A Naive Bayes Heuristic for Procedure Probability Lift

Peter Bouman

October 4, 2013

We want a formula to calculate the ratio between the probability of seeing a given procedure group in the seven days before an initial home health visit. This ratio, which we call a "lift," expresses the degree to which the procedure tends to cluster before home health visits, and therefore lift sufficiently above unity expresses support for the hypothesis that the home visits are justified by previous medical services.

Let's suppose that we observe the procedures P_i in the week preceding a home health visit. We can denote the event that a home health visit takes place when observed as H.

A useful measure of the plausibility of the commencement of home health care is the probability

$$P(H|\cap P_i). \tag{1}$$

(The notation $\cap P_i$ indicates the co-occurrence of all observed procedures.) If the conditional probability (1) is close to 1, then we believe that the home health care visit has good support from the preceding week's medical activity. If it is close to 0, the procedures observed do not provide a good explanation.

We can invoke Bayes' Theorem to rewrite this probability:

$$P(H|\cap P_i) = \frac{P(\cap P_i|H)P(H)}{P(\cap P_i)} \tag{2}$$

Equation (2) tells us we can restate equation (1)'s conditional probability in terms of the joint probabilities of observing all the procedures recorded. In practice, these joint probabilities will all be impractical to estimate from historical data, so we make the Naive Bayes simplifying assumption of marginal and conditional independence, an assumption that often leads to surprisingly good model performance despite the actual dependencies in the procedure probabilities.

Under Naive Bayes, the equation can be reorganized as:

$$P(H|\cap P_i) \approx \left(\prod_i \frac{P(P_i|H)}{P(P_i)}\right) P(H) \tag{3}$$

We recognize each ratio $P(P_i|H)/P(P_i)$ as the ratio of conditional to marginal probabilities of seeing each procedure in the seven days before the beginning of home health care. This ratio can be precomputed for each procedure or procedure group and used to score the group of procedures before each home health episode.

There are two considerations that come out of our scoring approach: First, we can allow procedure scores to be defined relatively, meaning we can ignore all constants like $P(H)$. Second, the relative comparison can just as well take place on the log scale as the linear scale (monotonicity), and so we can actually score the combination of observed procedures as:

$$\sum_i (\log P(P_i|H) - \log P(P_i)) \tag{4}$$

Note that we are defining $P(P_i)$ as the probability of seeing the procedure at least once in the seven days before home health care. We can approximate the log of this probability as $\log P(P_i) = \log(1 - \exp(-7P_i')) \approx \log(7P_i')$, where P_i' is the average daily rate of occurrence of the procedure in historical data.

Another minor consideration is the set of procedures that are never observed in the week before home health, and therefore have $P(P_i|H) = 0$. Since $\log(0)$ is undefined, we substitute in a value (like -6) guaranteed to put these procedures at the bottom of the sorted list.