Lecture 4

BAYES' THEOREM AND MAXIMUM LIKELIHOOD

From now on, instead of writing $p(A|B)$, I will often leave off the p, and write it simply as $(A|B)$. You can interpret this two ways. You can say I'm changing my notation; since it's always the function p that we're concerned with, I'll simply understand that it's always that function that is meant. Or, since it was an arbitrary function anyway, you can say that I've now adopted the convention that

$$p(x) \equiv x$$

by definition. It will make no difference at all which way you interpret this. Our fundamental rules of reasoning will then take the form:

Rule 1: $$(AB|C) = (A|BC)(B|C) = (B|AC)(A|C) \tag{4-1}$$

Rule 2: $$(A|B) + (a|B) = 1 \tag{4-2}$$

Rule 3: $$(A+B|C) = (A|C) + (B|C) - (AB|C) \tag{4-3}$$

Rule 4: If $\{A_1 \ldots A_n\}$ are mutually exclusive and exhaustive, and B does not favor any over any other, then

$$(A_i|B) = \frac{1}{n} \quad , \; i=1, 2, \ldots n. \tag{4-4}$$

4.1 Prior Probabilities.

Now out of all the propositions that this robot has to think about, there is one which is always in his mind. By X I mean all of his past experience since the day he left the factory to the time he started reasoning on the problem he's thinking about now. That is always part of the information

which is available to him, and obviously it would not be consistent for him to throw away what he knew yesterday in reasoning about his problems today. If human beings did that, education and civilization would be impossible. So for this robot there is no such thing as an "absolute" probability. All probabilities are conditional on X at least. X might be irrelevant to some problem and in that case this postulate would be unnecessary, but at least harmless. If it's irrelevant, it will cancel out mathematically. Any probabilities which are conditional on X alone we will call prior probabilities. If there is any additional evidence in addition to X, which the robot is now reasoning on, we will sometimes leave off the X. We'll understand that even when we don't write X explicitly, it's always built into all expressions:

$$(A|B) \equiv (A|BX) \quad .$$

But in a prior probability, I'll always put in X explicitly:

$$(A|X) \quad .$$

Because of some strange things that have been written about prior probabilities in the past, we have to point out that it would be a big mistake to think of X as some sort of hidden major premise, some universally valid proposition about nature, or anything of that sort. X is simply whatever initial information the robot had available up to the time we gave him his current problem. When we consider applications, you can think also that X stands for some set of hypotheses whose consequences we want to find out, plus the general conditions specified or implied in the statement of the problem.

4.2 Bayes' Theorem.

By far the most important rule which this robot uses in his everyday tasks is the one we get by dividing through the second equality of Rule 1

by, say, $(B|C)$:

$$(A|BC) = (A|C)\,\frac{(B|AC)}{(B|C)} \tag{4-5}$$

This is called Bayes' theorem, or the principle of inverse probability. You see it represents the process by which the robot learns from experience. He starts out with the probability of A, on the basis of evidence C; he is given new evidence B in addition, and this theorem tells how the probability of A changes as a result of this new evidence. Bayes' theorem comes from the fact that Rule 1 was symmetric in propositions A and B, which of course it had to be in order to be consistent. To this robot it is quite clear that if he wants to make any judgments about the truth of proposition A, the only correct way to do this is to calculate the probability of A, conditional on all the evidence he has. This will almost always mean that he will have to use Bayes' theorem.

Now let's imagine we let this robot examine some procedures that are used in statistical inference. A very large part of statistical inference is taken up with problems in which we are given certain evidence, which is typically the result of some experiment, and from this evidence we are supposed to do the best job we can of estimating some unknown parameter, or testing one hypothesis against another. All of these represent plausible reasoning on the basis of new evidence; the evidence of the experiment. Therefore, to our robot it's perfectly obvious that any such example of parameter estimation or hypothesis testing must be a special case of the application of Bayes' theorem. You see, his brain has been built so that this is the only possible way he can reason. To him, the fact that all these procedures must derive from Bayes' theorem is just as much a necessity of thought as the validity of a strong syllogism is to us.

Although this conclusion about Bayes' theorem is obvious to our robot,

it has not been at all obvious to most human statisticians. They largely regard Bayes' theorem as not having any logical basis except in the case where every probability in it can be interpreted as a relative frequency in some "random experiment." In that case, Bayes' theorem can be interpreted as selecting out of an original population of events some subpopulation in which the frequency of event A might be different from the frequency that it has in the population as a whole. But to the robot this is the only possible way of reasoning regardless of whether you can give the probabilities a frequency interpretation.

To a statistician of the "orthodox" school of thought, to be defined more completely later, the first thing he must do in solving a problem is to decide which quantities are "random," and which are not; the procedures he will use, and the whole way he will set up the problem, depend on which decision he makes. But our derivation of the rules for plausible reasoning in the last Lecture made no reference whatsoever to any random experiment. To the robot, therefore, whether any random experiment is or is not involved in the problem is totally irrelevant to the question of how he should reason.

Since this is perhaps <u>the</u> crucial issue in the controversies about probability theory, and the central point in most of the applications that I want to talk about later, we have got to meet it squarely right now. So let's ask the robot to make a strong, definite, and constructive statement about it. Here's what he has to say:

"Consider any procedure in statistical inference in which we reason on the basis of new information. If this procedure is fully consistent and in full qualitative agreement with common sense, then it is necessarily <u>exactly</u> derivable from Bayes' theorem. Conversely, if it is found to represent only some approximation to Bayes' theorem, then it follows that

(1) It is either inconsistent or it does qualitative violence to common sense, or both;

(2) These shortcomings can be exhibited by producing special cases; and

(3) Bayes' theorem will then represent a superior (and often simpler) way of handling the problem."

That is what the robot says. We've designed him in just such a way that it's the only thing he can say. It doesn't mean at all that what he says is right. We've got to put him to the test. For each particular procedure, this is a definite issue of fact; and not a vague matter of personal taste. Either the robot is right or he's wrong in the above statement, and it's in our power to find out whether he's right or wrong. So we'll browse through the statistical literature, and every time we see an example where the man says, "I'm not using Bayes' theorem," then we can look at it a little more carefully and see whether what he actually does can be derived from Bayes' theorem; and if not, whether we can exhibit the defects in his procedure.

4.3 Maximum Likelihood.

The first example is Sir Ronald A. Fisher's method of maximum likelihood. This is a way of estimating an unknown parameter, and I'll illustrate it by the problem of estimating the magnitude of a signal which is obscured by noise. You might be interested in some quotations from Fisher's book (Fisher, 1959). On page 9, he refers to "...my personal conviction which I have sustained elsewhere, that the theory of inverse probability is founded upon an error, and must be wholly rejected" (inverse probability and Bayes' theorem are the same thing as far as we're concerned). And later on he says on page 20 that "maximum likelihood has no real connection with

inverse probability." Well, let's illustrate the method. Suppose we have observed a voltage just at one instant, which is the sum of an unknown signal plus an unknown noise:

$$V = S + N \tag{4-6}$$

Our prior knowledge about the nature of the noise can be described by some probability distribution; the probability that the noise amplitude is in the range dN is

$$(dN|X) = W(N)dN \tag{4-7}$$

Now if we knew that the signal had a certain value S, then the probability of observing a voltage in the range dV would be given by some relation of the form

$$(dV|SX) = L(V,S)dV \tag{4-8}$$

where L(V,S) is called the likelihood function. In the present case, from the linearity of Eq. (4-6), this must be just the probability that the noise would have made up the difference; and so

$$L(V,S) = W(V-S). \tag{4-9}$$

But in the given problem, it's the voltage that's known and the signal that's unknown. The maximum likelihood estimate of the signal magnitude would then be the value of S which renders this likelihood function L an absolute maximum for the observed value of V:

$$\frac{\partial L}{\partial S} = 0 \quad , \qquad \frac{\partial^2 L}{\partial S^2} < 0. \tag{4-10}$$

Stated intuitively, the maximum likelihood estimate is the value according to which the observed voltage would appear as the least remarkable coincidence.

How would our robot go about handling this problem? To him the way of reasoning about the unknown signal is, of course, to calculate the probability that the signal has a certain amplitude, on the basis of all the avail-

able evidence. In other words, the robot says we should calculate $(dS|VX)$ by Bayes' theorem:

$$(dS|VX) = (dS|X)\,\frac{(dV|SX)}{(dV|X)}$$

$$= A\ (dS|X)\ L(V,S) \qquad (4\text{-}11)$$

where A is independent of S. So if we ask the robot what is the most probable value of the signal [more precisely, for what value of S is it most probable that the signal lies in the interval (S, S+dS) for a fixed dS], he will maximize not L but the product of L with the prior probability. So you see that if the robot's prior information didn't give him any reason to expect one signal magnitude more than another [i.e. if the prior probability $(dS|X)$ is independent of S in the range of interest], then the robot's estimate would be the same as the maximum likelihood estimate. If the robot has prior information about the signal, then of course he may easily get a very different value.

Now I think it's obvious not only to the robot, but also to us, that if we do have any prior information about the signal, then it would be screamingly inconsistent for us to refuse to take that information into account in estimating the magnitude of the signal. You see, we could describe the maximum likelihood estimate in another way as the value which we would obtain by <u>throwing away</u> all the prior information we had about the signal, and basing our estimate <u>only</u> on our prior information about the noise.

Suppose you went to a doctor and described your symptoms, and you wanted him to diagnose what was wrong. You tell him that when you raise your left arm you feel a pain in your right side and a few things like this, and the doctor is supposed to do some plausible reasoning to figure out what could be causing it. Suppose that after consultation had been underway for some time you notice that the doctor is not showing any interest in your

previous medical history. You ask him, "Well, aren't you going to look up my medical history?" And suppose the doctor said, "Why, no, I must not look at your medical history, because that would introduce a bias into my conclusions." What would you say? You'd say that the man is crazy. He shouldn't be allowed to practice medicine. To refuse to take the prior information you have into account in plausible reasoning, is not a consistent way of doing things.

Now, of course, a human statistician who uses maximum likelihood has just as much common sense as anybody else; and in a case where we do have prior information which is clearly relevant to the problem, common sense will tell all but the most pedantic not to use the method of maximum likelihood. In practice, he will avoid the bad errors of reasoning by inventing a different method when a different kind of problem comes up. In other words, he will use his prior information to tell him how to formulate the problem,* and he prefers to formulate it so this information no longer appears explicitly in his equations. The robot, however, doesn't need to invent a new procedure for every new kind of problem. To him, Bayes' theorem is always the only way of doing it.

I don't want to go into more details now because this is close to a problem which we are going to talk about a great deal later on; but for the present we'll just note that the robot's prediction was correct. Except in the case where it's clearly inconsistent, the method of maximum likelihood is exactly derivable from Bayes' theorem. After all polemics, there remains the simple fact that, mathematically, it is nothing but Bayes' theorem with uniform prior probability.

---

*An example of such a reformulation suggested by prior information is given in Lecture 9, Equations (9-18)-(9-22).