MDL exercises, first handout

(due February 18th)

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, $P(D \mid E)$, to the available knowledge, $P(E \mid D)$ and $P(E \mid D^c)$, 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 $P_M(x_{n+1} \mid x^n)$ can be computed as $\int_\theta P(x_{n+1} \mid \theta) w(\theta \mid x^n) d\theta$. 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) Suppose we observe an individual sequence X^n of 0s and 1s of length n, with n_1 ones. Let θ represent that the data are i.i.d. Bernoulli with probability of 1 given by θ. Suppose first (a) that $n_1/n = 1/3$. For what value of θ is $P_\theta(x^n)$ maximized? Now suppose (b) that $\theta = 1/3$. For what value of n_1 is $P_\theta(x^n)$ maximized? Finally, suppose (c) again that $\theta = 1/3$. We sample n outcomes from P_θ, and now we regard N_1 as a random variable. For what values of m is $P_\theta(N_1 = m)$

approximately maximized? (it is very hard to calculate this precisely, so it is sufficient to give two sets B_1 and B_2 (whose definition may depend on n) such that for large n, for some $m \in B_1$, the probability is much larger than for all $m \in B_2$).