

MDL exercises, third handout
(due March 6th)

1. The ELISA test for HIV (the AIDS virus) is used in America to screen blood donations. If a person actually carries HIV, experts estimate that the test gives a positive result 97.7% of the time. If a person does not carry HIV, ELISA gives a negative result 92.6% of the time. Estimates are that 0.5% of the American public carry HIV, 77% of which are male. Evelyn Average has just tested positive on ELISA and is scared out of her wits. What is the probability that she is infected? Hint: do this by relating the quantity of interest, $P(D | E)$, to the available knowledge, $P(E | D)$ and $P(E | D^c)$, where D is the event that she is infected with the disease, E is the event that she tests positive on ELISA, and D^c is the complement of event D .
2. In the Bayesian framework, the marginal probability of the $n + 1$ st outcome after observing the first n outcomes $P_M(x_{n+1} | x^n)$ can be computed as $\int_{\theta} P(x_{n+1} | \theta) w(\theta | x^n) d\theta$. Show that this is equal to $P_M(x^{n+1}) / P_M(x^n)$.
3. Consider the Bernoulli model, parameterized by the probability of observing one.
 - a) Let $P_M(n_0, n_1)$ abbreviate the Bayesian marginal probability $P_M(x^n)$ under a uniform prior of observing a sequence x^n with n_0 zeroes and n_1 ones. Use integration by parts to show that $(n_0 + 1)P_M(n_0, n_1 + 1) = (n_1 + 1)P_M(n_0 + 1, n_1)$.
 - b) Use this recurrence relation to prove Laplace's Law of Succession: the Bayesian probability under a uniform prior that the next outcome will be a one is $\frac{n_1 + 1}{n_0 + n_1 + 2}$, where n_0 and n_1 are the numbers of zeroes and ones that have been observed until now. Hint: also use Exercise 2.
 - c) According to the Law of Succession, the probability for the first outcome is $\frac{1}{2}$. Outcomes in a Bernoulli sequence are independent, so the probability of a sequence of length n is 2^{-n} . What is wrong with this argument? Use the Law of Succession and the chain rule of conditional probability (Section 2.2.2. in the book) to compute the real Bayesian probability of a sequence with n_0 zeroes and n_1 ones.
 - d) Consider two codes for coding sequences of 0s and 1s. One is the Bayesian code with lengths $-\log P_M(x^n)$, where P_M is the Bayesian probability based on a uniform prior over the Bernoulli model as above. The other is the two-stage code where you first code the number of 1s n_1 in x^n using a uniform code, and then you code the actual sequence with that number of 1's, using again a uniform code over all sequences of length n with n_1 1s.
Which code is better and why?