
DS 740
Data Mining
Bayes' Rule and Classification
Background for LDA

Learning Objectives

By the end of this lesson, you will be able to:

- Work with Bayes' rule to compute conditional probabilities.
- Understand sensitivity and specificity in context of predicting a qualitative response.
- Describe how such classification will be accomplished.

Bayes' Rule (from Bayes' Theorem)

	Notation	Denotes
Probability	*Pr*(A)	Probability that event A occurs
Conditional Probability	*Pr*(A | B)	Probability that event A occurs **given** that event B occurs

Basic form of Bayes' rule:

$$\Pr(A|B) = \frac{\Pr(B|A) \cdot \Pr(A)}{\Pr(B|A) \cdot \Pr(A) + \Pr(B|A') \cdot \Pr(A')}$$

A' denotes the **complement**: A does *not* occur.

Bayes' rule is a 253-year-old concept introduced by Thomas Bayes, an English minister from the 1700s. It forms the foundation for a way of thinking about statistics that is an alternative to the frequentist or traditional viewpoint of statistics. It has gained prominence, particularly in the last couple decades, due to the increased computational power that provide the ability to compute many of the posterior distributions.

Its basic form can be stated in terms of events or sets of occurrences, which we'll label A and B. Our goal is to obtain the conditional probability of A given B. And we can do so by using other probabilities, including opposite, so to speak, conditional probabilities to compute it.

This formula can be thought of as the posterior probability of A-- that is, given or knowing that B occurs. And it's computed as a ratio, with the top of the ratio being the conditional probability of B given A, the opposite conditional probability, times the prior probability of A. And the denominator is the sum of conditional probabilities times priors over the possible prior options.

Notes:

For a great basic introduction of Bayes' rule, see B.
Efron's lecture and paper (with abstract) for MAA Distinguished Lecture, "A 250-Year Argument (Belief, Behavior and the Bootstrap)."

Example of Bayes' Rule

Rare disease: affects only 1 in 1000 people.

- *D* denotes "person has disease"
- So *Pr*(*D*) = 1/1000 = 0.001 and *Pr*(*D'*) = 1 – 0.001 = 0.999

Test will diagnose disease for 99.5% of people with disease.

- *P* denotes "test is positive", thus *Pr*(*P*|*D*) = 0.995

Test will be negative for 95% of people without disease.

- *N* denotes "test is negative", *Pr*(*N*|*D'*) = 0.95

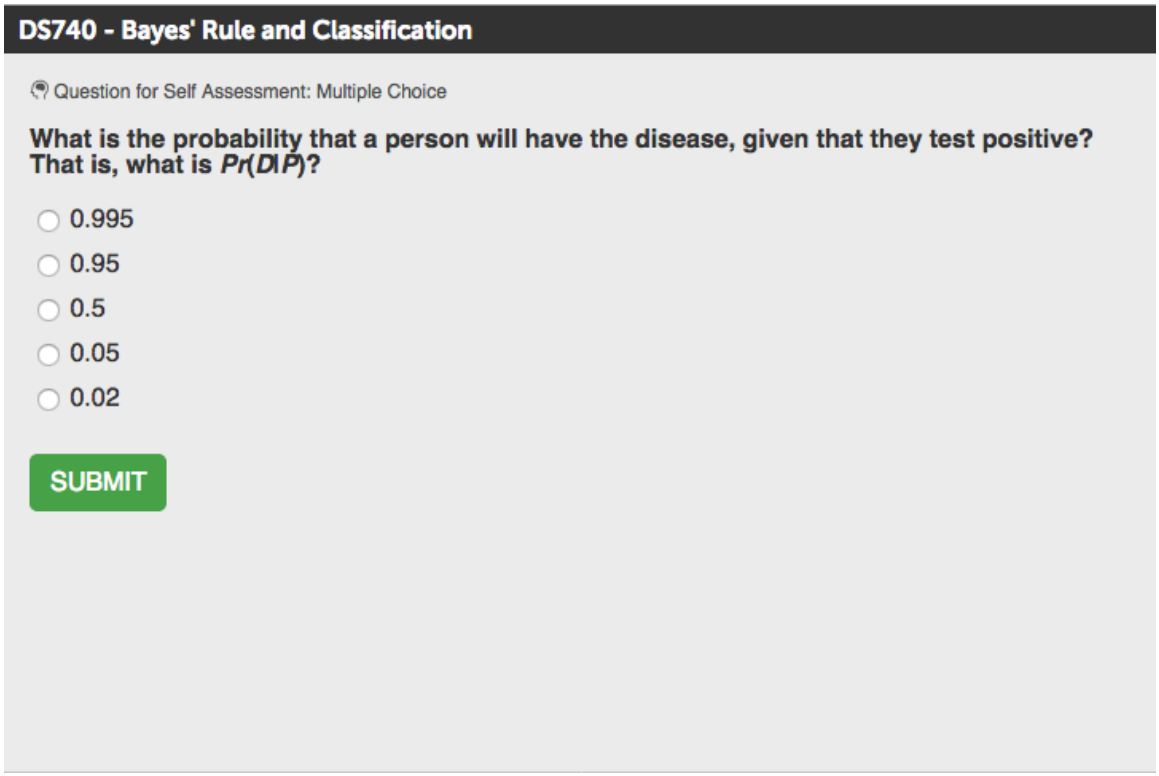
Bayes' rule: $\Pr(D|P) = \dfrac{\Pr(P|D) \cdot \Pr(D)}{\Pr(P|D) \cdot \Pr(D) + \Pr(P|D') \cdot \Pr(D')}$

A common example is to consider a test for a rare disease. Such a situation will allow us not only to discuss Bayes' rule, but also how observed versus predictive classifications are characterized. Relabeling appropriate to context, we let D represent the event, person has disease, and P represent the test for that disease is positive.

Prior probabilities-- that is, not having any test results involved-- tell us that there is a probability of 1 in 1,000 of having the disease, and thus a probability of 0.999, or 999 out of 1,000, of not having the disease.

Conditional probabilities of the test accuracies are also known and shown in the list. The probability of a positive diagnosis for a person, or given that a person has the disease, is 0.995. The probability of a negative diagnosis, given that a person does not have the disease, is 0.95. We'll discuss these test accuracies later.

Question 1



Answer is at the end of this transcript

A Different Organization of Data

1,000,000 individuals in a population.

	− "negative"	+ "positive"	Total
− "no disease"	949,050	49,950	999,000
+ "disease"	5	995	1,000
Total	949,055	50,945	1,000,000

Only ~ 2% of these individuals have disease

We will now reframe the situation in this example as count data. Consider a population that has a million individuals in it. Only a tenth of a percent, or 1,000 of these individuals, have the disease, according to the previous probability, since it is a rare disease.

So in turn, note that this means that 999,000 individuals do not have the disease. And this is where the counts and probabilities of individuals getting the disease based on their true situation come into play. Out of the 1,000 individuals with the disease, 99.5%, or a count of 995 individuals, will test positive, and five, the remaining five, will test negative.

Similarly, out of the 999,000 individuals without the disease, 95%, giving us a count of 949,050 individuals, will test negative. The remainder, for a count of 49,950, will test positive.

At the bottom of this table, we'll total the numbers of individuals who test negative and positive. And what is very interesting is taking a look at the numbers who test positive. So, note that only about 2% of the individuals who test positive are actually ones with the disease. And that's because there are so many people who don't have the disease, some of them will test positive, about 5%, and that gives us a big count.

Answer to Bayes' Rule Example

Out of those for whom the test is positive, only about 2% actually have the disease! Why?

$$\Pr(D|P) = \frac{\Pr(P|D) \cdot \Pr(D)}{\Pr(P|D) \cdot \Pr(D) + \Pr(P|D') \cdot \Pr(D')}$$

$$= \frac{0.995 \cdot 0.001}{0.995 \cdot 0.001 + 0.05 \cdot 0.999} \approx 0.019531$$

$$= \frac{\overbrace{995}^{\text{true positives}}}{995 + \underbrace{49950}_{\text{false positives}}} = \frac{995}{50945}$$

So, seeing this in equation form, if we fill in the probabilities to the formula for Bayes' rule, we calculate that only 2% of those who test positive actually have the disease. This may seem surprising. But going back to the rarity of the disease overall, along with the various accuracies of the test, we can help lay out the reason.

So, out of the 1,000 people who have the disease, we have a very accurate test for diagnosing the disease in people who actually have it. And so we get almost all of them-- that is, 995 people-- added in as positives. And these are the true positives.

But there is a very large group of people who do not have the disease, 999,000. And since the test is reasonably accurate, or specific, for diagnosing no disease, most of them will test negative. But there's still a decent chunk, about 5%, who test positive. This works out to be almost 50,000 people. And this larger count is the set of false positives. And so note that our probability as computed via the formula works out to be the same thing as taking the 995 true positives out of the total number of positives.

Classification

Observed Class	Predicted Class: "no"	Predicted Class: "yes"	Total
"no"	TN	FP	N
"yes"	FN	TP	P
Total	N*	P*	n

N: observed negatives

N*: predicted negatives

TN: true negatives from prediction.

Etc.

Sticking with the idea of classifying into two classes, we connect the relationship between predicted and observed classes through terminology. Letting N denote-- capital N-- the count of observed negatives, capital N* represent count of predicted negatives. We'll also let P and P* be the corresponding counts for the positives.

Then inside the table, we can break down the entries as true negatives, or TN, false positives, false negatives, and true positives. And so we see that the false positives and true positives add up to the total number of predicted positives, as the true negatives and false negatives add up to the total number of predicted negatives.

Notes:

N: observed negatives
P: observed positives
N*: predicted negatives
P*: predicted positives
n: total population
TN: true negatives from prediction
FN: false negatives from prediction
FP: false positives from prediction
TP: true positives from prediction

Summary of Classification Prediction

True Negative (TN) *rate*	=	**specificity**	=	TN/N
False Positive (FP) *rate*	=	1-**specificity**	=	FP/N
False Negative (FN) *rate*	=	1-**sensitivity**	=	FN/P
True Positive (TP) *rate*	=	**sensitivity**	=	TP/P

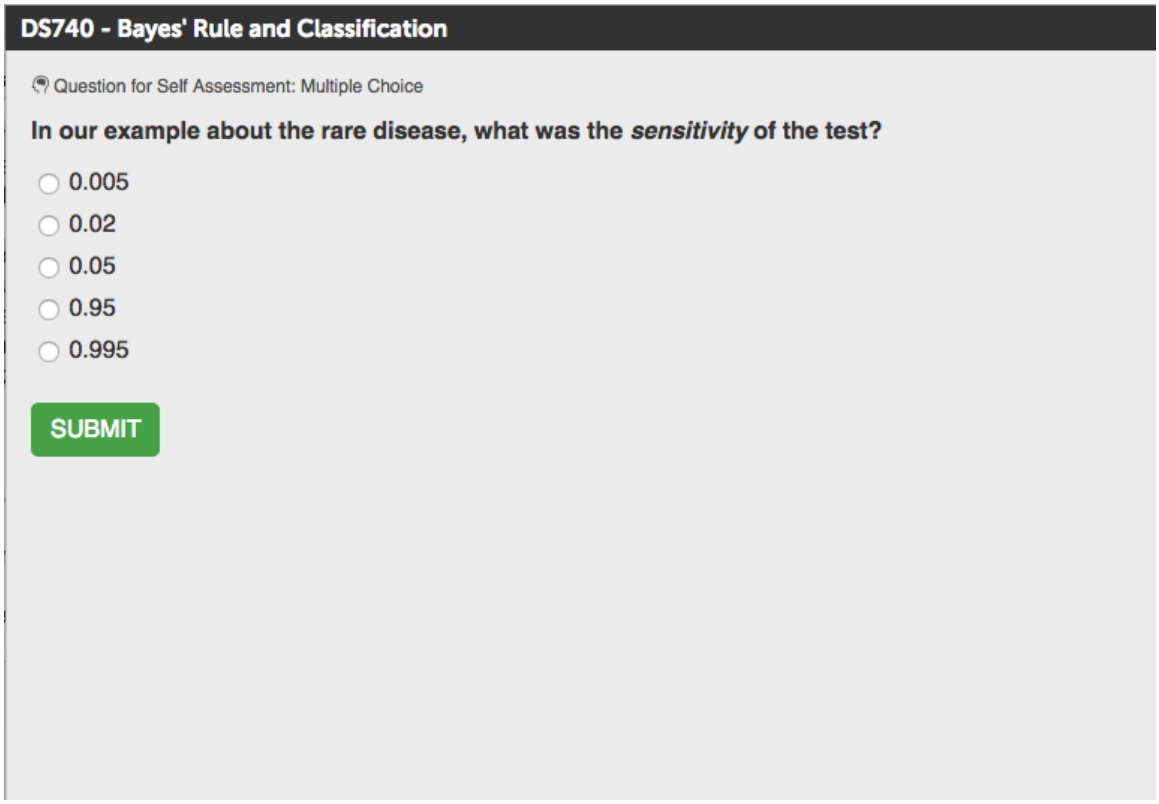
Example Revisited

- Rare disease affects only 1 in 1000 people.
- Test will diagnose disease for 99.5% of people with disease.
- Test will be negative for 95% of people without disease.
- 2% of positive tests actually have disease.

The rates of true identification are the focus in talking about classification models. Notably, we define specificity to be the true negative rate-- that is, the number of actually identified negatives over the total number of negatives. And similarly, sensitivity is the true positive rate.

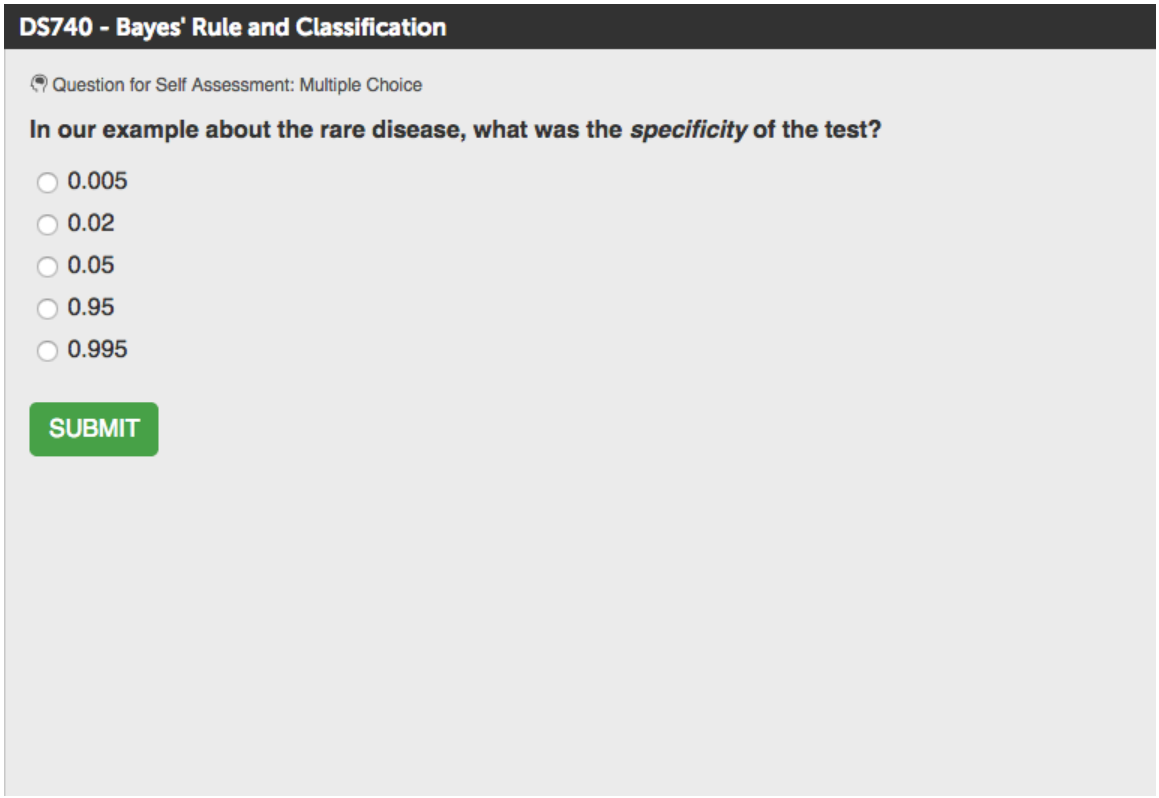
So to summarize, in this situation, we have a rare disease affecting only 1 in 1,000 people with a very high sensitivity in terms of a true positive rate and a high specificity in terms of a true negative rate. But because it is rare, we wind up with a lot more positive tests from those who are actually negative. So sensitivity and specificity continue to be very important to know, but results must be interpreted in light of the number of occurrences of each outcome, each class.

Question 2



Answer is at the end of this transcript

Question 3



Answer is at the end of this transcript

Goal

In example, want to place person into correct group/class of observed disease (disease/not disease) based on some information.

How might one predict? Eg. white blood cell count, blood sugar reading, etc. → continuous variable

General **goal** of discriminant analysis:

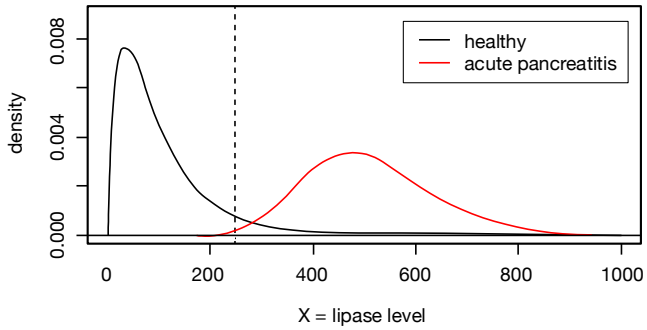
- classify a response *Y* into...
- one of *K* groups/classes...
- based on one or more continuous predictor variables *X*.

How might such a test be run? A diagnosis that is a positive or negative test result could be based off of some physiological measurement, such as white blood cell count, blood sugar reading, level of pancreatic enzymes, et cetera. For the application of linear discriminant analysis, we will focus on continuous predictors. In the next lecture, we will discuss why the predictors must be continuous, as well as what additional assumptions are made about the distribution of the predictors.

We will describe extension to more classes in the next lecture. And so we'll state our generic goal here. We wish to classify a response Y into one of K groups or classes based off one or more continuous predictor variables X.

Understanding Classification by *X*

$$Y = \begin{cases} 0 & \text{healthy} \\ 1 & \text{acute pancreatitis} \end{cases}$$



X = Levels of lipase levels (units/liter)

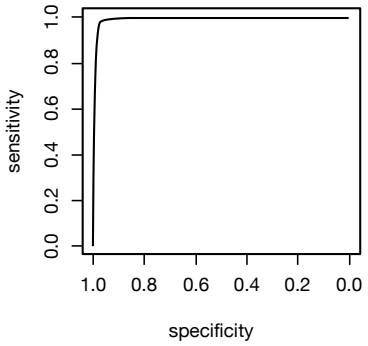
Going back to our disease example, suppose the situation is such that we are attempting to diagnose acute pancreatitis as an illness. One indicator of this illness is lipase level in units per liter. Suppose the distribution shown in the picture below represent healthy individuals, seen in the black line distribution, versus individuals with acute pancreatitis, as represented by the distribution with the red line.

Notice that these two distributions are of the lipase levels. And so we're going to use that X predictor variable to classify individuals into either healthy versus acute pancreatitis.

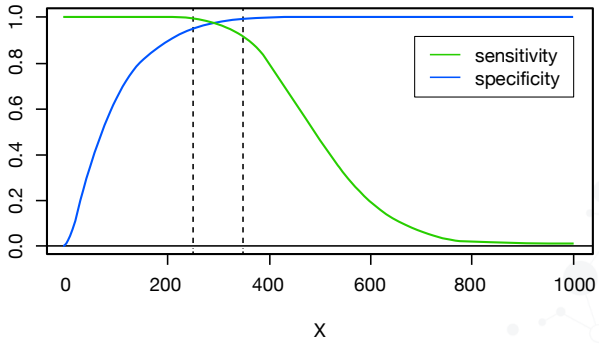
Note also that there is not much overlap. For example, looking at the value X equals 250, only a few, a small proportion, of the healthy individuals have a lipase level higher than that value. And only a very few individuals with acute pancreatitis have a lipase level below that. So, if these were the distributions, lipase level appears that it would do a good job distinguishing between healthy individuals and those with acute pancreatitis.

ROC Curve for Test

ROC curve:
highly accurate test.



Test highly sensitive and specific
for *X*-values from about 250-350.



We take a further look at two images to help us understand how this test is performing. The first image is the previously introduced ROC curve. And because of the area underneath the curve being very close to 1, we note that this is a highly accurate test. And that can be further illustrated by looking at the sensitivity and specificity values that we could get for different choices for the X value as a cutoff to classify individuals.

And if we take an X value anywhere between about 250 to 350, we notice that both the green line, representing sensitivity, and the blue line, representing specificity, will have values quite close to 1 for X's anywhere over that range. There will be a trade-off, of course, as there is between sensitivity and specificity. But over that range, it's a pretty even match between the two.

Summary

We have seen Bayes' rule as a basic probability computation tool. It allows a user to incorporate **sensitivity** and **specificity** of a prediction, together with the **prevalence** of an outcome, to compute the probability of that outcome, given input information.

We have also introduced the goal of discriminant analysis: prediction by classifying a response Y into one of K classes, based off one or more continuous predictor variables X.

Connection between these two concepts will be made in the next presentation.

Question 1 Answer

DS740 - Bayes' Rule and Classification

Feedback for Self Assessment

✓ Correct!

What is the probability that a person will have the disease, given that they test positive? That is, what is $Pr(D|P)$?

Your answer:

0.02

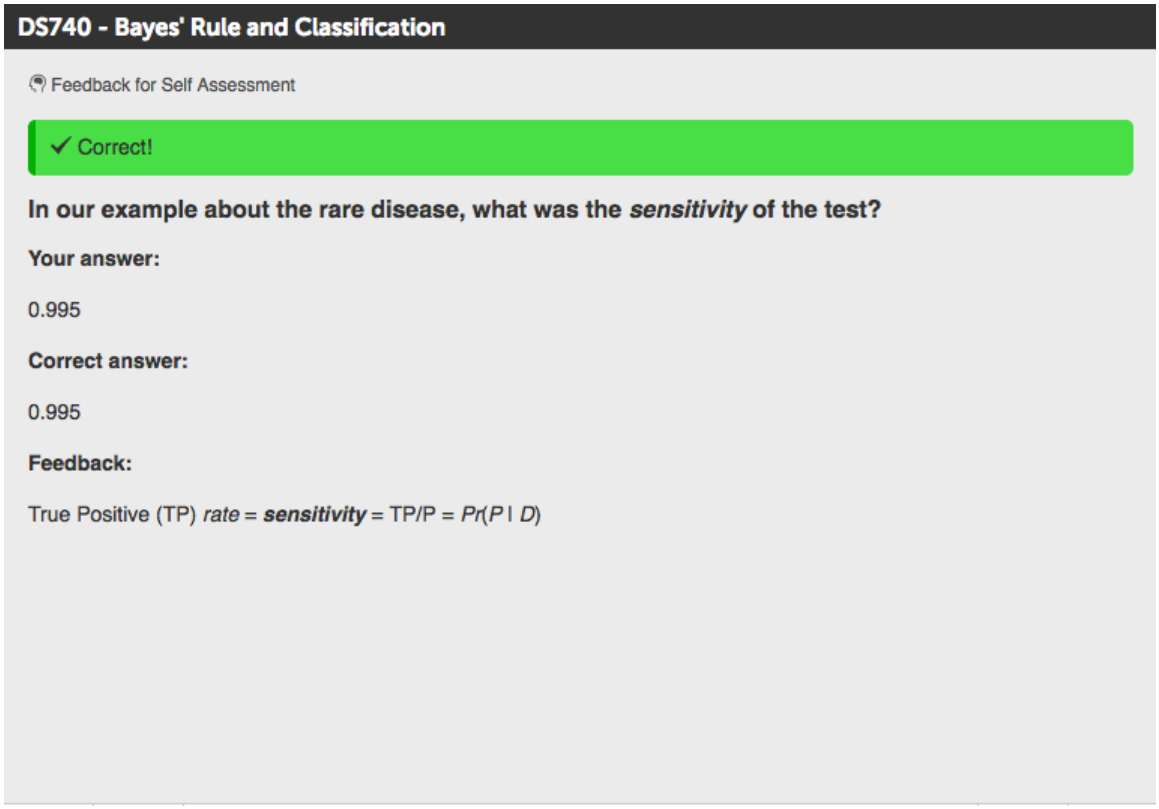
Correct answer:

0.02

Feedback:

Since $Pr(M|D') = 0.95$, then $Pr(P \mid D') = 1 - 0.95 = 0.05$.

Question 2 Answer



Question 3 Answer