5$.

To train the Joint Bayes classifier, we need to see every value of $\bar{X}$ "enough" times; training the Naive Bayes classifier only requires seeing both values of X_i "enough" times.

Main Question: How many "observations"/instances are needed until, with probability $1 - \varepsilon$, we have seen every variable we need to see at least once?

Note: To train the classifiers well would require more than this, but for this problem we only require one observation.

Hint: You may want to use the following *inequalities*:

- **For any $k \geq 1$, $(1 - 1/k)^k \leq e^{-1}$**
- **For *any* events $E_1, ..., E_k$, $Pr(E_1 \cup \ldots \cup E_k) \leq \sum_{i=1}^{k} Pr(E_i)$.**
  (This is called the "union bounds" property.)

Consider the Naive Bayes classifier.

a. Show that if N observations have been made, the probability that a given value of X_i (either 0 or 1) has *not* been seen is $\leq \frac{1}{2^{N-1}}$.

b. Show that if more than $N_{NB} = 1 + \log_2\left(\frac{d}{\varepsilon}\right)$ observations have been made, then the probability that *any* X_i has not been observed in both states is $\leq \varepsilon$.

Solution:

a. $P(\textbf{component } X_i \textbf{ not seen in both states}) = \left(\frac{1}{2}\right)^N + \left(\frac{1}{2}\right)^N = \frac{2}{2^N} = \frac{1}{2^{N-1}}$

b. $P(\textbf{any component not seen in both states})$

$\leq \sum_{i=1}^{d} P(\textbf{component } X_i \textbf{ not seen in both states})$

$= \sum_{i=1}^{d} \frac{1}{2^{N_{NB}-1}} = d \cdot \frac{1}{2^{N_{NB}-1}} = d \cdot \frac{1}{2^{1+\log_2 \frac{d}{\varepsilon}-1}} = d \cdot \frac{1}{2^{\log_2 \frac{d}{\varepsilon}}} = d \cdot \frac{1}{\frac{d}{\varepsilon}} = d \cdot \frac{\varepsilon}{d} = \varepsilon$

Consider the Joint Bayes classifier.

c. Let $\bar{x}$ be a particular value of $\bar{X}$. Show that after N observations, the probability that we have never seen $\bar{x}$ is $\leq e^{-N/2^d}$.

d. Using the "union bounds" property, show that if more than $N_{JB} = 2^d \ln\left(\frac{2^d}{\varepsilon}\right)$ observations have been made, then the probability that an arbitrarily chosen (but fixed) value of $\bar{X}$ has not been seen is $\leq \varepsilon$.

Solution:

c. $P(\bar{x}$ **not seen in** N **observations**)

$$= \left(1 - \frac{1}{2^d}\right)^N = \left[\left(1 - \frac{1}{2^d}\right)^{2^d}\right]^{N/2^d} \leq \left(\frac{1}{e}\right)^{N/2^d} = e^{-N/2^d}$$

d. $P($**any** $\bar{x}$ **not seen in** N_{JB} **observations**)

$\leq \sum_{\bar{x}} P(\bar{x}$ **not seen in** N_{JB} **observations**)

$$= \sum_{\bar{x}} e^{-N_{JB}/2^d} = 2^d \cdot e^{-N_{JB}/2^d} = 2^d \cdot e^{-\ln \frac{2^d}{\varepsilon}} = 2^d \cdot \frac{1}{e^{\ln \frac{2^d}{\varepsilon}}} = \frac{2^d}{\frac{2^d}{\varepsilon}} = \varepsilon$$

e. Let $d = 2$ and $\varepsilon = 0.1$. What are the values of N_{NB} and N_{JB}?
What about $d = 5$?
And $d = 10$?

Solution:

$$\varepsilon = 0.1,\ d = 2 \quad \Rightarrow \quad \begin{cases} N_{NB} = & 1 + \log_2 \dfrac{2}{0.1} = & 1 + \log_2 20 \approx & 5.32 \\ N_{JB} = & 2^2 \cdot \ln \dfrac{2^2}{0.1} = & 4 \cdot \ln 40 \approx & 14.75 \end{cases}$$

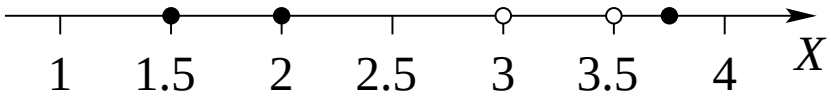
$$\varepsilon = 0.1,\ d = 5 \quad \Rightarrow \quad \begin{cases} N_{NB} = & 1 + \log_2 \dfrac{5}{0.1} = & 1 + \log_2 50 \approx & 6.64 \\ N_{JB} = & 2^5 \cdot \ln \dfrac{2^5}{0.1} = & 32 \cdot \ln 320 \approx & 184.58 \end{cases}$$

$$\varepsilon = 0.1,\ d = 10 \quad \Rightarrow \quad \begin{cases} N_{NB} = & 1 + \log_2 \dfrac{10}{0.1} = & 1 + \log_2 100 \approx & 7.64 \\ N_{JB} = & 2^{10} \cdot \ln \dfrac{2^{10}}{0.1} = & 1024 \cdot \ln 10240 \approx & 9455.67 \end{cases}$$

Exemplifying

ML hypotheses* and *MAP hypotheses

CMU, 2009 spring, Tom Mitchell, midterm, pr. 2.3-4



Let's consider the 1-dimensional data set shown above, based on the single real-valued attribute X. Notice there are two classes (values of Y), and five data points.

Consider a special type of *decision trees* where leaves have *probabilistic labels*. Each leaf node gives the probability of each possible label, where the probability is the fraction of points at that leaf node with that label.

For *example*, a decision tree learned from the data set above with zero splits would say $P(Y = 1) = 3/5$ and $P(Y = 0) = 2/5$. A decision tree with one split (at $X = 2.5$) would say $P(Y = 1) = 1$ if $X < 2.5$, and $P(Y = 1) = 1/3$ if $X \geq 2.5$.

a. For the above data set, draw a tree that maximizes the *likelihood* of the data.

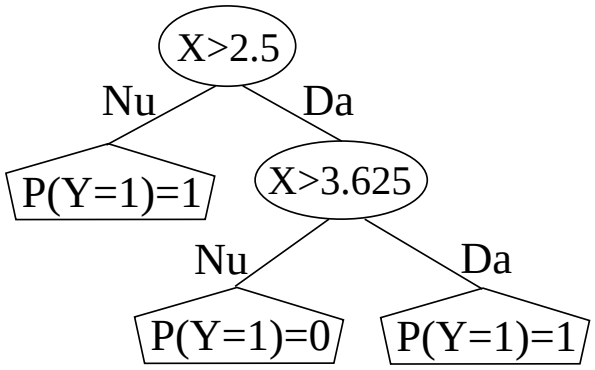
$T_{ML} = \text{argmax}_T \, P_T(D)$, **where**

$P_T(D) \overset{def.}{=} P(D|T) \overset{i.i.d.}{=} \prod_{i=1}^{5} P(Y = y_i | X = x_i, T)$,

where y_i is the label/classs of the instance x_i

$(x_1 = 1.5,\ x_2 = 2,\ x_3 = 3,\ x_4 = 3.5,\ x_5 = 3.75.)$

Solution:



b. Consider a prior probability distribution $P(T)$ over trees that penalizes the number of splits in the tree.

$$P(T) \propto \left(\frac{1}{4}\right)^{splits(T)^2}$$

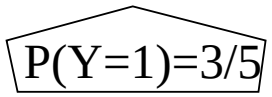
where T is a tree, $splits(T)$ is the number of splits in T, and $\propto$ means "is proportional to".

For the same data set, give the MAP tree when using this prior, $P(T)$, over trees.

Solution:

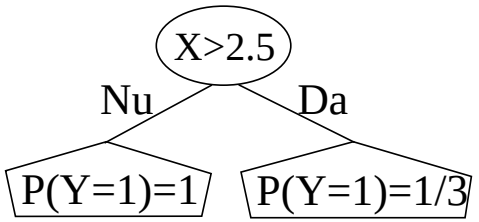
0 nodes:

$$P(T_0 \mid D) \propto \left(\frac{3}{5}\right)^3 \cdot \left(\frac{2}{5}\right)^2 \cdot \left(\frac{1}{4}\right)^0 = \frac{3^3 \cdot 2^2}{5^5} = \frac{108}{3125} = 0.0336$$



1 node:

$$P(T_1 \mid D) \propto 1^2 \cdot \left(\frac{2}{3}\right)^2 \cdot \frac{1}{3} \cdot \left(\frac{1}{4}\right)^1 = \frac{1}{27} = 0.037$$



2 nodes:

$$P(T_2) \propto \left(\frac{1}{4}\right)^4 \Rightarrow P(T_2 \mid D) \propto 1 \cdot \left(\frac{1}{4}\right)^4 = \frac{1}{256} = 0.0039 \Rightarrow \textbf{the MAP tree is } T_1.$$

Logistic Regression: introductory issues

Stanford, 2016 spring, Chris Piech,

Introduction to Probability for Computer Scientists course,

lecture notes #40

Bayesian classification is a task concerned with choosing a value of y that maximizes $P(Y|X)$. The *Naive Bayes* algorithm works by approximating that probability using the *naive assumption* that the features / attributes are mutually independent given the class label. For all classification algorithms you are given n i.i.d. training datapoints $(x^{(1)}, y^{(1)})$, $(x^{(2)}, y^{(2)}), \ldots,$ $(x^{(n)}, y^{(n)})$, where each vector $x^{(i)}$ has d features / attributes. Here we aill assume that $y^{(i)} \in \{0, 1\}$ for $i = 1, \ldots, n$.

***Logistic Regression* is also a classification algorithm that works by trying to learn a function that approximates $P(Y|X)$. It makes the central *assumption* that $P(Y|X)$ can be approximated as a *sigmoid function* (also called logistic function) applied to a linear combination of input features. Mathematically, for a single training datapoint (x, y) Logistic Regression assumes:**

$$P(Y = 1|X = x) = \sigma(z) \text{ and, equivalently } P(Y = 0|X = x) = 1 - \sigma(z)$$

$$\text{where } z \overset{not.}{=} \theta_0 + \sum_{i=1}^{d} \theta_i x_i \overset{not.}{=} \theta \cdot x, \text{ assuming } x_0 = 1,$$

$$\text{and } \sigma(z) = \frac{1}{1 + e^{-z}}.$$

Starting from the above formulas for the probability of $Y|X$, we can create an algorithm that selects values of θ that maximize that probability for all the training data.

a. Prove that the log-likelihood of all the training data (under the Logistic Regression assumption) is:

$$L(\theta) = \sum_{i=0}^{n} \left(y^{(i)} \ln \sigma(\theta \cdot x^{(i)}) + (1 - y^{(i)}) \ln(1 - \sigma(\theta \cdot x^{(i)})) \right). \tag{1}$$

Solution:

To start, here is a super slick way of writing the probability of one datapoint:

$$P(Y = y|X = x) = \sigma(\theta \cdot x)^y \left[1 - \sigma(\theta \cdot x)\right]^{1-y} \textbf{ assuming } y \in \{0, 1\}.$$

Since each datapoint is independent, the probability of all the data is:

$$L(\theta) = \prod_{i=1}^{n} P(Y = y^{(i)}|X = x^{(i)}) = \prod_{i=1}^{n} \sigma(\theta \cdot x^{(i)})^{y^{(i)}} \left[1 - \sigma(\theta \cdot x^{(i)})\right]^{1-y^{(i)}}. \tag{2}$$

If you take the log of this function, you get the reported Log Likelihood for Logistic Regression.

Comment

Now that we have a function for log-likelihood, we simply need to chose the values of θ that maximize it. Unlike in other situations, here there is no closed form way to calculate θ. Instead we will chose it using optimization, so we will employ an algorithm called *gradient ascent*. That algorithm claims that if you continuously take small steps in the direction of your gradient, you will eventually make it to a local maxima. In the case of Logistic Regression you can prove that the result will always be *a global maxima*. The small step that we continually take given the training dataset can be calculated as:

$$\theta_j^{new} = \theta_j^{old} + \eta \frac{\partial}{\partial \theta_j^{old}} L(\theta^{old}),$$

were η is the magnitude of the step size ("learning rate") that we take.

b. Show that the partial derivative of the log-likelihood function with respect to each parameter θ_j is:

$$\frac{\partial}{\partial \theta_j} L(\theta) = \sum_{i=0}^{n} [y^{(i)} - \sigma(\theta \cdot x^{(i)})] x_j^{(i)} \text{ for } j = 0, 1, \ldots, d. \tag{3}$$

Solution:

To start, here is a property for the derivative of σ with respect to its inputs:

$$\frac{\partial}{\partial z} \sigma(z) = \sigma(z)[1 - \sigma(z)] \text{ for } \forall z \in \mathbb{R}.$$

The partial derivative of the log-likelihood for *only* one datapoint (x, y) w.r.t. the θ_j component is computed as follows:

$$\begin{aligned}
&\frac{\partial}{\partial \theta_j} \ln[\sigma(\theta \cdot x)^y \, [1 - \sigma(\theta \cdot x)]^{1-y}] \\
&= \frac{\partial}{\partial \theta_j} y \ln \sigma(\theta \cdot x) + \frac{\partial}{\partial \theta_j} (1 - y) \ln[1 - \sigma(\theta \cdot x)] \\
&= \left[\frac{y}{\sigma(\theta \cdot x)} - \frac{1 - y}{1 - \sigma(\theta \cdot x)} \right] \frac{\partial}{\partial \theta_j} \sigma(\theta \cdot x) \\
&= \left[\frac{y}{\sigma(\theta \cdot x)} - \frac{1 - y}{1 - \sigma(\theta \cdot x)} \right] \sigma(\theta \cdot x)[1 - \sigma(\theta \cdot x)] x_j \\
&= \frac{y - \sigma(\theta \cdot x)}{\sigma(\theta \cdot x)[1 - \sigma(\theta \cdot x)]} \sigma(\theta \cdot x)[1 - \sigma(\theta \cdot x)] x_j \\
&= [y - \sigma(\theta \cdot x)] x_j. \qquad (4)
\end{aligned}$$

Because the derivative of sums is the sum of derivatives, the partial derivative of the log-verosimility function w.r.t. θ_j is simply the sum o this term for each training datapoint. More exactly, after applying the $\ln$ function to the (2) equality, and then calculating its partial derivative w.r.t. θ_j (for $j \in \{0, 1, \ldots, d\}$) we will get the (3) result, due to the (4) relation proven above.

The relationship between [the decision rules of] *Naive Bayes* and *Logistic Regression*; the case of Boolean input variables

CMU, 2005 fall, Tom Mitchell, HW2, pr. 2
CMU, 2009 fall, Carlos Guestrin, HW1, pr. 4.1.2
CMU, 2009 fall, Geoff Gordon, HW4, pr. 1.2
CMU, 2012 fall, Tom Mitchell, Ziv Bar-Joseph, HW2, pr. 3.a

a. [Equivalence of NB and LR]

In Tom's draft chapter (*Generative and discriminative classifiers: Naive Bayes and logistic regression*) it has been proved that when Y follows a Bernoulli distribution and $X = (X_1, \ldots, X_n)$ is a vector of Gaussian variables, then under certain assumptions the Gaussian Naive Bayes classifier implies that $P(Y|X)$ is given by the logistic function with appropriate parameters W. So,

$$P(Y = 1|X) = \frac{1}{1 + \exp(w_0 + \sum_{i=1}^{n} w_i X_i)}.$$

and therefore,

$$P(Y = 0|X) = \frac{\exp(w_0 + \sum_{i=1}^{n} w_i X_i)}{1 + \exp(w_0 + \sum_{i=1}^{n} w_i X_i)}$$

Consider instead the case where Y is Boolean (more generally, Bernoulli) and $X = (X_1, \ldots, X_n)$ is a vector of Boolean variables. Prove for this case also that $P(Y|X)$ follows this same form and hence that Logistic Regression is also the discriminative counterpart to a Naive Bayes generative classifier over Boolean features.

***Note*:**

***Discriminative classifiers* learn the parameters of $P(Y|X)$ directly, whereas *generative classifiers* instead learn the parameters of $P(X|Y)$ and $P(Y)$.**

***Hints*:**

1. Simple notation will help. Since the X_i's are Boolean variables, you need only one parameter to define, $P(X_i|Y = y_k)$, for each $i = 1, \ldots, d$.
Define $\theta_{i1} = P(X_i = 1|Y = 1)$, in which case $P(X_i = 0|Y = 1) = 1 - \theta_{i1}$. Similarly, use θ_{i0} to denote $P(X_i = 1|Y = 0)$.

2. Notice that with the above notation you can represent $P(X_i|Y = 1)$ as follows:

$$P(X_i = 1|Y = 1) = \theta_{i1}^{X_i}(1 - \theta_{i1})^{(1-X_i)}$$

Note that when $X_i = 1$ the second term is equal to 1 because its exponent is zero. Similarly, when $X_i = 0$ the first term is equal to 1 because its exponent is zero.

Solution

$$
\begin{aligned}
P(Y=1|X=x) &\overset{B.F.}{=} \frac{P(X=x|Y=1)\,P(Y=1)}{\sum_{y'\in\{0,1\}} P(X=x|Y=y')\,P(Y=y')} \\
&= \frac{1}{1+\dfrac{P(X=x|Y=0)P(Y=0)}{P(X=x|Y=1)P(Y=1)}} \\
&= \frac{1}{1+\exp\left(\ln\dfrac{P(X=x|Y=0)P(Y=0)}{P(X=x|Y=1)P(Y=1)}\right)} \\
&= \frac{1}{1+\exp\left(\ln\dfrac{P(X_1=x_1,\ldots,X_d=x_d|Y=0)P(Y=0)}{P(X_1=x_1,\ldots,X_d=x_d|Y=1)P(Y=1)}\right)} \\
&\overset{cond.\ indep.}{=} \frac{1}{1+\exp\left(\ln\frac{P(Y=0)}{P(Y=1)}+\sum_{i=1}^{d}\ln\frac{P(X_i=x_i|Y=0)}{P(X_i=x_i|Y=1)}\right)}
\end{aligned}
$$

Prior probabilities are: $P(Y=1)=\pi$ **and** $P(Y=0)=1-\pi$**.**

Also, each X_i **follows a Bernoulli distribution:**
$P(X_i|Y=1)=\theta_{i1}^{X_i}(1-\theta_{i1})^{(1-X_i)}$**, and** $P(X_i|Y=0)=\theta_{i0}^{X_i}(1-\theta_{i0})^{(1-X_i)}$**.**

So,

$$\begin{aligned}
P(Y=1|X=x) &= \frac{1}{1+\exp\left(\ln\frac{1-\pi}{\pi}+\sum_{i=1}^{d}\ln\frac{\theta_{i0}^{X_i}(1-\theta_{i0})^{(1-X_i)}}{\theta_{i1}^{X_i}(1-\theta_{i1})^{(1-X_i)}}\right)} \\
&= \frac{1}{1+\exp\left(\ln\frac{1-\pi}{\pi}+\sum_{i=1}^{d}\left(X_i\ln\frac{\theta_{i0}}{\theta_{i1}}+(1-X_i)\ln\frac{1-\theta_{i0}}{1-\theta_{i1}}\right)\right)} \\
&= \frac{1}{1+\exp\left(\ln\frac{1-\pi}{\pi}+\sum_{i=1}^{d}\ln\frac{1-\theta_{i0}}{1-\theta_{i1}}+\sum_{i=1}^{d}X_i\left(\ln\frac{\theta_{i0}}{\theta_{i1}}+\ln\frac{1-\theta_{i0}}{1-\theta_{i1}}\right)\right)}
\end{aligned}$$

Therefore, in order to reach $P(Y=1|X=x)=1/(1+\exp(w_0+\sum_{i=1}^{d}w_iX_i)$**, we can set**

$$w_0=\ln\frac{1-\pi}{\pi}+\sum_{i=1}^{d}\ln\frac{1-\theta_{i0}}{1-\theta_{i1}} \quad \textbf{and} \quad w_i=\ln\frac{\theta_{i0}}{\theta_{i1}}+\ln\frac{1-\theta_{i0}}{1-\theta_{i1}} \ \textbf{for}\ i=1,\ldots,d.$$

b. [Relaxing the conditional independence assumption]

To capture interactions between features, the Logistic Regression model can be supplemented with extra terms. For example, a term can be added to capture a dependency between X_1 and X_2:

$$P(Y=1|X) = \frac{1}{1+\exp(w_0 + w_{1,2}X_1X_2 + \sum_{i=1}^{n} w_iX_i)}$$

Similarly, the conditional independence assumptions made by Naive Bayes can be relaxed so that X_1 and X_2 are not assumed to be conditionally independent. In this case, we can write:

$$P(Y|X) = \frac{P(Y)\,P(X_1, X_2|Y)\,\prod_{i=3}^{n} P(X_i|Y)}{P(X)}$$

Prove that for this case, that $P(Y|X)$ follows the same form as the logistic regression model supplemented with the extra term that captures the dependency between X_1 and X_2 (and hence that the supplemented Logistic Regression model is the discriminative counterpart to this generative classifier).

***Hints*:**

1. Using simple notation will help here as well. You need more parameters than before to define $P(X_1, X_2, Y)$. So let's define $\beta_{ijk} = P(X_1 = i, X_2 = j, Y = k)$, for each i, j and k.

2. The above notation can be used to represent $P(X_1, X_2|Y = k)$ as follows:

$$P(X_1, X_2|Y = k) = (\beta_{11k})^{X_1 X_2} (\beta_{10k})^{X_1(1-X_2)} (\beta_{01k})^{(1-X_1)X_2} (\beta_{00k})^{(1-X_1)(1-X_2)}$$

Solution

$$
\begin{aligned}
P(Y=1|X) &\overset{B.\ F.}{=} \frac{P(X|Y=1)P(Y=1)}{P(X|Y=1)P(Y=1)+P(X|Y=0)P(Y=0)} \\
&= \frac{1}{1+\dfrac{P(X|Y=0)P(Y=0)}{P(X|Y=1)P(Y=1)}} \\
&= \frac{1}{1+\exp\left(\ln\dfrac{P(X|Y=0)P(Y=0)}{P(X|Y=1)P(Y=1)}\right)} \\
&\overset{cdtl.\ indep.}{=} \frac{1}{1+\exp\left(\ln\dfrac{P(X_1,X_2|Y=0)\prod_{i=3}^{d}P(X_i|Y=0)P(Y=0)}{P(X_1,X_2|Y=1)\prod_{i=3}^{d}P(X_i|Y=1)P(Y=1)}\right)} \\
&= \frac{1}{1+\exp\left(\ln\dfrac{1-\pi}{\pi}+\sum_{i=3}^{d}\ln\dfrac{P(X_i|Y=0)}{P(X_i|Y=1)}+\ln\dfrac{P(X_1,X_2|Y=0)}{P(X_1,X_2|Y=1)}\right)} \\
&= \frac{1}{1+\exp\left(\ln\dfrac{1-\pi}{\pi}+\sum_{i=3}^{d}\ln\dfrac{\theta_{i0}^{X_i}(1-\theta_{i0})^{(1-X_i)}}{\theta_{i1}^{X_i}(1-\theta_{i1})^{(1-X_i)}}+\ln\dfrac{(\beta_{110})^{X_1X_2}\,(\beta_{100})^{X_1(1-X_2)}\,(\beta_{010})^{(1-X_1)X_2}\,(\beta_{000})^{(1-X_1)(1-X_2)}}{(\beta_{111})^{X_1X_2}\,(\beta_{101})^{X_1(1-X_2)}\,(\beta_{011})^{(1-X_1)X_2}\,(\beta_{001})^{(1-X_1)(1-X_2)}}\right)} \\
&= \frac{1}{1+\exp\left(\ln\dfrac{1-\pi}{\pi}+\sum_{i=3}^{d}\left(X_i\left(\ln\dfrac{\theta_{i0}}{\theta_{i1}}+\ln\dfrac{1-\theta_{i1}}{1-\theta_{i0}}\right)+\ln\dfrac{1-\theta_{i1}}{1-\theta_{i0}}\right)+\ln\dfrac{\beta_{000}}{\beta_{001}}+w_1X_1+w_2X_2+w_{1,2}X_1X_2\right)}
\end{aligned}
$$

with

$$
\begin{aligned}
w_0 &= \ln \frac{1-\pi}{\pi} + \sum_{i=3}^{d} \ln \frac{1-\theta_{i1}}{1-\theta_{i0}} + \ln \frac{\beta_{000}}{\beta_{001}} \\
w_1 &= \ln \frac{\beta_{100}}{\beta_{101}} + \ln \frac{\beta_{001}}{\beta_{000}} \\
w_2 &= \ln \frac{\beta_{010}}{\beta_{011}} + \ln \frac{\beta_{001}}{\beta_{000}} \\
w_{1,2} &= \ln \frac{\beta_{110}}{\beta_{111}} + \ln \frac{\beta_{000}}{\beta_{001}} \ln \frac{\beta_{101}}{\beta_{100}} + \ln \frac{\beta_{011}}{\beta_{010}} \\
w_i &= \ln \frac{\theta_{i0}}{\theta_{i1}} + \ln \frac{1-\theta_{i1}}{1-\theta_{i0}} \ \textbf{for} \ i = 3, \ldots, d.
\end{aligned}
$$

Gaussian Baysian Classification

Some properties

Proving

the relationship between the decision rules for

***Gaussian Naive Bayes* and the *Logistic Regression* algorithm**

when the covariance matrices are diagonal and identical

i.e., $\sigma_{0i}^2 = \sigma_{1i}^2$ **for** $i = 1, \ldots, d$

CMU, 2009 spring, Ziv Bar-Joseph, HW2, pr. 2

Assume a two-class ($Y \in \{0, 1\}$) Naive Bayes model over the d-dimensional real-valued input space $\mathbb{R}^d$, where the input variables $X|Y = 0 \in \mathbb{R}^d$ are distributed as

$$Gaussian(\mu_0 =< \mu_{01}, \ldots, \mu_{0d} >,\ \sigma =< \sigma_1, \ldots, \sigma_d >)$$

and $X|Y = 1 \in \mathbb{R}^d$ as

$$Gaussian(\mu_1 =< \mu_{11}, \ldots, \mu_{1d} >,\ \sigma =< \sigma_1, \ldots, \sigma_d >)$$

i.e., the inputs given the class have different means but identical variance for both classes.

Prove that, given the conditions stated above, the conditional probability $P(Y = 1|X = x)$, where $X = (X_1, \ldots, X_d)$ and $x = (x_1, \ldots, x_d)$ can be written in a simiar form to Logistic Regression:

$$\frac{1}{1 + \exp(w_0 + w \cdot x)}$$

with the parameters $w_0 \in \mathbb{R}$ and $w = (w_1, \ldots, w_d) \in \mathbb{R}^d$ chosen in a suitable way.

As a consequence, the decision rule for the Gaussean Bayes classifier supported by this model the desion rule has a linear form.

Solution

$$
\begin{aligned}
P(Y=1|X=x) &\overset{B.F.}{=} \frac{P(X=x|Y=1)\,P(Y=1)}{\sum_{y'\in\{0,1\}} P(X=x|Y=y')\,P(Y=y')} \\
&= \frac{1}{1+\dfrac{P(X=x|Y=0)P(Y=0)}{P(X=x|Y=1)P(Y=1)}} \\
&= \frac{1}{1+\exp\left(\ln\dfrac{P(X=x|Y=0)P(Y=0)}{P(X=x|Y=1)P(Y=1)}\right)} \\
&= \frac{1}{1+\exp\underbrace{\left(\ln\dfrac{P(X_1=x_1,\ldots,X_d=x_d|Y=0)P(Y=0)}{P(X_1=x_1,\ldots,X_d=x_d|Y=1)P(Y=1)}\right)}_{exponent}}
\end{aligned}
$$

$$
\begin{aligned}
\textit{exponent} \overset{\textit{cond. indep.}}{=} & \ln \frac{P(Y=0)}{P(Y=1)} + \sum_{i=1}^{d} \ln \frac{P(X_i = x_i | Y=0)}{P(X_i = x_i | Y=1)} \\
= & \ln \frac{P(Y=0)}{P(Y=1)} + \sum_{i=1}^{d} \ln \left(\frac{\frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(x_i-\mu_{i0})^2}{2\sigma_i^2}\right)}{\frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(x_i-\mu_{i1})^2}{2\sigma_i^2}\right)} \right) \\
= & \ln \frac{P(Y=0)}{P(Y=1)} + \sum_{i=1}^{d} \left(\frac{(x_i-\mu_{i1})^2}{2\sigma_i^2} - \frac{(x_i-\mu_{i0})^2}{2\sigma_i^2} \right) \\
= & \ln \frac{P(Y=0)}{P(Y=1)} + \sum_{i=1}^{d} \frac{2x_i(\mu_{0i}-\mu_{1i}) + (\mu_{1i}^2 - \mu_{0i}^2)}{2\sigma_i^2} \\
= & \ln \frac{P(Y=0)}{P(Y=1)} + \sum_{i=1}^{d} \left(\frac{x_i(\mu_{0i}-\mu_{1i})}{\sigma_i^2} + \frac{(\mu_{1i}^2 - \mu_{0i}^2)}{2\sigma_i^2} \right) \\
= & \underbrace{\ln \frac{P(Y=0)}{P(Y=1)} + \sum_{i=1}^{d} \frac{(\mu_{1i}^2 - \mu_{0i}^2)}{2\sigma_i^2}}_{w_0} + \sum_{i=1}^{d} \underbrace{\frac{\mu_{0i}-\mu_{1i}}{\sigma_i^2}}_{w_i} x_i
\end{aligned}
$$

In conclusion,

$$P(Y = 1|X = x) = \frac{1}{1 + e^{(w \cdot x + w_0)}}$$

with

$$w_0 = \ln \frac{P(Y = 0)}{P(Y = 1)} + \sum_{i=1}^{d} \frac{(\mu_{1i}^2 - \mu_{0i}^2)}{2\sigma_i^2} \textbf{ and } w_i = \frac{\mu_{0i} - \mu_{1i}}{\sigma_i^2}, i = 1, \ldots, d$$

Note that

$$P(Y = 0|X = x) = \frac{e^{(w \cdot x + w_0)}}{1 + e^{(w \cdot x + w_0)}}$$

and

$$P(Y = 1|X = x) > P(Y = 0|X = x) \Leftrightarrow w \cdot x + w_0 < 0$$

Since the coefficients w_i for $i = 1, \ldots, d$ do not depend on x_i, it follows that this *decison rule* of Gaussian Naive Bayes [in the conditions stated in the beginning of this problem] is a linear rule, like in Logistic Regression.

However, this relationship does not mean that there is a one-to-one correspondence between the parameters w_i of Gaussian Naive Bayes (GNB) and the parameters w_i of logistic regression (LR) because LR is discriminative and therefore doesn't model $P(X)$, while GNB does model $P(X)$.

To be more specific, note that the coefficients w_i in the GNB decision rules should be devided by $P(x_1, \ldots, x_d)$ in order to correspond to $P(Y = 1|X = x)$, which means that then they will not anymore be independent of x_i, like the LR coefficients.

Estimating the parameters for Gaussian Naive Bayes and Full/Joint Gaussian Naive Bayes algorithms

CMU, 2014 fall, W. Cohen, Z. Bar-Joseph, HW2, pr. 5.ab

Consider a Gaussian Naive Bayes model, where the conditional distribution of each feature is a one-dimensional Gaussian, $X^{(j)}|Y \sim N(\mu_Y^{(j)}, (\sigma_Y^{(j)})^2)$, $j = 1, \cdots, d$.

a. Given n **independent training data points,** $\{(X_1, Y_1), \cdots, (X_n, Y_n)\}$**, give a maximum-likelihood estimate (MLE) of the conditional distribution of feature** $X^{(j)}, j = 1, \ldots, d$.

Solution:

The likelihood of the samples in Class 0 is

$$\begin{aligned} L(X_{i,0}^{(j)}|\mu_0^{(j)},(\sigma\mu_0^{(j)})^2) &= \prod_{i=1}^{n_0}\frac{1}{\sqrt{2\pi}\sigma_0^{(j)}}\exp\left(-\frac{(X_{i,0}^{(j)}-\mu_0^{(j)})^2}{2(\sigma_0^{(j)})^2}\right) \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma_0^{(j)}}\right)^{n_0}\exp\left(-\sum_{i=1}^{n_0}\frac{(X_{i,0}^{(j)}-\mu_0^{(j)})^2}{2(\sigma_0^{(j)})^2}\right) \end{aligned}$$

and the log-likelihood is

$$\ln L = -n_0\ln\sigma_0^{(j)} - \frac{1}{2(\sigma_0^{(j)})^2}\sum_{i=1}^{n_0}(X_{i,0}^{(j)}-\mu_0^{(j)})^2 + constant$$

Taking the partial derivatives of the log-likelihood, we have

$$\frac{\partial\ln L}{\partial\mu_0^{(j)}} = 0 \Leftrightarrow \sum_{i=1}^{n_0}(X_{i,0}^{(j)}-\mu_0^{(j)}) = 0 \Leftrightarrow \mu_0^{(j)} = \frac{1}{n_0}\sum_{i=1}^{n_0}X_{i,0}^{(j)}$$

$$\frac{\partial\ln L}{\partial\sigma_0^{(j)}} = 0 \Leftrightarrow -\frac{n_0}{\sigma_0^{(j)}} + \frac{1}{(\sigma_0^{(j)})^3}\sum_{i=1}^{n_0}(X_{i,0}^{(j)}-\mu_0^{(j)})^2 = 0 \Leftrightarrow (\sigma\mu_0^{(j)})^2 = \frac{1}{n_0}\sum_{i=1}^{n_0}(X_{i,0}^{(j)}-\hat{\mu}_0^{(j)})^2$$

Similarly, one can derive the MLE for the parameters in Class 1.

b. Suppose the prior of Y is already given. How many parameters do you need to estimate in Gaussian Naive Bayes model?

Solution:

For each class, there are 2 parameters (the mean and variance) for each feature, therefore there are $2 \cdot 2d = 4d$ parameters for all features in the two classes.

c. In a full/Joint Gaussian Bayes model, we assume that the conditional distribution $\Pr(X|Y)$ is a multidimensional Gaussian, $X|Y \sim \mathcal{N}(\mu_Y, \Sigma_Y)$, where $\mu \in \mathbb{R}^d$ is the mean vector and $\Sigma \in \mathbb{R}^{d \times d}$ is the covariance matrix.

Again, suppose the prior of Y is already given. How many parameters do you need to estimate in a full/Joint Gaussian Bayes model?

Solution:

For each class, there are d parameters for the mean, $d(d+1)/2$ parameters for the covariance matrix, because the covariance matrix is symmetric. Therefore, the number of parameters is $2 \cdot (d + d(d+1)/2) = d(d+3)$ in total for the two classes.

Proving

the relationship between

The full Gaussian Bayes* algorithm and *Logistic Regression

when $\Sigma_0 = \Sigma_1$

CMU, 2011 spring, Tom Mitchell, HW2, pr. 2.2

Let's make the following *assumptions*:

1. Y **is a boolean variable following a Bernoulli distribution, with parameter** $\pi = P(Y = 1)$ **and thus** $P(Y = 0) = 1 - \pi$**.**

2. $X =< X_1, X_2, \ldots, X_n >$ **is a vector of random variables *not* conditionally independent given** Y**, and** $P(X|Y = k)$ **follows a *multivariate normal distribution*** $N(\mu_k, \Sigma)$**.**

Note that μ_k **is the** $n \times 1$ **mean vector depending on the value of** Y**, and** Σ **is the** $n \times n$ **covariance matrix, which does not depend on** Y**. We will write/use the density of the multivariate normal distribution in vector/matrix notation.**

$$\mathcal{N}(x; \mu\Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)$$

Is the form of $P(Y|X)$ **implied by such this [not-so-naive] Gaussian Bayes classifier [LC: similar to] the form used by logistic regression?**
Derive the form of $P(Y|X)$ **to prove your answer.**

We start with:

$$\begin{aligned} P(Y=1|X) &= \frac{P(X|Y=1)\,P(Y=1)}{P(X|Y=1)\,P(Y=1)+P(X|Y=0)\,P(Y=0)} \\ &= \frac{1}{1+\dfrac{P(Y=0)\,P(X|Y=0)}{P(Y=1)\,P(X|Y=1)}} = \frac{1}{1+\exp\left(\ln\dfrac{P(Y=0)\,P(X|Y=0)}{P(Y=1)\,P(X|Y=1)}\right)} \\ &= \frac{1}{1+\exp\left(\ln\dfrac{P(Y=0)}{P(Y=1)}+\ln\dfrac{P(X|Y=0)}{P(X|Y=1)}\right)} \end{aligned}$$

Next we will focus on the term $\ln\dfrac{P(X|Y=0)}{P(X|Y=1)}$**:**

$$\ln\frac{P(X|Y=0)}{P(X|Y=1)} = \ln\frac{\dfrac{1}{\sqrt{(2\pi)^{d/2}|\Sigma|^{1/2}}}}{\dfrac{1}{\sqrt{(2\pi)^{d/2}|\Sigma|^{1/2}}}} + \ln\exp[(\star)] = \ln\exp[(\star)] = (\star)$$

where $(\star)$ **is the formulation obtained as the difference between the exponential parts of two multivariate Gaussian densities** $P(X|Y=0)$ **and** $P(X|Y=1)$**.**

$$
\begin{aligned}
(\star) &= \frac{1}{2}[(X-\mu_1)^\top \Sigma^{-1}(X-\mu_1) - (X-\mu_0)^\top \Sigma^{-1}(X-\mu_0)] \\
&= (\mu_0^\top - \mu_1^\top)\Sigma^{-1}X + \frac{1}{2}\mu_1^\top \Sigma^{-1}\mu_1 - \frac{1}{2}\mu_0^\top \Sigma^{-1}\mu_0
\end{aligned}
$$

As a result, we have:

$$
\begin{aligned}
P(Y=1|X) &= \frac{1}{1+\exp\left(\ln\frac{1-\pi}{\pi} + \frac{1}{2}\mu_1^\top \Sigma^{-1}\mu_1 - \frac{1}{2}\mu_0^\top \Sigma^{-1}\mu_0 + (\mu_0^\top - \mu_1^\top)\Sigma^{-1}X\right)} \\
&= \frac{1}{1+\exp(w_0 + w^\top X)}
\end{aligned}
$$

where $w_0 = \ln\frac{1-\pi}{\pi} + \frac{1}{2}\mu_1^\top \Sigma^{-1}\mu_1 - \frac{1}{2}\mu_0^\top \Sigma^{-1}\mu_0$ **is a scalar,**
and $w = \Sigma^{-1}(\mu_0 - \mu_1)$ **is a** $d \times 1$ **a parameter vector.**

Note that $((\mu_0^\top - \mu_1^\top)\Sigma^{-1})^\top = ((\mu_0 - \mu_1)^\top \Sigma^{-1})^\top = (\Sigma^{-1})^\top((\mu_0 - \mu_1)^\top)^\top = \Sigma^{-1}(\mu_0 - \mu_1)$ **because** Σ^{-1} **is symmetric.**
(Σ **is symmetric because it is a covariance matrix, and therefore** Σ^{-1} **is also symmetric.)**

In conclusion, $P(Y|X)$ **has the form of the logistic regression (in vector and matrix notation).**

Gaussian Baysian Classification

Some exercises

Exemplifying the Gaussian [Naive] Bayes algorithm on data from $\mathbb{R}$

CMU, 2001 fall, Andrew Moore, midterm, pr. 3.a

Suppose you have the nearby training set with one real-valued input X and a categorial output Y that has two values.

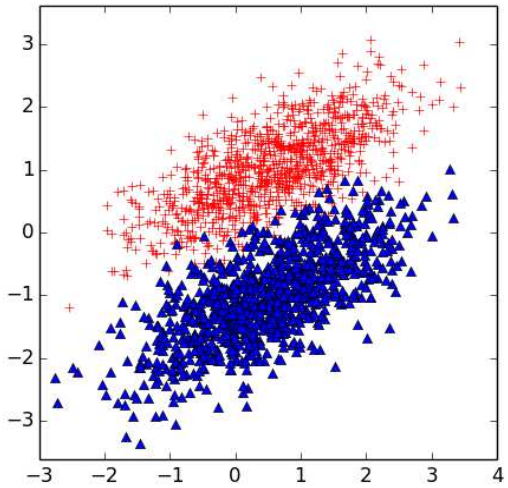
X	Y
0	A
2	A
3	B
4	B
5	B
6	B
7	B

a. You must learn from this data the parameters of the Gaussian Bayes classifer. Write your answer in the following table.

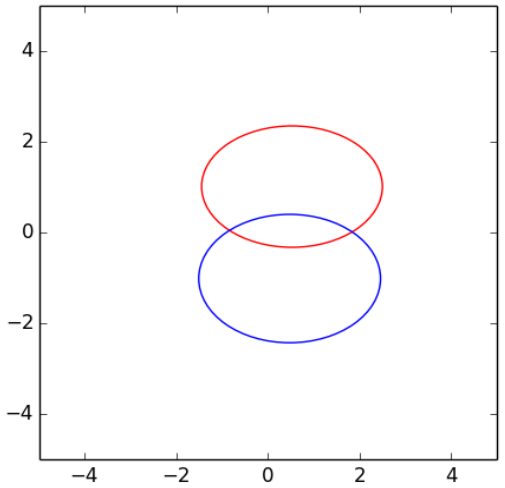
$\mu_A =$	$\sigma_A^2 =$	$P(Y = A) =$
$\mu_B =$	$\sigma_B^2 =$	$P(Y = B) =$

b. Using the notation $\alpha = p(X = 2|Y = A)$ and $\beta = p(X = 2|Y = B)$,

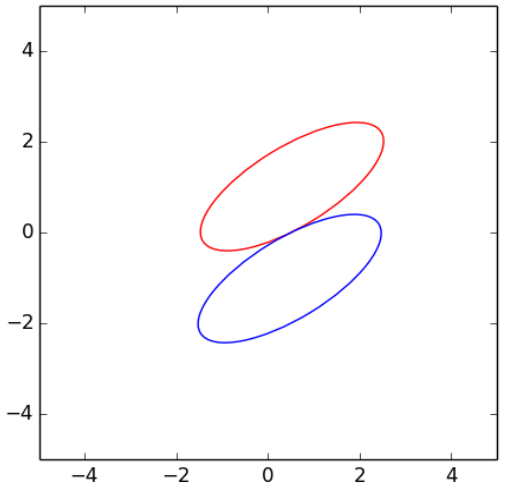
- **What is $p(X = 2, Y = A)$? (Answer in terms of α.)**
- **What is $p(X = 2, Y = B)$? (Answer in terms of β.)**
- **What is $p(X = 2)$? (Answer in terms of α and β.)**
- **What is $p(Y = A|X = 2)$? (Answer in terms of α and β.)**
- **How would the point $X = 2$ be classified by the Gaussian Bayes algorithm? (Answer in terms of α and β.)**



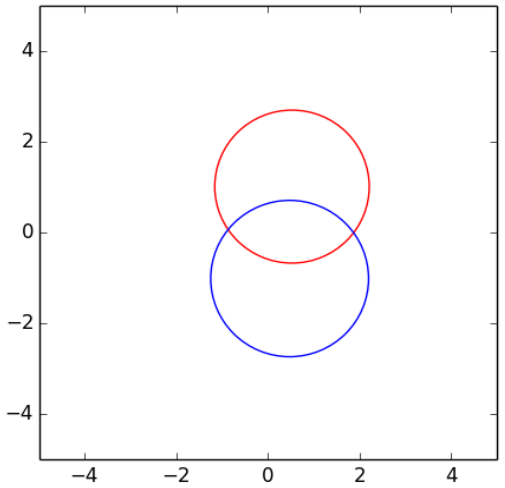
(A) Data



(B)



(C)



(D)

Solution:

a. (C) is the truth.

b. (B) corresponds to the Gaussian Naive Bayes estimates. [LC: Here follows the explanation:]
Because the Gaussian Naive Bayes model assume independence of the two features conditioned on the class label, the estimated model should be aligned with the axies. Both (B) and (D) satisfy this, but only in (B) the width and height of the oval, which are proportional to the standard deviation of each axis, matched the data.

c. (C) gives the lowest training error.

Exemplifying the Gaussian [Naive] Bayes algorithm on data from $\mathbb{R}^2$

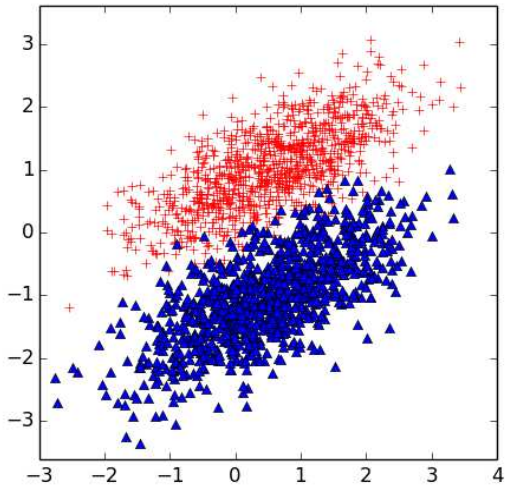
CMU, 2014 fall, W. Cohen, Z. Bar-Joseph, HW2, pr. 5.c

In a two dimensional case, we can visualize how Gaussian Naive Bayes behaves when input features are correlated. A data set is shown in Figure (A), where red points are in Class 0, blue points are in Class 1. The conditional distributions are two-dimensional Gaussians. In (B), (C) and (D), the ellipses represent conditional distributions for each class. The centers of ellipses show the means, and the contours show the boundary of two standard deviations.

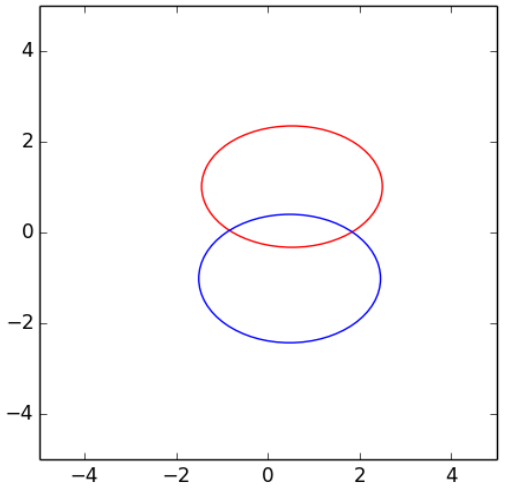
a. Which of them is most likely to be the true conditional distribution?

b. Which of them is most likely to be estimates by a Gaussian Naive Bayes model?

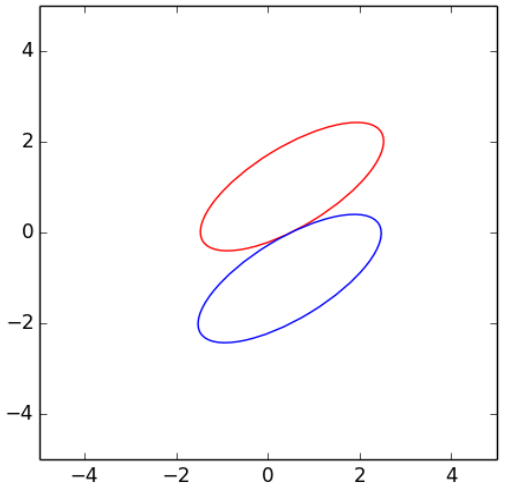
c. If we assume the prior probabilities for both classes are equal, which model will achieve a higher accuracy on the training data?



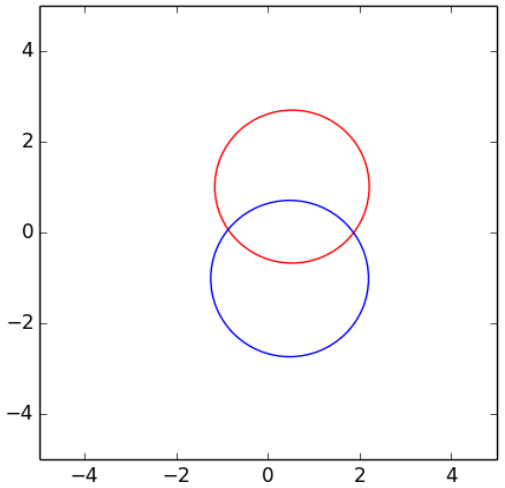
(A) Data



(B)



(C)



(D)

Solution:

a. (C) is the truth.

b. (B) corresponds to the Gaussian Naive Bayes estimates. [LC: Here follows the explanation:]
Because the Gaussian Naive Bayes model assume independence of the two features conditioned on the class label, the estimated model should be aligned with the axies. Both (B) and (D) satisfy this, but only in (B) the width and height of the oval, which are proportional to the standard deviation of each axis, matched the data.

c. (C) gives the lowest training error.