Why Bayesian statistics?

1. Bayes' theorem
2. The table game
3. Why Bayesian statistics?
4. Bayesian sequence analysis

1. Bayes' theorem

- Conditional and joint probabilities

  $P(A,B) = P(A|B)P(B) = P(B|A)P(A)$

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

- Bayes' throrem (1763)

  A=D (data)

  B=M (model)

  $$P(M|D) = \frac{P(D|M)P(M)}{P(D)}$$

- P(M|D): posteriori （后验概率）
- P(D|M): data Likelihood（似然概率）
- P(M): Priori（先验概率）
- P(D): Evidence probability（事实概率）

The parameters for a probabilistic model are typically estimated from large sets of trusted examples, often called a training set.

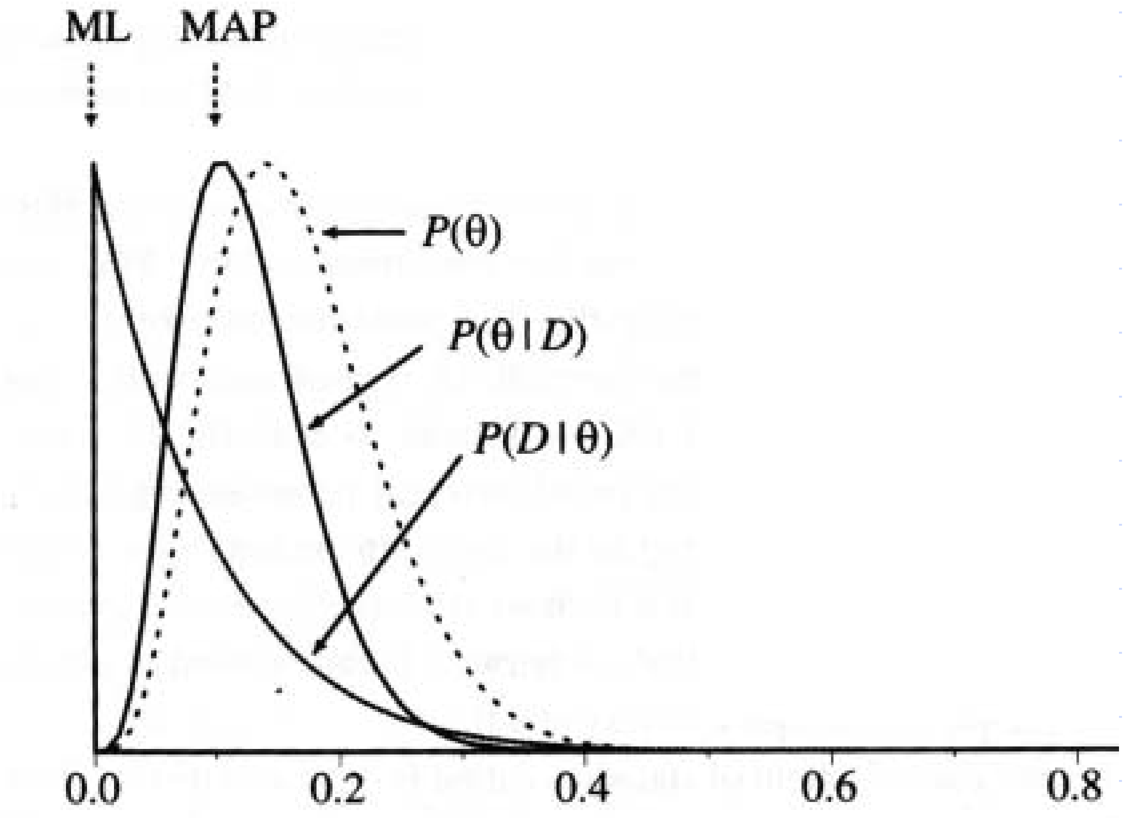
- ML: Maximum likelihood estimation
- MAP: Maximum a posteriori

$$P(\theta \mid D) = \frac{P(\theta)P(D \mid \theta)}{\int_{\theta'} P(D \mid \theta')P(\theta')}$$

Note that since our parameters are usually continuous rather than discrete quantities, the denominator is now an integral rather than a sum:

$$P(D) = \int_{\theta'} P(D \mid \theta')P(\theta')$$

A dice example



2. The table game

- Alice and Bob: a game in which the first one to six points wins. But the way each point is decided is a little strange…
- Now, Alice is already winning 5 points to 3 (A leads B 5:3), and what is the expected probability that Alice will win?

- The table game is an example of a scientific inference problem. It was controversial for centuries after first being proposed in the 1300's. Published solutions included 2:1 and 3:1 odds. In the mid-1600's, Blaise Pascal's 7:1 solution is considered to be one of the origins of probability theory.

- If the probability *p* were known, this would be easy. For example, if the mark was exactly in the middle of the table, probability (*p*) that Bob wins is 1/8, and fair odds would be 7:1 (Pascal);
- One approach would be to make a ML of the unknown parameter p. Bob's probability of winning is $(3/8)^3=27/512$, and Alice's probability of winning to be 485/512; fair odds would be about 18:1;
- How about Bayesian solution?

The Bayesian solution

$$E(Bob-wins)=\int_0^1 (1-p)^3 P(p\,|\,A=5,B=3)\,dp$$

$$P(p\,|\,A=5,B=3)=\frac{P(A=5,B=3\,|\,p)P(p)}{\int_0^1 P(A=5,B=3\,|\,p)P(p)\,dp}$$

$$P(A=5,B=3\,|\,p)=\frac{8!}{5!3!}P^5(1-p)^3$$

$$E(Bob-wins)=\frac{\int_0^1 p^5(1-p)^6\,dp}{\int_0^1 p^5(1-p)^3\,dp}$$

$$\int_0^1 p^{m-1}(1-p)^{n-1}\,dp=\frac{\Gamma(n)\Gamma(m)}{\Gamma(m+n)}$$

$$\Gamma(n+1)=n!$$

$$E(Bob - wins) = \frac{5!6!}{12!} / \frac{5!3!}{9!} = \frac{1}{11}$$

And Alice's expected probability is 10/11.

Thus, the Bayesian fair odds to be 10:1.

The Table game is adapted by S. R. Eddy from the key example in the landmark, posthumous 1763 paper by the Reverend Thomas Bayes.

- Distinctive features: (1)If we need to invoke uncertain parameters in the problem, we don’t attempt to make point estimates of these parameters; (2)The use of inverse probability calculation; (3)Using probability to represent a degree of belief.

Exercise

- Consider an occasionally dishonest casino that uses two kinks of dice. Of the dice 99% are fair but 1% are loaded so that a six comes up 50% of the time. We pick a dice at random and roll it three times, getting three consecutive sixes. We are suspicious that this is a loaded die. How can we evaluate whether that is the case?

3. Why Bayesian statistics?

- There seem to be a lot of computational biology papers with “Bayesian” in their titles these days.
- At least two reasons: (1) its explicit use of probabilistic models to formulate scientific problems (i.e. a quantitative storytelling); (2) its coherent way of incorporating all sources of information and of treating nuisance parameters and missing data (Liu, 2002).
- There is no shortage of problems in biology where we want to infer something from observed data, but the inference depends on <u>uncertain parameters</u> or <u>missing data</u> in a <u>probability model</u>.

4. Bayesian sequence analysis

- Significance of alignment scores

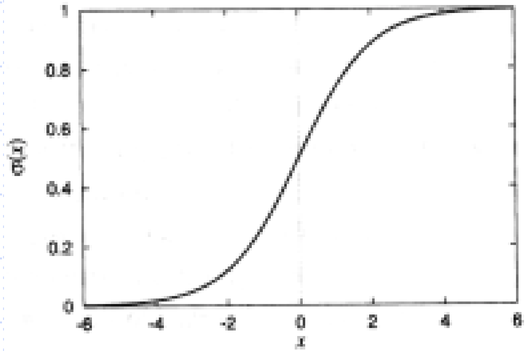
$$P(M \mid x, y) = \frac{P(x, y \mid M)P(M)}{P(x, y)}$$

$$P(M \mid x, y) = \frac{P(x, y \mid M)P(M)}{P(x, y \mid M)P(M) + P(x, y \mid R)P(R)}$$

Let $S' = S + \log(\frac{P(M)}{P(R)})$ where $S = \log(\frac{P(x, y \mid M)}{P(x, y \mid R)})$

Then $P(M \mid x, y) = \sigma(S')$

$$\sigma(x) = \frac{e^x}{1 + e^x}$$



(R.S.A.K.'s book, pp36-37)

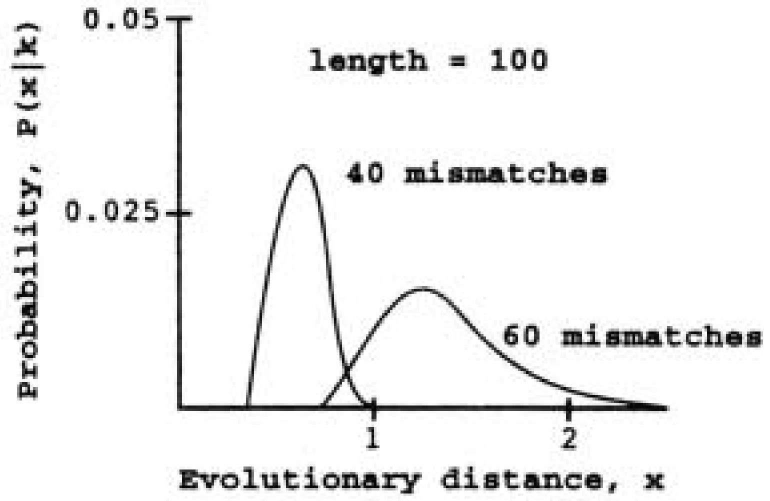
Bayesian evolutionary distance

$$P(x \mid k) = P(k \mid x)P(x) / P(k) = P(k \mid x)P(x) / \sum_x P(k \mid x)$$

where P(x|k) is the probability of distance x given the sequence with k mismatches (and n-k matches), P(k|x) is the odds score for the sequence with k mismatches using the log odds scores in the DNA PAM100x matrix, and P(x) is the prior probability of distance x). The denominator is the sum of the odds scores over the range of x, which is 0.01- 4, representing PAM1 to PAM400.

(Mount's book, pp122-123)

An example



(Mount's book, pp122-123)