IPL

INDUCTION
and the PROBLEM of LEARNING

Induction and th. problem of learning.

Gen. outline of presentation, and abstract, emphasizing positive results.

1. Outline normal learning and induction process

a) Selection of ~~axxx~~ classes

b) case counting

c) combining data of ~~xx~~ case counting (B.G.)

d) modifications on case selection.

---

1) First describe ~~[illegible]~~ th. particular learning problem, telling about i.p. (?), describe just what is presented to th. machine, and just what is expected of it.

2) Show that what th. machine must do, is a prediction problem, and is a special case of th. d.d.t.s. Also Mention that most, if not all learning problems ~~are~~ expressable as a t.s. prediction.

Mention that approach to mechanics of machine behavior will be attempt at recreating in it, what I believe to be close to my own mental processes in solving th. problem.

3) Give a few cases of ddts prediction, using ngms.

4) " cases in which gps must be formed, because of small sample size (say, usually zero sample size).

5) Dwell on necessity of forming good gps. — how ad-hock gps. usually result in poor prediction.

6) How th. B.G. problem arises. Note that I will not go into B.G. — that I think a very poor approx. soln. is adequate.

Since one is, in learning problems gn. a history of correct stim-responses, and is gn. a new stim. and asked for a response. Another learning problem: gn. pairs of S.R's with their G's, gn a new S, to devise a R of maximum expected G.

U avail

7) Some ways in which gps are formed. Some gp. parameters (U, sample size, freq. ≡, U may have several components — depending on sample size available). Factoring of Gps ≡ into atpst, strsts

8) Why U is imp. and some ways in which ~~how~~ it may be evaluated

9) Th. specific problem; how $=, \sim, \beta\Sigma, \beta\Pi, +$, (with carry line, then without carry line in linear notation) is accomplished.

10) Describe present difficulties, future expectations.

Rewrite this outline, changing th. ordering somewhat.

1) should come (after ~9)

(2)

1. Describe some learning problems, then try to generalize to characteristics of all/or many learning problems.

Some cases

(1) Simple rote learning: e.g. a) spelling b) Arithmetic tables.

(2) Problems involving somewhat more complex abstractions:
a) Arithmetic addition (without carry line) Mult, & sub. — depending on how taught. b) algebraic or trigonometric simplifications c) soln. of alg. equs. d) ~~[illegible]~~ literal integration. e) interpolation, extrapolation of functions. f) Proofs [involving th. idea of searching]

[SN] Trouble is, it is not clear from th. above that th. machine would ever solve any really difficult, or interesting problems. However, After some of these mathical probs. it mite be possible to get th. machine to answer questions in English. Note that th. math T.M. was undertaken mainly as a method of studying th. abstraction process, and not ~~as a goal in itself~~

Th. point is, most learning problems can be presented as extrapolation, or induction problems.

2. How these are examples of prediction: that a time series is ~~[illegible]~~ one of th. more general kinds of prediction problem. A t.s. may be continuous or discrete, or a mixture. Th. d.d.t.s. is of most interest, and is most difficult.

3: Th. ngms, and ngmst as used in ddts prediction. The problem of specificity v.s. sample size.

4. How ~~[illegible]~~ gps ($gp \equiv$ ngmst) are nec. to give sample size > 0.

4. How good gps. give good prediction — how ad hoc gps. are O.K., but of little value in prediction.

5. How B.G. ~~[illegible]~~ arises (th. "combination problem") — That it probably isn't an imp. problem, in th. sense that a poor soln. is probably O.K.

.7 Some ways that gps are formed — factoring them into strs, strsts, ntpsts. Some dimensional examples of strs, ntpsts. Some gp params: U, samp. size, freq., aprip.

.6 Why U is imp.: How it corresponds to th. elevation of a gp. to a "word".

8. Describe th. specific problem of Math T.M. $= \sim, \beta \leq, \beta\Pi, +$ (with carry line)

9. Show some of th. gps, strs, utpsts that have been useful.

10. Give some future problems, e.g. mult., literal alg., solving eqns., differentiating, Integration, iteral and numerical soln. of diff. eqns., proofs, etc.

Th. problem of getting T.M. to understand English, and how "concept formation" may be induced by R.W. continuous problems.

11. Index.

1. Many learning problems can be looked upon as problems in induction, or extrapolation. One is given several solved examples similar to the problem of interest, and then one is expected to solve it. I believe that in general, the problem of induction, contains most of the significant, difficult problems of learning, and much can be learned about learning by studying the induction process. An aspect of importance in many learning processes, that does not seem to appear in ~~many~~ most induction processes, is the concept of the "sub-goal". I believe, however, that the concepts useful in solving induction problems will also be adequate in the suggesting of trial sub-goals, in certain learning problems.

* Some clarification of learning problems with sub-goals, and how abs. process would be helpful. see (4.03

2. Some examples of learning are as follows.

a) simple rote learning, as spelling or arithmetic tables. In th. former, a spoken word is presented along with its graphic representation. The learner must then reproduce the graphic representation, after hearing the spoken word. Similarly with arithmetic table ~~learning as 3+4=7.~~

* Th. above example is very poor. The true learning process in spelling e.g. is much more complex, and has a longer history of cases than th. single case ~~given. Similarly in Arithmetic table learning.~~

~~b) A very general kind of arithmetic learning problem is described on Page~~

c b) Trigonometric identity proving and theorem proving. These involve search processes in which no specific exhaustive routine is taught. The general technique is obtained by the student thru induction.

-03 * Also poor example, since they involve sub-goals (usually) and we may not want to deal with such problems. Actually, ~~th~~ restriction of induction to problems without sub-goals, is a bit artificial. I must investigate ~~th~~ ~~[illegible]~~ θ process of sub-goal formation in these induction problems ~~see [illegible]~~ →

d c) ~~[illegible] Word and langu~~

Initial word and language learning by a child. The child notes that certain sound sequences are associated with the presence of certain situations. The process by which th. child is able to "name" a situation (e.g. presence of mother ~~[illegible]~~ or father → "Mama", "Dada") is a rather complex induction ~~[illegible]~~ problem. The conclusion, by the child, that trying to imitate his parents' sounds in these situations, is desirable, is an induction process.

* The above paragraph is a rather questionable way of looking at the phenomenon discussed. 'Imitation' (at least, many people will question it.) may be a built-in response, requiring no reinforcement. However, even if it is built-in, this simply gives th. child's conclusions by aprip., and so he must still use induction.

* ~~[illegible]~~ Th. formation of sub-goals in th. soln. of "search" problems, is a technique T.M. can discover for himself, or it can be taught to him by a tng. seq. designed for such a discovery. However, th. exact abstractions made in this inductive process, and their plausibility — i.e. how readily they stem from other abss., must be looked into. Th. formation of sub-goals may ~~b~~ involve an induction technique of a significantly different ("hyper") type than other previous inductions. I think we may, however, be able to take this in our stride.

e d) Learning problem involving maximizing reward recieved for ~~respon~~ behavior. The organism is given a stimulus, S_i, and it presents a response, R_j to the environment. The environment presents a reward G_{ij} to the organism. After many S_i, R_j ~~pairs~~'s are tried, and G_{ij}'s recieved, the organism tries to find R_j's such that the G_{ij}'s they will recieve are maximal. The pure induction problem involves predicting the expected ~~This problem is more complex~~ value of G_{ij} from a S_i, R_j pair.

* This problem is more complex than the simple induction problem. The organism is always faced with the occasional choice of doing an experimental ~~S_i~~ R_j to gain information, rather than to try to get an immediate large G_{ij}. Also, as the problem has been formulated above, even if one <u>could</u> predict the expected G_{ij} from any S_i, R_j pair, ~~one still therefore~~ one still has the search problem (of the 2nd kind), that ~~wants~~ for fixed S_i, we must find a R_j such that G_{ij} is maximal. In a well-designed machine, the soln. to the problem may not proceed in the step-wise manner as above. i.e. finding a formula for an expected G_{ij} from an arbitrary S_i, R_j pair, then optimizing R_j for fixed S_i. The machine may proceed <u>directly</u> to find ~~optimum~~ a method to obtain ~~R_j~~ optimum R_j, given S_i.

In such a case, it is not clear that we have a simple induction problem.

We <u>can</u>, however, separate the problem as given, into an induction-problem, plus a search problem.

f e) The Problem of a scientist. One of the problems of a scientist, is to predict what will follow from arbitrary initial conditions. He ~~does this~~ attempts to do this by observing many cases and drawing inductive conclusions. He may or may

~~2. The above examples are all cases of prediction.~~
~~As such, they~~
not draw up a formal "law of nature" to aid in the extrapolation. In either case, the process may be looked upon as either induction or learning.

~~[illegible]~~

2. The above ~~cases~~ examples ~~are~~ all involve ~~cases of~~ prediction. As such, they ~~can be looked upon as examples of time series.~~ all bear a strong structural resemblance to time series'. We shall

* The formal resemblance between time series' and other prediction problems should be clarified. i.e. A time series problem usually involves prediction of the next element, ~~when~~ following a long known sequence. In many prediction problems, however, ~~the problem is to predict~~ one is given many members of the series, and then one is given a member that is incomplete, and asked to complete it. e.g. one's Given many correct, completed arithmetic addition problems, then one is given 1 + 7 = . The problem is to complete the ~~[illegible]~~ equation.

Actually, ~~[illegible]~~ all of the induction problems, including time series' are special cases of the following: Given a large object ~~[illegible]~~ that may be composed of many related or unrelated parts: Suppose that part of this object is unknown to you — the problem is to predict just what it is.

At any rate the concepts ~~[illegible]~~ useful in time series prediction are also useful in the general induction problem.

discuss time series' in some detail, since ~~the concepts used will be useful in the the easy~~ it is easy to find simply describable examples, that ~~have many does~~ exhibit properties that are very useful in discussing the general induction problem.

Time series' are of several types. One method of classification is on the basis of whether time ~~a) Continuous in time, with continuous variation in parameters of members.~~

~~b) Discrete~~

is a discrete or continuous variable, and another ~~is~~ depends on whether the elements of the time series have discrete or continuous values of the parameters that describe them.

'Examples:'

a) Discrete time, discrete elements: Printed English, Morse code signals.

b) Discrete time, continuous elements: Peak daily temperature readings.

c) Continuous time, discrete elements: stock market readings

d) Continuous time, Continuous elements: most Sound or electrical signals; noise.

The type of series' we will study most extensively is type a): discrete time, discrete elements. The limitations of this choice will be ~~dealt with later.~~ It is believed that the

* This limitation is in the "Real world to-language conversion problem, and the evolution of spacial concepts" in an organism.

> doubley' discrete time series is by far, most difficult to deal with, has had little work done on it, and is most important, in terms of ~~...~~ the lite that it throws on other ~~...~~ induction problems.

3. Let use consider the problem of predicting printed English. The concept that I would like to use is that of the "n-gram" / e.g. as used by Shannon in PEPE. An n-gram is an ordered sequence of n elements. Elements may be letters, small or large, punctuation and small and large spaces. In printed English we may have on the order of 60 different possible "elements".

Suppose we have a long English text, ~~[illegible]~~ that is abruptly truncated and we want to predict the next element. For a rough approximation, we can make a frequency count of each of the 60 elements, and make a set of probabilistic predictions ~~directly~~ based ~~on~~ directly upon them.

For a more accurate prediction, we may note the ~~[illegible]~~ element with which the truncated text ends, and make frequency counts for all digrams that begin with that ~~[illegible]~~ element. We may on this basis, ~~[illegible]~~ usually make a more accurate prediction.

In a similar way, we may look at the last n-1 elements of the text, and ~~count in~~ make frequency counts of n-grams that begin with those n-1 elements.

AS n increases, we might expect some increase in prediction accuracy, providing the text was long enuf to give a large enuf ~~statistical~~ sample of these n-grams.

~~* Shannon has shown, however, that prediction accuracy increases only very ~~but~~ slowly for n > say, 100~~

Unfortunately, however, since our text is of finite length, n cannot be very large, or else the advantages of good predictions thru large n, will be overbalanced by the ~~small sampl~~ low

insert * > accuracy due to small sample size.

Let us examine this situation more carefully and try to generalize. In any prediction problem we are confronted by a set of conditions ~~that~~ ~~we~~ from which we try to make a prediction. Ideally, this set of conditions will have occured many times before, and accurate probability estimates can be made. More often, however, the set of conditions has never before occured in exactly the same way. In such a case, we try to classify the event to be predicted, within a larger class of events, of which we have a sufficiently large sample, so that relyable probability estimates may be made.

insert: ~~E~~ (α), page 8.

~~[illegible]~~ Suppose our text ended with the ~~[illegible]~~ elements $\ldots e_4\ e_7\ e_2\ e_3$, ~~and that the rest~~

We may represent the rest of the previous text (which contains r elements) by A_r, so the entire truncated text becomes $A_r\ e_4\ e_7\ e_2\ e_3$. We wanted to know ~~When we made frequency counts on the elements e_i ($i = 1, \ldots 60$)~~ the relative probability of the various sequences $A_r\ e_4\ e_7\ e_2\ e_3\ e_i$, for all 60 values of i.

When we made our predictions on frequency counts of the e_i elements only, we essentially inserted the sequence $A_r\ e_4\ e_7\ e_2\ e_3\ e_i$ in the set of all $r+5$-grams that end in e_i.

In a similar way, when we made our predictions by making frequency counts on the ~~trigram~~ tetragrams, $e_7\ e_2\ e_3\ e_i$, we inserted our sequences in the sets of all $r+5$-grams that end in $e_7\ e_2\ e_3\ e_i$.

It should be noted that usually there are fewer $r+5$-grams that end in $e_7\ e_2\ e_3\ e_i$, than there are $r+5$-grams that end in e_i.

---

* Actually, we shouldn't let $A_r\ e_4\ e_7\ e_2\ e_3$ be the entire truncated text. This tends to make sample sizes too small. Better let $r = 0$ or some small number. — No, on 2nd thot, this use of $A_r\ e_4 \ldots e_3$ is O.K. — perhaps it should be added, for clarity, that one knows that th. truncated text is at least $2(r+5)$ elements long, but one only knows the last $r+5$ elements.

Loosely speaking, we try to classify the event in a class as "close" to it as possible, yet which has ~~[illegible]~~ happend sufficiently frequently in the past so as to give it relyable statistics. In general, we must compromise between an extremely "close" class with few members, and a desire for a class that has occured many times in the past, that often has many members, and is not as closely related to the event of interest.

The problem of ~~[illegible]~~ assigning events to classes, ~~[illegible]~~ and inventing useful classes, will be the main problems to which we will direct our selves ~~in this presentation~~. In addition, we shall concern ourselves with the assignment of parameters to these classes, so that they might be used in prediction.

As might be suspected, the ideal solution to the problem of classification of ~~events~~ an event, lies in its assignment to all classes that contain it. The prediction problem involved is rather complex, however, and usually only the roughest approximations to the ideal solution will be attempted.

4. The devising of ~~[illegible]~~ suitable classes is of most importance in prediction, and prediction ~~[illegible]~~ effectiveness will depend upon how well one has chosen them. For instance if one is accosted by a grizzley bear, one's prediction of ~~[illegible]~~ its future behavior would be best, if one catagorized it ~~the event with~~ as a grizzley bear, rather than ~~with~~ as a mammal.

The confidence that one has in a classification will depend upon how the class was constructed, how many times events have been observed in the class, how many members there are in the class, ~~[illegible]~~ and ~~[illegible]~~ several other factors.

Usually one tries to chose a class ~~as~~ containing as few members as possible, ~~yet has [illegible] occured reasonabley frequently~~ yet whose members have occured with reasonable frequency. Also, it is usually expedient that the description of the class ~~in a language that has been~~

be "simple." This "simplicity" will be with respect to the definition of the class within a language that has proved useful in the past. ~~[illegible]~~ ~~[illegible]~~ the simplicity of class description is formally equivalent to ~~[illegible]~~ a scientific methodology rule of thumb called "Occam's Razor". These concepts will later be used as a guide ~~[illegible]~~ to the more ~~[illegible]~~ precise formulation of class prediction parameters.

Another example of a very poor kind of class, is the ad-hock class...... Suppose we want to predict the next ~~[illegible]~~ element of a sequence ~~[illegible]~~ that terminates in $a_1\, a_3\, a_1\, a_2$, and there are only 3 different kinds of elements; a_1, a_2 and a_3. What we want, is the relative frequencies of the 3 ~~[illegible]~~ sequences

$$A_1 \equiv a_1\, a_3\, a_1\, a_2\, a_1$$
$$A_2 \equiv a_1\, a_3\, a_1\, a_2\, a_2$$
$$\text{and } A_3 \equiv a_1\, a_3\, a_1\, a_2\, a_3.$$

If none of these sequences have occured sufficiently frequently to give a good probability estimate, we must express each of them as members of a ~~[illegible]~~ ~~[illegible]~~ set of sequences, such that the ~~[illegible]~~ ~~members have sufficiently frequently~~ total frequency of occurance of members of each of these sets is sufficiently high. A possible set would, as was suggested before, be the placing of $a_1\, a_3\, a_1\, a_2\, a_1$ in the set of all pentagrams that end in a_1.

Suppose, however, that we place $a_1\, a_3\, a_1\, a_2\, a_1$ in the set of pentagrams whose members are, ad-hock, defined to be $a_1\, a_3\, a_1\, a_2\, a_1$ ~~[illegible]~~ and $a_2\, a_2\, a_2\, a_2\, a_2$

Similarly, we can form the class of members

$a_1\, a_3\, a_1\, a_2\, a_2$
and $a_3\, a_3\, a_3\, a_3\, a_3$

and the class of members $a_1\, a_3\, a_1\, a_2\, a_3$ and $a_1\, a_1\, a_1\, a_1\, a_1$.

This "ad-hock" discu. may be an unnec. digression

Suppose, further, that $a_2 a_2 a_2 a_2 a_2$, $a_3 a_3 a_3 a_3 a_3$ and $a_1 a_1 a_1 a_1 a_1$ are all pentagrams of rather hy frequency of occurance — much hyer than A_1, A_2 or A_3. Our class frequencies would then be largely governed by the frequencies of the abitrarily chosen sequences and to very little extent by the frequencies of the sequences A_1, A_2 and A_3. ~~The reason this method of prediction would not be used, is that the 3 ad-hock classes would be given~~

Fortunately our final prediction of the probability distribution of the next element, would be determined by the membership of A_1, A_2 and A_3 in several classes other than the ad-hock classes. In particular, the classes which were constructed in "acceptable ways", would be given much more weight in the prediction, than would the arbitrarilly constructed ad-hock classes.

A simple example of a set which is constructed in an "acceptable way" is "the set of all pentagrams that end in $a_1 a_2 a_1$."

Various methods of set construction, and their degree of "acceptability" will be dealt with at some **length.**

5. A problem of much importance, is that of combining data due to the membership of an event in several different sets. For example, fighter A has won 80% of his matches, ~~fighter B has won and~~ and fighter B has won **70%** of his matches. Now A is matched against B. What is the best way that we can use the available data to determine the probability that A will beat B?

Altho this problem is interesting, difficult, and important, I shall not dwell up it at this point. Although a tentative solution has been obtained, I think that it is very probable that an extremly rough approximation will be sufficient to achieve quite useable results.

6. A parameter of sets that is of much importance in prediction, is the "Usefulness" of the set. This parameter ~~describes~~ tells how much more effective prediction is when one uses this set, that it is without the use of that set. In order to evaluate usefulness, it is necessary to evaluate the effectiveness of a prediction run. Any of several methods may be used to do this. Probably the methods devised by J. McCarthy in "Paying the weatherman" are more than adequate.

Sets with high Usefulness values might be thot of as corresponding to often used words or phrases in a language. In the problem of combining statistics of events that belong to several different sets, usefulness of a set is an important parameter in determining how much weight it is to get.

7. At this point we will give some examples of important kinds of prediction sets, and how one may obtain new ones that have a reasonable likely hood of being useful, by suitable operations upon old useful sets. On a more intuitive level, this amounts to combining old scientific concepts and theories to obtain new theories that have a reasonable ~~likelyhood~~ apriori probability of usefulness.

To give these examples, we will investigate a particular induction problem that seems to illustrate them.

~~Consider the following~~ The problem consists of presenting the machine with correctly worked examples of arithmetic problems. After the machine has been given many such examples, it will be presented with a problem in which some of the ~~figures~~ digits are missing, and it will be asked to find the probability distributions for the missing digits.

Altho this approach has been applied to arithmetic, for the most part, it is felt that most, if not all, mathematical problems, could be presented to the machine in this form.

The first kind of problem we will present will be equality. The ~~computer~~ machine will be presented with the following set of examples:

$$= 1001 \quad , \quad = 1010 \quad , \quad = 0110 \quad , \quad = 0001$$
$$ 1001 \quad \quad 1010 \quad \quad 0110 \quad \quad 0001$$

etc. We present the machine with a ~~series~~ set of examples in which the first line starts with =, and is followed by a random sequence of 4 0's and 1's. The next line contains a duplication of the sequence of 0's and 1's.

We then present the machine with a problem as

$$= 1101$$
$$1\boxed{?}01.$$

We want the machine to reply by giving the probability distribution of the various elements that mite be in the square marked $\boxed{?}$

A first approximation by the machine would simply give the relative frequencies of the four different possible elements 1, 0, =, space. If the examples were presented on a 7 x 10 array, each example would have one equals sign, on the average of 4 0's and 4 1's, and 61 spaces. The relative probabilities as computed from this data, would be:

$= \ : \ \frac{1}{70}$

$0 \ : \ \frac{4}{70}$

$1 \ : \ \frac{4}{70}$

space : $\frac{61}{70}$.

We note that in our own thinking about the problem, we would not give such hy probability to the space. This is because we have some reason to beleave that the 0's, and 1's are of more interest than the spaces, and are more likely to be the rite answer.

If we had asked the machine many

questions in the past, and had given him the correct answers, he would, indeed, soon find out that most of the answers were zeros or ones..

We can present this information to the ~~computer~~ machine by making every example have a ? space in it, and then indicate later, to the machine, what the rite answer is. For a machine of a deterministic kind, in which we know its internal workings, there is no essential difference between ~~[illegible]~~

a) Asking the machine a question, letting it guess at the answer, and then telling it the correct answer

or b) Giving the machine the correct answer to a question, then telling it what the question was.

For this reason, we will present each example as if it were a question, and we will indicate the correct answer. This can be done by giving our examples in the following form:

$$\frac{=1001}{1\boxed{0}01},\quad \frac{=1010}{101\boxed{0}},\quad \frac{=0\boxed{1}10}{0110},\quad \frac{=0001}{000\boxed{1}},$$

etc.

Here we have indicated at what position the question would have been asked, and we have indicated the correct answer.

This device is used to direct the machine's attention to the factors of interest. In the real world, ~~this is done by learning from experience~~ one learns that the factors of interest in examinations tend to be the factors that one was asked questions about in previous examinations. The above device is an attempt to simulate this situation. It is felt that it will be possible, later in the life of the machine, to discard this device, and the machine will be able to decide by itself, what factors are likely to be of importance. In such a case, the machine would, on the average, gain more information from an example than at present. This is because usually there are many squares in an example that would be equally likely candidates for interrogation with respect to probability distribution of digits.

If we do not ~~[illegible]~~ put our "interrogation square" around the = sign, the relative element frequencies will be 50% for 0 and 50% for 1.

These ~~frequencies~~ probabilities were obtained in a manner analogous to ~~[illegible]~~ direct frequency counts of characters in printed English. We shall now employ the analogs of n-grams, as used in Shannon's "Prediction and Entropy of Printed English".

We will first observe ~~[illegible]~~ the frequencies of oriented pairs of digits that include the interrogation square. If we observe n examples, ~~the~~

~~digram 1 $\boxed{1}$ occures $\frac{n}{4}$ times~~

~~the digram~~ ~~0 $\boxed{1}$ no times~~

~~space $\boxed{1}$ $\frac{n}{4}$ times~~

~~1 $\boxed{1}$ $\frac{n}{8}$ times~~ ~~0 $\boxed{1}$ $\frac{n}{8}$ times~~ ~~$\boxed{1}$ space $\frac{5n}{8}$ times~~

the following digrams will occure with the noted frequencies:

digram:	1 / $\boxed{1}$	0 / $\boxed{1}$	space / $\boxed{1}$	1 / $\boxed{1}$	0 / $\boxed{1}$	space / $\boxed{1}$	0 / $\boxed{0}$	$\boxed{1}$ / 1	$\boxed{1}$ / 0	$\boxed{1}$ / space	$\boxed{1}$ / 1
no. of times it occures	$\frac{n}{4}$	0	$\frac{n}{4}$	$\frac{3n}{32}$	$\frac{3n}{32}$	$\frac{5}{16}n$	$\frac{n}{4}$	$\frac{n}{4}$	0	$\frac{n}{4}$	$\frac{3}{16}n$

Suppose we are now given the problem:

$$\begin{array}{c} = 1\,1\,0\,1 \\ 1\,1\,\boxed{?}\,1 \end{array}$$

If we use the digrams 0 / $\boxed{1}$ and 0 / $\boxed{0}$ we will arrive at the conclusion that the digit in question is 0, with a probability close to unity. On the other hand, there are many other digrams that might, at first glance, be equally likely to give

Suppose we ~~are~~ now given the problem

$$\begin{array}{l} = 1101 \\ 11\boxed{?}1. \end{array}$$

We want to know the relative probabilities of

$$A \equiv \left\{ \begin{array}{l} = 1101 \\ 11\boxed{0}1 \end{array} \right\} \quad \text{and} \quad B \equiv \left\{ \begin{array}{l} = 1101 \\ 11\boxed{1}1. \end{array} \right\}$$

We may classify A is by means of the digram $\overset{0}{\boxed{0}}$. This digram corresponds to the class of all examples that contain $\overset{0}{\boxed{0}}$ in them. Similarly, we may classify B with the digram $\overset{0}{\boxed{1}}$. The first digram would occur about $\frac{n}{2}$ times in n examples, The second digram, never. The probability obtained the relative frequency of for A, would be about unity, for B, about zero.

[SN] It mite be well to include my reasons for chosing ~~...~~ ugms and objects of hy U, to ~~make~~ construct objects of greater complexity, which "hy" expected U: This would involve explaining why Occam's razor is trivial — how new ideas are always made of a "minimal combination" of ~~old useful ideas, since~~ old words. I.E. A theory is "simple" ~~...~~ in str." if it uses few "words" in th. old lang. Th. "words" in th. old lang. are ugnsts that have proved useful.

If we observe n examples, we will note the following approximate digit frequencies, for large n: ["s" denotes a blank space]

Digram Configuration	$\overset{1}{\square}$	$\overset{0}{\square}$	$\overset{s}{\square}$	$\square^{\,1}$	$\square^{\,0}$	$\square^{\,s}$	$\square 1$	$\square 0$	$\square s$
no. of times 0 appeared in interrog. square	0	$\frac{n}{4}$	$\frac{n}{4}$	$\frac{3n}{32}$	$\frac{3}{32}n$	$\frac{5}{16}n$	$\frac{3}{16}n$	$\frac{3n}{16}$	$\frac{n}{8}$
no. of times 1 appeared in interrog. sq.	$\frac{n}{4}$	0	$\frac{n}{4}$	$\frac{3n}{32}$	$\frac{3}{32}n$	$\frac{5}{16}n$	$\frac{3}{16}$	$\frac{3}{16}n$	$\frac{n}{8}$

Suppose we are now given the problem

$$\begin{array}{l} = 1101 \\ 11\boxed{?}1 \end{array}$$

It is clear that we can use any one of the digrams

0 , 1 , ~~s~~ , □ , ⊡ , □ , 1□ , 1
□ , □ , □1 , ~~s~~ 5 , 5 , 5 , □

to obtain the probability distribution for the occupant of ~~the~~ the interrogation square. If we look at our history of examples, however, we will find that only the digram $\begin{array}{c}0\\ \square\end{array}$ has been of any value in prediction.

Footnote or Appendix: The mean value of a series of predictions may ~~will~~ be taken as ~~[illegible]~~

$$\frac{1}{n}\sum_{i=1}^{n} \ln q_i$$

n is the number of predictions,

q_i is the probability given by the predictor, that the event that did, indeed, occure at the i^{th} case in the series, would occure. Other criteria of value of ~~[illegible]~~ a series of predictions may be used. One kind of criteria is suggested by J. McCarthy in "Paying the Weatherman". It is ~~[illegible]~~ possible that the exact nature of the criterion that is used is not very critical.

~~[illegible]~~ This digram has been 100% correct in all of its predictions - which were alway 0. All other digrams gave ~~[illegible]~~ .5 probability for 0 or ~~[illegible]~~ 1, each time.

It is only reasonable that we give the digram $\begin{array}{c}0\\ \square\end{array}$ much weight, in solving the particular example that was given.

In general, if we are given other examples of this simple arithmetic operation, we will find that only the digrams $\begin{array}{c}0\\ \square\end{array}$ and $\begin{array}{c}1\\ \square\end{array}$ are of any value in prediction. We will take ~~[illegible]~~ special note of these digrams and proceed.

Our next series of examples will be

$\sim$ 1001 , $\sim$ 1010 , $\sim$ $\boxed{0}$110 , $\sim$ 0001
0$\boxed{1}$0 , 010$\boxed{1}$, 1001 , 1$\boxed{1}$10