Bayesian Algorithms

At the core of Bayes' theorem is the simple concept of starting with a hypothesis in the form of an estimated probability that the hypothesis is correct and then adjusting the probability as new evidence (i.e., data) becomes available. For example, if we flip a true coin a sufficiently large number of times the number of heads will be approximately equal to the number of tails. Therefore, if our hypothesis is that the coin is true then our initial estimated probability of the flipped coin coming down heads would be 50%. As we continue to flip the coin each heads will increase and each tails will decrease the probability of the coin being true. The initial probability, which is essentially a subjective belief, is referred to as the prior probability and the adjusted probability based on new data is referred to as the posterior probability. The posterior probability is determined with Bayes' theorem as follows, where P stands for probability:

P(cause | effect) = P(cause) x P(effect | cause) / P(effect)

Pedro Domingos in his excellent book The Master Algorithm (Basic Books 2015, page 147) provides the following medical example where influenza is the cause and fever is the effect. If out of 100 patients, 14 had influenza, 20 had a fever that was not necessarily associated with influenza, and 11 had both influenza and a fever. Therefore:

P(cause-influenza)	**=**	**14 / 100**
P(effect-fever) | cause-influenza)	**=**	**11 / 14**
P(effect-fever)	**=**	**20 / 100**
P(cause-influenza | effect-fever)	**=**	**(14/100) x (11/14) / (20/100)**
posterior probability	**=**	**0.55 or 55%**

We usually know the probability of the effects given the cause. For example, the probability of a fever if the patient has influenza. However, what we would like to know is the probability of the cause given the effect, such as the probability that the patient has influenza if the patient has a fever. If in the above example the physician had started with an intuitive estimate of the prior probability of 70% that a patient with a fever has influenza, then this probability would now be reduced to 55%. Additional data would either increase or decrease the 55% that has now become the prior probability.

In reality, the application of Bayes' theorem to this machine-learning example would require multiple effects such as sore throat, fatigue, prevalence of influenza in proximity of the patient, and so on, to be taken into account. The computational burden increases exponentially. If there are n effects and each carries the Boolean value of yes or no, then there are 2n combinations that have to be calculated for each data set. In the case of only 10,000 data sets and only ten effects there are already over 10 million calculations ($10000 \times 2^{10} = 10000 \times 1024 = 10{,}024{,}000$).

However, we are faced with two further problems. First, there are dependencies that require combinations of symptoms to be considered. A patient who has a fever and also a sore throat is more likely to have influenza than a patient with only one of those symptoms. Second, we need an enormous amount of data to have some confidence that the available data covers the combinatorial range of different cases and that there are a sufficient number of instances of each case. To deal with the combinatorial explosion problem we are forced to make concessions such as, we assume that the effects are independent of each other given the cause. This is referred to as the Naïve Bayes Classifier. Naïve Bayes is widely applied to email spam filters, search engines, text classifiers, and in many other domains.