Dept. of Math. University of Oslo
Statistical Research Report No. 4
ISSN 0806–3842 February 2004

Representing Human Uncertainty by Subjective Likelihood Estimates

Fredrik A. Dahl
Department of Mathematics, University of Oslo
P. O. Box 1053 Blindern, 0316 Oslo, Norway
Email: fadahl@math.uio.no

Abstract: *We give a definition of human uncertainty through subjective likelihood estimates. The subject is asked for his estimated likelihood of a discrete variable, given a present piece of uncertain observation, under the hypothetical assumption that the variable was uniformly distributed prior to the new observation. With this interpretation of human uncertainty, we are able to perform consistent inference about our target variable, by formally treating the input as likelihood factors. The algorithm has been successfully implemented in an expert system for classification of wildwood mushrooms.*

Introduction

People frequently make statements like: "I'm 90% sure that the taxi driver spoke Swedish in his cell phone." The purpose of the present article is to give a probabilistic interpretation of statements of this type, so that we can combine them, and produce consistent inference.

A simple approach would be to claim that when a person makes the given statement, he is right 90% of the times. This makes some sense when there are only two alternatives, if we assume symmetry in his errors, so that his 10% error rate applies whether the driver actually speaks Swedish or not. However, we would like to generalize our interpretation to cases with more than two possible answers. Suppose our subject estimates the taxi driver's language to be:
"*90% Swedish, 5% Norwegian, 3% Danish and 2% Icelandic*"
Then the error rate interpretation fails to make sense.

The article is laid out as follows: First we review different established models of human uncertainty. Then we define our model of subjective likelihoods and give a Bayesian inference rule for combining statements. Then we describe an application of the algorithm in an expert system that helps a user classify wildwood mushrooms. The last section concludes the article.

Established models of human uncertainty

In this section we give a broad overview of models that have been used for quantifying human uncertainty.

Certainty factors

In the early days of artificial intelligence, expert systems were built that were imitating human inference (Shortliffe, 1976). The typical expert system consisted of a set of facts, a set of rules, and an inference engine. The inference engine applied a sequence of rules to the set of facts, thereby producing new facts. Uncertainty was modelled through certainty factors associated to facts and rules. Although some expert systems of this kind worked quite well, certainty factors are not popular nowadays, because they tend to produce contradictions.

Fuzzy logic

Fuzzy logic attempts to model uncertainty through vagueness, rather than probabilities. In a fuzzy logic context, our taxi driver example statement would be interpreted as: "On a swedishness scale from 0 to 100, the taxi driver's language was 90." This is an interesting and useful semantic model in many cases, but our goal is to model the fact that the subject's observation may be wrong, not that he is correct to a certain degree. For a discussion on how fuzzy logic relates to probability theory, see (Dubois & Prade, 1997). More fundamental connections between the theories of standard and fuzzy sets are made in (Indahl, 2000).

Dempster-Schaefer theory

In Dempster-Schafer theory, uncertainty is modelled by an interval $(a,b) \subset [0,1]$ (Shafer, 1976). The idea is that the span of the interval reflects the degree of uncertainty. One might e.g. assign [0,1] to a statement of extra terrestrial intelligent life, while the event that the future flipping of a fair coin gives "heads", would have a collapsed interval $\{0.5\}$. The theory gives a consistent calculus for combining statements. It is not readily applicable to our setting, though, because our subject does not convey his uncertainty in the form of intervals.

Lower previsions

The theory of lower (and upper) previsions can be seen as a generalization of D-S theory (Walley, 1996). The lower prevision of a statement can be interpreted as a lower limit of the probability of the statement being true. The theory is related to gambling situations where one assumes that the opponent may have more information than oneself. As an example, you might assign a 0.4 lower prevision on the event that the flipping of a coin gives "heads", if you suspect that the coin may be unfair, but you are sure that even an unfair coin will give heads at least 40% of the time. The theory is by nature pessimistic, as it always works through worst-case values of probabilities. This is good for the purpose of making robust inference, but does not capture the meaning of out taxi driver example statement.

Subjective Probability

A natural interpretation of our example statement is that the subject's subjective probability of the driver's conversation being in Swedish is 0.9. The term subjective probability (as opposed to frequency based probability) means that the subject merely assigns numbers to different events and statements, which obey the rules of probability calculus.

A problem with subjective probabilities is that one cannot easily combine different subjective probability statements in a meaningful way, because the statement is derived from the subject's internal probability model. Suppose we want to combine the given statement with the fact that the event took place in Sweden, we would first need to know whether the subject had already included this important piece of information in his 0.9 probability estimate.

Also, it is very hard for people to produce consistent subjective probabilities in cases where they simply do not know. It is well known that the attempt of assigning uniform probabilities to reflect ignorance fails. The question of extra terrestrial intelligent life is a good example: If you assign a 0.5 probability of extra terrestrial intelligent life in our galaxy, you cannot readily assign the same probability for the left arm of the galaxy, or for entire universe. The difficulty in representing ignorance is a big problem with subjective probability models.

Bayesian networks

In a Bayesian network (Jensen, 1996), nodes represent random variables, which are connected through edges that represent causal relations. When new evidence is presented, probabilities are propagated through the network in a consistent way.

Bayesian networks represent a different perspective than that of classical expert systems: Rather than imitating the human thought process, with uncertainty associated to inference rules, one creates a consistent causal probability model, and uses probability calculus for inference. Under this paradigm, certain and uncertain human knowledge is included in the model of the world, rather than in the automatic reasoning. Hence, Bayesian network modelling does not offer any immediate interpretation of our taxi driver statement, but it gives a framework within which we would like our interpretation to fit.

Subjective likelihood

Our interpretation of the taxi-driver statement, which we introduce in this article, is this: "The probability of me hearing what I heard, if he did speak Swedish, is nine times higher than the probability of me hearing what I hear if he didn't speak Swedish."

With this interpretation, the statement only refers to the *present observation*, not the subject's overall judgment concerning the driver's language. By only referring to the subject's present observation, and not to his personal beliefs about the probability of meeting Swedish-speaking people in this given situation, the statement is made *context free*. This enables us to use it in a formal probabilistic Bayesian model, and combining it with other statements, without worrying about the statement's context.

Now assume that the subject is in Sweden, where the a priori probability of a taxi driver speaking Swedish on the phone is, say, 95%. Then the likelihood of the conversation having been in Swedish is the prior probability of 0.95 multiplied by the observation weight 0.9, while the likelihood of the opposite is 0.05 times 0.1. This gives:

$$P\left[\text{Conversation in Swedish}\right] = \frac{0.95 \cdot 0.9}{0.95 \cdot 0.9 + 0.05 \cdot 0.1} \approx 0.994$$

This high estimate is reasonable, because the conversation both sounded Swedish to the subject, and took place in Sweden.

We formalize this calculation for observations with *n* different values. Let $q \in \Theta = \{q_1, q_2, ..., q_n\}$ be the true state of Nature, and let the prior distribution p be a probability vector of length *n*, so that $p_i = P(q_i)$. Let $\{o_1, o_2, ..., o_n\}$ be a vector of random variables with values in some space *X*. We interpret *X* as the set of possible observations that the subject can make, and the random variable o_i represents the random observation that the subject makes, given $q = q_i$. Assume now that the subject has made observation *x*. We can then calculate the conditional distribution $(p \mid x)$ using Bayes formula:

$$(p \mid x)_i = P(q_i \mid x) = \frac{P(q_i) P(x \mid q_i)}{\sum_{j=1}^{n} P(q_j) P(x \mid q_j)} = \frac{p_i P(o_i = x)}{\sum_{j=1}^{n} p_j P(o_j = x)} \qquad (1)$$

In this formula, we use the standard statistical convention of interpreting $P(o_i = x)$ as a probability if the o_i's are discrete variables, and as probability density (likelihood) otherwise. (In theory, one should link this to the structure of the observation space *X*, but we will return to this below.)

So far, our construction is one of standard Bayesian inference. The next step in an applied Bayesian analysis would usually be to collect data (*x*), and compute $(p \mid x)$, treating the distributions of the o_i's as given. Our approach is simpler mathematically, as we leave the assessment of *o*, *x* and *X* to the subject. We define the *subjective likelihood vector* of observation $x \in X$ by:

$$q = \left[P(o_1 = x), P(o_2 = x), ..., P(o_n = x)\right] \qquad (2)$$

Again, we either treat $P(o_i = x)$ as a probability or a probability density. Note that the observation *x* and the random variables o_i are "private" for the subject, which is the reason why we can disregard the mathematical structure of the domain *X*. For our purpose, a rescaling of the vector *q* is also of no importance, as only the components' relative values affect our computation below.

In a sense, we condition *p* by the vector *q* (which is what our subject reports), so write $(p \mid q)$ instead of $(p \mid x)$. This is a slightly abuse of notation, but we prefer to hide the private variable *x*.

With our definitions, equation (1) now simplifies to:

$$(p \mid q)_i = \frac{p_i q_i}{\sum_{j=1}^{n} p_j q_j} \qquad (3)$$

If the denominator is zero, the observation contradicts *p*, in which case we define $p \mid q = p$, for convenience.

Observe that if $q_k = 1$ for some *k*, in which case *q* represents certainty, then (3) implies $p \mid q = q$. On the other hand, if q is uniform, in which case *q* represents complete ignorance, then $p \mid q = p$.

The (normalized) subjective likelihood vector also has a different but equivalent interpretation of *hypothetical posterior distribution of a uniform prior conditioned by the observation*: Assume $p = \frac{1}{n}[1,1,...,1]$. Then we easily see that $(p \mid x) = \frac{q}{\sum_{i=1}^{n} q_i}$.

Hence, *q* can be thought of as the relative probabilities of the q s, given *x*, assuming a uniform prior distribution.

Underlying model

We now proceed to the more complex case with an underlying model.

Again, let the true state of Nature be $q \in \{q_1, q_2, ..., q_n\}$ with a prior distribution vector *p*. We also have a set of *m* features: $\{F_1, F_2, ..., F_m\}$. Each feature F_j has a domain of n_j feature values: $\{v_1^j, v_2^j, ..., v_{n_j}^j\}$. For each q_i, we have a probability distribution $P_{q_i} = P_i$ over $\prod_{j=1}^{m} \{1,2,...,n_j\}$. We use the compact notation $P_i(j,k) = P(F_j = v_k^j \mid q_i)$, where $1 \le j \le m$ and $1 \le k \le n_j$. Hence, each state of nature q_i gives a probability vector $P_i(j, \cdot)$ over the n_j different values, for each feature *j*. We assume that the *m* feature distributions of P_i are independent.

Let q^j be a subjective likelihood vector for feature *j*. Just like in the previous section, this means that there exist random variables $\{o_1^j, o_2^j, ..., o_{n_j}^j\}$ (corresponding to the n_j different values for feature *j*), with values in some space *X*, such that $q^j = [P(o_1^j = x), P(o_2^j = x), ..., P(o_{n_j}^j = x)]$. Again, we need not worry about what *X* and the distributions of o_i^j look like, because our subject supplies us with q^j directly.

Then the distribution of p conditioned by *x*, through q^j, is given by:

$$\left(p \mid q^j\right)_i = \frac{p_i \sum_{k=1}^{n_j} P_i(j,k) q_k^j}{\sum_{l=1}^{m} p_l \sum_{k=1}^{n_j} P_l(j,k) q_k^j} \qquad (4)$$

Here too, we disregard the subjective likelihood vector, if it contradicts p:

If $\sum_{l=1}^{m} p_l \sum_{k=1}^{n_j} P_l(j,k) q_k^j = 0$, then $\left(p \mid q^j\right) = p$.

Observe that if q^j is uniform, then $p \mid q^j = p$. Again, this means that a uniform subjective likelihood successfully represents complete ignorance, because conditioning by it makes no difference. The intuition behind this is clear: A uniform subjective likelihood means that the subject reports that his observation is equally likely for each possible feature value.

If $q_k^j = 1$ for some k, then q^j represents certainty. In this case, (4) simplifies to standard conditioning: $\left(p \mid q^j\right)_i = P\left(q_i \mid F_j = v_k^j\right)$, which is what we want.

The following simple proposition states that the order in which we condition by subjective likelihood vectors makes no difference. We apply an abbreviated inner product notation: $P_i q^j = \sum_{k=1}^{n_j} P_i(j,k) q_k^j$.

Proposition 1: Let $j, \bar{j} \in \{1,2,...,m\}$, and let q^j and $q^{\bar{j}}$ be corresponding subjective likelihood vectors that do not contradict the prior p. Then we have:

$$\left(\left(p \mid q^j\right) \mid q^{\bar{j}}\right)_i = \left(\left(p \mid q^{\bar{j}}\right) \mid q^j\right)_i = \frac{p_i P_i q^j P_i q^{\bar{j}}}{\sum_{l=1}^{m} \left(p_l P_l q^j P_l q^{\bar{j}}\right)}.$$

The result is a trivial consequence of Bayesian inference theory (Press, 2003), but for readers unfamiliar with this, we give a direct proof.

Proof: We apply the formula (4):

$$\left(\left(p \mid q^j\right) \mid q^{\bar{j}}\right)_i =$$

$$\frac{\left(p \mid q^j\right)_i P_i q^{\bar{j}}}{\sum_{l=1}^{m} \left(p \mid q^j\right)_l P_l q^{\bar{j}}} = \frac{\frac{p_i P_i q^j}{\sum_{l=1}^{m} p_l P_l q^j} P_i q^{\bar{j}}}{\sum_{l=1}^{m} \frac{p_l P_l q^j}{\sum_{ll=1}^{m} p_{ll} P_{ll} q^j} P_l q^{\bar{j}}} = \frac{\sum_{ll=1}^{m} \left(p_{ll} P_{ll} q^j\right) p_i P_i q^j P_i q^{\bar{j}}}{\sum_{l=1}^{m} \left(p_l P_l q^j\right) \sum_{l=1}^{m} p_l P_l q^j P_l q^{\bar{j}}} = \frac{p_i P_i q^j P_i q^{\bar{j}}}{\sum_{l=1}^{m} p_l P_l q^j P_l q^{\bar{j}}}$$

QED.

Mushroom application

We now give an application in the domain of mushroom classification. The setting is this: The subject has found a mushroom in the woods, and needs help in determining to which species it belongs.

The Model

Each q_i corresponds to a species (or in some cases a union of similar species). The prior probability distribution p over the q's corresponds to how frequent the different species are in the woods. Features are observable properties of mushrooms. An example is "Color of the cap", with a given listing of colors, as its value set. Other features, such as "Has white spots on the cap" have the binary value set of "yes" and "no".

Each species has a given probability distribution for each feature, which represent its variability. As an example, the well-known Fly Agaric (amanita muscaria) very often has white spots on the cap, but not always. Therefore, P(yes)=0.95 and P(no)=0.05 is a reasonable distribution for the feature "Has white spots on the cap". The main color of the cap may also vary; a reasonable distribution is P(red)=0.7, P(orange)=0.2, and P(yellow)=0.1. Also, it normally has a collar on the stalk, but it sometimes falls off, and P(yes)=0.9 and P(no)=0.1 is our distribution of the feature "Has collar on the stalk" for the Fly Agaric species. Currently the implementation includes about 100 different species and 20 features.

In order to handle otherwise contradictive evidence, we have also defined a "default species" with uniform distribution for all features, and low prior probability. When the computation gives a high probability to this "species", it either means that the mushroom in question is of a species not included in the model, or the user has made incorrect observations.

Choice of features

So far we have focused on calculations used for updating the probability distribution over species, given input from the user. A different problem is the order in which the system asks its questions. For this problem of feature choice, we have implemented an optimization algorithm, which seeks to minimize the expected posterior total variance.

We define the total variance of a probability distribution p by $V(p) = \sum_{i=1}^{n} p_i (1 - p_i)$.

The minimum value of V is zero, which is realized if and only if p places all probability on one species.

Assume that the system chooses feature j. The current probability distribution p generates a distribution f over the n_j different feature values: $f_k = \frac{\sum_{i=1}^{n} p_i P_i(j,k)}{\sum_{k'=1}^{n_j} \sum_{i=1}^{n} p_i P_i(j,k')}$.

For the purpose of feature choice, we assume that the user is able to observe the feature j without uncertainty, so that his subjective likelihood vector q^j will be a unit

vector with weight 1 on component $\hat{k}$, denoted by $e_{\hat{k}}$. Under this assumption, the distribution of $\hat{k}$ is given by f above. Now we can calculate the expected total variance after conditioning by the user's response to feature j:

$E\left[V\left(p \mid q^j\right)\right] = \sum_{k=1}^{n_j} f_k V\left(p \mid e_k\right)$. The chosen feature is the one minimizing the posterior expected total variance: $j^* = \underset{j}{\operatorname{argmin}} E\left[V\left(p \mid q^j\right)\right]$ (with some arbitrary rule for breaking ties).

The following proposition states that the total expected variance cannot increase by posing question j.

Proposition 2: $E\left[V\left(p \mid q^j\right)\right] \le V\left(p\right)$

Proof: Due to the linearity of the expectancy, it suffices to show $E\left[\left(p_i \mid \hat{k}\right)\left(1-p_i \mid \hat{k}\right)\right] \le p_i\left(1-p_i\right)$. The result follows from Jensen's inequality (Ferguson, 1967), because $E\left[p_i \mid \hat{k}\right] = p_i$ and $p\left(1-p\right)$ is concave in p.

QED.

This property is rather important from a practical point of view. We have experimented with other objective functions than V, which appear intuitively reasonable, such as the probability of the most probable species, negated. This often works fine, but in some cases it asks completely irrelevant questions, that offer no information, because the relevant questions give an expected increase in the objective function. Hence the program avoids the critical questions, for fear of what it might discover. Proposition 2 guarantees that this will not happen with the objective function V.

User interface issues

It turns out to be impractical for a user to assign numbers to his uncertain observations. We have therefore implemented a user interface where he checks the different values he considers possible, in descending order of likelihood. We assign a weight of 1.0 to his first choice, 1/3 to his next, 1/5 to his third, and so on. This, of course, is not the only reasonable choice, but it appears to capture human uncertainty reasonably well. The user may choose not to check any values, which gives the uniform distribution (or equivalently: passes the question).

Experience

Over all, the system works very well, as even complete novices in the area of mushroom classification have classified a broad range of mushrooms successfully with the support of our system. A big improvement in performance came when we included sample pictures of feature values, rather than mere text descriptions.

Practical problems

The biggest practical problem we encountered was convincing the subjects to report their uncertainty by checking more than one feature value, particularly for yes/no questions. People find it easier to describe a color as "most likely beige, but possibly brown, grey or white" than to answer both "yes" and "no" to a question.

Problems of uncertainty interpretation

We have not found it necessary to explain the exact interpretation of subjective likelihood vectors to our test subjects. However, if the subject has prior experience with mushroom classification, he may start by making up his mind about which species the present mushroom belongs to, and then bias his responses toward what he knows to be typical features values for that species. This is a problem of separating observation and judgement, which one gets with the use of subjective probabilities, and which we try to avoid with subjective likelihoods. It is therefore important to instruct the subject to observe each feature individually, and leave the overall judgement to the program. Fortunately, this is only a real problem for subjects that do not need the expert system support.

Dependency problems

Our calculation scheme relies on independence of the different features for each species. This has given some problems.

We have mentioned the binary feature “Has a collar on the stalk”. A few species of the *amanita* family have a collar with clearly visible stripes. In order to distinguish the edible Blusher (*Amanita rubescens*) from poisonous *amanita* species, we therefore included the binary feature “Has collar with stripes”. Clearly, a mushroom with striped collar has a collar, so these features are not independent, which may cause problems. Suppose the user reports that that his mushroom has a collar, with 75% certainty, and then reports that it has a striped collar, also with 75% certainty. The computation would then place too much weight on the striped collar species, as both of the collar related observations appear to count in their favour. A good solution to this problem would be to merge these two features into one: “Has collar on the stalk” with values “no”, “yes, without stripes”, and “yes, with stripes”.

A different dependence problem arose with the Cantrell (Cantharellus cibarius) species. We use three different features for the color of a mushroom: its color on the cap, underneath the cap, and on the stalk. The Cantrell is normally yellow, but it may vary from whitish to orange. However, it invariably has the same color all over, so the independence assumption fails for the three color-related features. This is best solved by splitting the species into different variants, each with the same color all over.

These dependency problems are mainly of academic interest, as they do not appear to affect the frequency of misclassifications significantly.

Conclusion

We conclude that our interpretation of uncertainty in human observation through subjective likelihood estimates is successful. It appears to capture human semantic in a reasonable way, and in particular models complete ignorance successfully. Our interpretation also has the advantage of being firmly rooted in Bayeseian statistics.

Our successful application to Mushroom classification confirms the methods practical usefulness.

In our opinion, our inference model combines the “modern expert system approach” of building formally sound probabilistic models of the world, with the “classical expert system approach” of modelling human uncertainty explicitly.

References

Dubois, D., Prade, H. (1997): Bayesian conditioning in possibility theory, *Fuzzy sets and systems*, 92 (2): 223-240.

Ferguson, T. S. (1967): *Mathematical Statistics, a decision theoretic approach*, Academic Press.

Indahl, U. G. (2000): Crisp analogs of fuzzy sets, *Fuzzy sets and systems*, 110 (2): 293-298.

Jensen, F. V. (1996): *An introduction to Bayesian Networks*, UCL Press.

Press, S. J. (2003): *Subjective and objective Bayesian statistics: principles, models, and applications*, John Wiley and Sons. Hoboken, New Jersey.

Shafer, G. (1976): *A Mathematical theory of evidence*, Princeton University Press, Princeton, NJ.

Shortliffe, E. H. (1976): *Computer-based medical consultation: MYCIN*. Amsterdam: Elsevier Science.

Walley, P. (1996): Measures of uncertainty in expert systems, *Artificial Intelligence*, 83: 1-58.