**Course: Advanced Bioinformatics**

**Module title: Naïve Bay’s Classifier**

**Module no. : 180**

In probability theory and statistics, Bayes' theorem (alternatively Bayes' law or Bayes' rule) describes the probability of an event, based on conditions that might be related to the event. For example, suppose one is interested in whether Addison has cancer, and that she is 65. If cancer is related to age, information about Addison's age can be used to more accurately assess the probability of her having cancer using Bayes' Theorem.

When applied, the probabilities involved in Bayes' theorem may have different probability interpretations. In one of these interpretations, the theorem is used directly as part of a particular approach to statistical inference. With the Bayesian probability interpretation the theorem expresses how a subjective degree of belief should rationally change to account for evidence: this is Bayesian inference, which is fundamental to Bayesian statistics. However, Bayes' theorem has applications in a wide range of calculations involving probabilities, not just in Bayesian inference.

Bayes' theorem is stated mathematically as the following equation

P(A|B) = \frac{P(A)\, P(B | A)}{P(B)},

where A and B are events.

P(A) and P(B) are the probabilities of A and B without regard to each other.

P(A | B), a conditional probability, is the probability of observing event A given that B is true.

P(B | A), is the probability of observing event B given that A is true.

Drug testing

Tree diagram illustrating drug testing example. U, U bar, "+" and "−" are the events representing user, non-user, positive result and negative result. Percentages in parentheses are calculated.

Suppose a drug test is 99% sensitive and 99% specific. That is, the test will produce 99% true positive results for drug users and 99% true negative results for non-drug users. Suppose that 0.5% of people are users of the drug. If a randomly selected individual tests positive, what is the probability he or she is a user?


\begin{align}
P(\text{User}\mid\text{+}) &= \frac{P(\text{+}\mid\text{User}) P(\text{User})}{P(\text{+}\mid\text{User}) P(\text{User}) + P(\text{+}\mid\text{Non-user}) P(\text{Non-user})} \\[8pt]
&= \frac{0.99 \times 0.005}{0.99 \times 0.005 + 0.01 \times 0.995} \\[8pt]
&\approx 33.2\%
\end{align}

Despite the apparent accuracy of the test, if an individual tests positive, it is more likely that they do not use the drug than that they do. This again illustrates the importance of base rates, and how the formation of policy can be egregiously misguided if base rates are neglected.

This surprising result arises because the number of non-users is very large compared to the number of users; thus the number of false positives (0.995%) outweighs the number of true positives (0.495%). To use concrete numbers, if 1000 individuals are tested, there are expected to be 995 non-users and 5 users. From the 995 non-users, 0.01 × 995 ≃ 10 false positives are expected. From the 5 users, 0.99 × 5 ≃ 5 true positives are expected. Out of 15 positive results, only 5, about 33%, are genuine.

Note: The importance of specificity can be illustrated by showing that even if sensitivity is 100% and specificity is at 99% the probability of the person being a drug user is ≈33% but if the specificity is changed to 99.5% and the sensitivity is dropped down to 99% the probability of the person being a drug user rises to 49.8%.