1 Subjective probability modeling and its relation to the likelihood

Handling uncertainty is undoubtedly a major part of all human activities, both scientific and non-scientific ones. We have to make decisions and inference in situations where direct knowledge is not available to us. Particularly important for science, is the *logical process of decision making in situations of uncertainty*, where we face the problem of choosing an action from a set of different alternatives, each involving uncertain consequences. Our concern is to behave *rationally*, thus avoiding an "illogical" choice in such a scenario, which would typically involve worse expected consequences than other choices. It will be assumed that in presence of complete information we can always choose the *best* alternative, hence we are not dealing primarily here with the sophisticated mathematical or computational machinery that is often needed for the delivery of the solution. An example of this type of decision making under certainty is the following traveling salesman problem (TSP): Given a finite number of "cities" along with the cost of travel between each pair of them, find the cheapest way of visiting all the cities and returning to your starting point.

Subjective probability, concerns the judgements of a given person, conveniently called You, about uncertain events or propositions. We start our journey to the field of subjective probability by considering a simple problem from the frequentist point of view (which You are likely to be familiar with).

Example 1 ***Thumbtack tossing****. Consider an old-fashioned thumbtack, which is of metal with a round curved head, rather than with a colored plastic one. The thumbtack will be tossed onto a soft surface (in order not damage it), while we keep track of whether it comes to stop with the point up or point down. In the absence of any information to distinguish the tosses or to suggest that tosses occurring close together in time are any more or less likely to be similar to or different from each other than those that are far apart in time, it seems reasonable to treat the different tosses symmetrically. We might also believe that although we might only toss the thumbtack a few times, if were to toss it many more times, the same judgement of symmetry would continue to apply to the future tosses. Under such conditions, it is traditional to model the outcomes of the individual tosses as independent and identically distributed (IID) Bernoulli random variables with* $X_i = 1$ *meaning that toss* i *is point up and* $X_i = 0$ *meaning that toss* i *is point down. In the frequentist framework, one invents a* ***parameter****, say* θ*, which is assumed to be a fixed value in* $[0, 1]$ *not yet known to us (see the remark below). Then one says that the* X_i *are IID with* $P(X_i = 1) = \theta$*. The so called* ***likelihood*** *function of a sequence of* n *tosses will under this assumption take the form*

$$P(X_1 = x_1, ..., X_n = x_n) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} \tag{1}$$

which is the joint distribution of the observed values x_i *conditional on* θ*. The value of* θ *maximizing this function is the relative frequency of observing tosses*

point up, that is $\sum_{i=1}^{n} x_i/n$. *Given an observed sequence, our best guess of the probability of observing point up in the next toss, equals the relative frequency as well. You immediately see what happens under scarce information, for instance, when the only two recorded tosses we have available are point down.*

Given the earlier description of our simple thumbtack tossing problem, the assumptions made in the above frequentist approach (IID and fixed unknown θ*) may appear unnecessary stringent. In fact, this is remarkably true. To derive a subjective probabilistic description of the behavior of the tosses, we need a minimal assumption of symmetry, called* ***exchangeability****. Recall that we considered the information to be obtained from any one toss in exactly the same way we would consider the information from any other toss. Similarly, we would treat the information to be obtained from any two tosses in exactly the same way we would consider the information from any other two tosses, regardless of where they appear in our sequence of tosses. The same argument continues to apply to any subsequence of tosses. These remarks of symmetry informally define the concept of exchangeability, which lies in the heart of the subjective probability description. The concept and its generalization will be investigated formally later on. As You might already have guessed, subjective probability description of the current situation, will require Your probabilistic description about the uncertainty related to the tosses (this will considered after the introduction of some formal concepts).*

Remark 2 ***Meaning of the parameter in the thumbtack tossing problem****. A great deal of controversy in statistics arises out of the question of the meaning of such parameters as in the above example. De Finetti (1974) argues persuasively that one need not assume the existence of such things. Sometimes they are just assumed to be undefined properties of the experimental setup which magically make the outcomes behave according to our probability models. Sometimes they are defined in terms of the sequence of observations themselves (such as limits of relative frequencies). The last one is particularly troublesome because the sequence of observations does not yet exist and hence the limit of relative frequency cannot be a fixed value yet.*

From the above example we see the close connection between frequency probability and so called classical inference, because the latter requires the data to be repeatable. An unbiased estimator, for instance, is defined to have expected value equal to the parameter being estimated. Such statement is conditional on the parameter taking a fixed but unknown value, while the data are imagined as repeatable. Typically, experimental data is thought to be repeatable, thus having frequency probability distribution, while parameters governing the data behavior in such framework are considered unique and unrepeatable.

2 Predictive modeling

In the framework presented below probabilities are always *personal degrees of belief*, in that they are a numerical representation of an analyst's or decision

maker's personal uncertainty relation between events. Moreover, probabilities are always conditional on the information available. It makes thus no sense to qualify the word probability with adjectives such as "objective", "correct" or "unconditional". The term *random quantity* is here used to signify a numerical entity whose value is uncertain. The term *probability measure* (P) will be used in a rather loose manner (to avoid technicalities) to describe the way in which probability is "distributed" over the possible values of a random quantity. For a real-valued random quantity X, this may *e.g.* be given in terms of the distribution function $F(x) = P(X \leq x)$. When the probability distribution concentrates on a countable set of values, X is called a *discrete* random quantity, and we have the probability mass function $p(x) = P(X = x)$. For *continuous* random quantities we have the regular *density* function representation $P(X \in B) = \int_B p(x)dx$. Thus, to keep notation simple, $p(\cdot)$ is used both for mass and density functions.

As clearly stated in de Finetti (1974), to be able to use probability calculus as a normative tool for the description of the characteristics of interest for random quantities, one *has to* express individual degrees of belief (*i.e.* subjective opinions), expressed as probabilities about the uncertainty involved in the considered situation. That is, phrases such as "I don't know", "I can't" or "I don't want to" cannot be accepted as answers to the question concerning what one's beliefs are. The failure to express these probabilities will lead us outside the Bayesian paradigm (in the stringent sense). However, in the literature Bayesian paradigm is often understood more widely, including even cases where the subjective probabilities are replaced by formally derived functions (this aspect will be considered more in depth later).

Using generic notation we assume that the subjective degrees of belief correspond to the specification of the joint distribution $P(x_1, ..., x_n)$ of a set of random quantities $\mathbf{x} = x_1, ..., x_n$, represented by the joint density (or mass) function $p(x_1, ..., x_n)$. This specification automatically leads, for $1 \leq m < n$, to the marginal joint density

$$p(x_1, ..., x_m) = \int p(x_1, ..., x_n)dx_{m+1} \dots dx_n \tag{2}$$

and the joint density of $\mathbf{y} = x_{m+1}, ..., x_n$ (thought as yet unobserved), conditional on having observed the particular values of $\mathbf{z} = x_1, ..., x_m$, is

$$p(x_{m+1}, ..., x_n | x_1, ..., x_m) = \frac{p(x_1, ..., x_n)}{p(x_1, ..., x_m)} \tag{3}$$

A *predictive probability model* for random quantities can be defined according to the following.

Definition 3 ***Predictive probability model.*** *A predictive model for a sequence of random quantities $x_1, x_2, ...$ is a probability measure P, which specifies the joint belief distribution for any subset of $x_1, x_2, ...$.*

Consider now a sequence $x_1, x_2, ...$ under the assumption of a predictive model stating that for any n the joint density is given by

$$p(x_1, ..., x_n) = \prod_{i=1}^{n} p(x_i) \tag{4}$$

This model thereby states that the uncertain quantities are independent. If we now consider the conditional density for $1 \leq m < n$, it takes the form

$$p(x_{m+1}, ..., x_n | x_1, ..., x_m) = p(x_{m+1}, ..., x_n) \tag{5}$$

meaning that we cannot learn from experience within this sequence of interest. In other words, past data provide us with no additional information about the possible outcomes of future observations in the sequence.

A predictive model specifying such independence is clearly inappropriate in contexts where we believe that the successive accumulation of data will provide increasing information about future events. Thus, in most cases a useful predictive model, *i.e.* the structure of $p(x_1, ..., x_m)$, ought to encapsulate some form of dependence among the individual random quantities. In general, there are a vast number of possible subjective assumptions about the form of such dependencies, and here we are able to consider some commonly used canonical forms.

Suppose that, in thinking about $P(x_1, ..., x_n)$, the joint degree of belief distribution for a sequence of random quantities $x_1, ..., x_m$, an individual makes the judgement that the subscripts or the labels identifying the individual random quantities, are "uninformative". The uninformativeness is in the sense that the same marginal distribution would be specified for all possible singletons, pairs, triples etc., regardless of which labels were happened to be picked from the original sequence (recall the thumbtack tossing in Example 1). This leads us to the concept of exchangeability, formally defined below.

Definition 4 ***Exchangeability***. *Random quantities $x_1, ..., x_n$ are said to be (finitely) exchangeable under a probability measure P when the corresponding joint belief distribution satisfies*

$$p(x_1, ..., x_n) = p(x_{\pi(1)}, ..., x_{\pi(n)})$$

for an arbitrary permutation π of the labels $\{1, ..., n\}$. Further, an infinite sequence $x_1, x_2, ...$ is said to be infinitely exchangeable when every finite subsequence is finitely exchangeable.

For example, suppose that $x_1, ..., x_{100}$ are exchangeable. It follows from the above definition that they all have the same marginal distribution. Also, (x_1, x_2) has the same joint distribution as (x_{99}, x_1), and (x_5, x_2, x_{48}) has the same joint distribution as (x_{31}, x_{32}, x_{33}), and so on. The notion of exchangeability involves a judgement of complete symmetry among all the observables $x_1, ..., x_n$ under consideration. Clearly, in many situations this might be too restrictive an assumption, even though a partial judgement of symmetry is present, which should be evident from the following example.

Example 5 ***Tossing with different thumbtacks****. Consider a scenario which is similar to that of Example 1, except that we make n_i, $i = 1, ..., k$, tosses with thumbtacks of different material. For instance, the first thumbtack is made of metal, the second of plastic, the third of kevlar and so on. We might be involuntary to describe a sequence of observations under this scenario using the complete symmetry assumption leading to exchangeability. On the other hand, as before it should be reasonable to treat tosses made with the same thumbtack as exchangeable.*

We now formally treat the subjective modeling problem of an infinitely exchangeable sequence of $0-1$ (binary) random quantities (say, thumbtack tosses) $x_1, x_2, ...$ with $x_i = 0$ or $x_i = 1$, for all $i = 1, 2, ...$.

Theorem 6 ***Representation theorem for binary random quantities.*** *If $x_1, x_2, ...$ is an infinitely exchangeable sequence of binary random quantities with probability measure P, there exists a distribution function Q such that the joint mass function $p(x_1, ..., x_n)$ for $x_1, ..., x_n$ can be written as*

$$p(x_1, ..., x_n) = \int_0^1 \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} dQ(\theta) \tag{6}$$

where

$$Q(\theta) = \lim_{n\to\infty} P[y_n/n \leq \theta]$$

and $y_n = \sum_{i=1}^{n} x_i$, $\theta = \lim_{n\to\infty} y_n/n$.

The interpretation of this representation theorem is of profound significance from the point of view of subjectivist modeling philosophy. It is *as if*:

- The x_i are judged to be independent, Bernoulli random quantities conditional on a random quantity θ.
- θ is itself assigned a probability distribution $Q(\theta)$.
- By the strong law of large numbers $\theta = \lim_{n\to\infty} y_n/n$, so that Q may be interpreted as "beliefs about the limiting frequency of 1's".

What the above says is that, under the assumption of exchangeability, we may act *as if*, conditional on θ, the quantities $x_1, ..., x_n$ are a *random sample* from a Bernoulli distribution with parameter θ which corresponds to the joint sampling distribution (the *likelihood*)

$$p(x_1, ..., x_n|\theta) = \prod_{i=1}^{n} p(x_i|\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} \tag{7}$$

where the parameter θ is given a *prior distribution* $Q(\theta)$. Notice that under this interpretation the prior states beliefs about what we would anticipate observing as the limiting relative frequency. Further, the assumption of exchangeability

in the current framework considerably limits via Theorem 6 our alternatives in the specification of a predictive probability model. Any choice must be of the form given by (6), where we have the freedom of choosing the subjective beliefs about θ. By ranging over all possible choices of the prior $Q(\theta)$, we build all possible predictive probability models for the current framework.

We have thus established a justification for the conventional model building procedure of combining a likelihood and a prior. The likelihood is defined in terms of an assumption of conditional independence of the observations given a parameter. This, and its associated prior distribution, acquire an operational interpretation in terms of a limiting average of observables (here limiting frequency).

In many applications involving binary random quantities, we may be more interested in a summary random quantity, such as $y_n = x_1 + \cdots + x_n$, than in the individual sequences of x_i's. The representation $p(y_n)$ follows easily from (6), since

$$p(y_n) = \binom{n}{y_n} p(x_1, ..., x_n),$$

for all $x_1, ..., x_n$ such that $x_1 + \cdots + x_n = y_n$. We thus get

$$p(y_n) = \int_0^1 \binom{n}{y_n} \theta^{y_n} (1-\theta)^{n-y_n} dQ(\theta)$$

This provides a justification, when expressing beliefs about y_n, for acting *as if* we have a binomial likelihood with a prior distribution $Q(\theta)$ for the binomial parameter θ.

The Bayesian learning process in this simple situation is compactly represented by the following corollary.

Corollary 7 ***Corollary to the Representation theorem for binary random quantities****. If $x_1, x_2, ...$ is an infinitely exchangeable sequence of binary random quantities with probability measure P, the conditional probability function $p(x_{m+1}, ..., x_n | x_1, ..., x_m)$ for $x_{m+1}, ..., x_n$ given $x_1, ..., x_m$, has the form*

$$\int_0^1 \prod_{i=m+1}^{n} \theta^{x_i} (1-\theta)^{1-x_i} dQ(\theta | x_1, ..., x_m) \tag{8}$$

where

$$dQ(\theta | x_1, ..., x_m) = \frac{\prod_{i=1}^{m} \theta^{x_i} (1-\theta)^{1-x_i} dQ(\theta)}{\int_0^1 \prod_{i=1}^{m} \theta^{x_i} (1-\theta)^{1-x_i} dQ(\theta)}$$

and

$$Q(\theta) = \lim_{n \to \infty} P[y_n/n \leq \theta]$$

and $y_n = \sum_{i=1}^{n} x_i$, $\theta = \lim_{n \to \infty} y_n/n$.

We thus see that the basic form of representation of beliefs does not change. All that has happened, expressed in conventional terminology, is that the *prior*

distribution $Q(\theta)$ for θ has been revised into the *posterior* distribution $dQ(\theta|x_1, ..., x_m)$. The conditional probability function $p(x_{m+1}, ..., x_n|x_1, ..., x_m)$ is called the *posterior predictive* probability function. This provides the basis for deriving the conditional predictive distribution of any other random quantity defined in terms of the future observations.

In a more general setup the representation theorem states for an infinitely exchangeable sequence of real valued quantities $x_1, x_2, ...$ with probability measure P, that there exists a probability measure Q over the space $\mathcal{Q}$ of all distribution functions for the observable quantity, such that the joint distribution function of $x_1, ..., x_n$ can be written as

$$P(x_1, ..., x_n) = \int_{\mathcal{Q}} \prod_{i=1}^{n} F(x_i) dQ(F) \tag{9}$$

where

$$Q(F) = \lim_{n\to\infty} P(F_n) \tag{10}$$

where F_n is the empirical distribution function defined by $x_1, ..., x_n$. Thus, we may act *as if* we have independent observations $x_1, ..., x_n$ conditional on F, which is an unknown distribution function playing the role of an infinite-dimensional parameter. The belief distribution Q has in this case the interpretation of what we believe the empirical distribution function F_n would look like for a "large" number of observations. This result can be analogously extended to a finite dimensional Euclidean space for vector valued random quantities.

If, in particular, our beliefs are such that the distribution function F can be defined in terms of a finite-dimensional parameter $\boldsymbol{\theta}$, the joint density of our observations can be written as

$$p(x_1, ..., x_n) = \int_{\boldsymbol{\Theta}} \prod_{i=1}^{n} p(x_i|\boldsymbol{\theta}) dQ(\boldsymbol{\theta}) \tag{11}$$

where $p(\cdot|\boldsymbol{\theta})$ is the density function corresponding to the unknown parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. By taking a step yet further, and letting $p(\boldsymbol{\theta})$ correspond to the density representation of $Q(\boldsymbol{\theta})$, we obtain

$$p(x_1, ..., x_n) = \int \prod_{i=1}^{n} p(x_i|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{12}$$

From the above we may deduce that

$$\begin{aligned} p(x_{m+1}, ..., x_n|x_1, ..., x_m) &= \frac{\int \prod_{i=1}^{n} p(x_i|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int \prod_{i=1}^{m} p(x_i|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}} \\ &= \frac{\int \prod_{i=m+1}^{n} p(x_i|\boldsymbol{\theta}) \prod_{i=1}^{m} p(x_i|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int \prod_{i=1}^{m} p(x_i|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}} \end{aligned} \tag{13}$$

where

$$\frac{\prod_{i=1}^{m} p(x_i|\boldsymbol{\theta}) p(\boldsymbol{\theta})}{\int \prod_{i=1}^{m} p(x_i|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}} = p(\boldsymbol{\theta}|x_1, ..., x_m) \tag{14}$$

so that

$$p(x_{m+1},...,x_n|x_1,...,x_m) = \int \prod_{i=m+1}^{n} p(x_i|\boldsymbol{\theta})p(\boldsymbol{\theta}|x_1,...,x_m)d\boldsymbol{\theta} \qquad (15)$$

The relation in (14) is just *Bayes' theorem*, which expresses the posterior density for $\boldsymbol{\theta}$ in the context of parametric model for $x_1,...,x_m$ given $\boldsymbol{\theta}$. By using the more compact notations about the "future" $\mathbf{y}$ and the "current" observations $\mathbf{z}$, we see that

$$\begin{aligned} p(\mathbf{x}) &= \int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} \qquad (16) \\ p(\mathbf{y}|\mathbf{z}) &= \int p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{z})d\boldsymbol{\theta} \\ p(\boldsymbol{\theta}|\mathbf{z}) &= \frac{p(\mathbf{z}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{z})} \end{aligned}$$

In particular the role of Bayes' theorem is identified as a coherent learning step about the unobservables when we pass from $p(\mathbf{z})$ to $p(\mathbf{y}|\mathbf{z})$.

References

[1] Bernardo, J. M. and Smith, A. F. M. (1994). Bayesian theory. Chichester: Wiley.

[2] de Finetti, B. (1974). Theory of probability **1**. Chichester: Wiley.