Probabilities and Probabilistic Models

Probabilistic models

- A *model* means a system that simulates an object under consideration.
- A *probabilistic model* is a model that produces different outcomes with different probabilities – it can simulate a *whole class* of objects, assigning each an associated probability.
- In bioinformatics, the objects usually are DNA or protein sequences and a model might describe a family of related sequences.

Examples

1. The roll of a six-sided dice – six parameters p_1, p_2, …, p_6, where p_i is the probability of rolling the number i. For probabilities, $p_i > 0$ and
$$\sum_i p_i = 1$$
2. Three rolls of a dice: the model might be that the rolls are independent, so that the probability of a sequence such as $[2, 4, 6]$ would be $p_2\, p_4\, p_6$.
3. An extremely simple model of any DNA or protein sequence is a string over a 4 (nucleotide) or 20 (amino acid) letter alphabet. Let q_a denote the probability, that residue a occurs at a given position, at random, independent of all other residues in the sequence. Then, for a given length n, the probability of the sequence $x_1, x_2, \ldots, x_n$ is
$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} q_{x_i}$$

Conditional, joint, and marginal probabilities

- two dice D_1 and D_2. For $j = 1,2$, assume that
  - the probability of using die D_j is $P(D_j)$, and for $i = 1,2,\ldots, 6$, and
  - the probability of rolling an i with dice D_j is $P_{D_j}(i)$.

- In this simple two dice model, the *conditional probability* of rolling an i with dice D_j is: $P(i \mid D_j) = P_{D_j}(i)$.

- The *joint probability* of picking die D_j and rolling an i is:

$$P(i, D_j) = P(D_j)P(i \mid D_j).$$

- The probability of rolling i – *marginal probability*

$$P(i) = \sum_{j=1}^{2} P(i, D_j) = \sum_{j=1}^{2} P(D_j)P(i \mid D_j)$$

Maximum likelihood estimation

(maximálne vierohodný odhad)

- Probabilistic models have parameters that are usually estimated from large sets of trusted examples, called a training set.
- For example, the probability q_a for seeing amino acid a in a protein sequence can be estimated as the observed frequency f_a of a in a database of known protein sequences, such as SWISS-PROT.
- This way of estimating models is called Maximum likelihood estimation, because it can be shown that using the observed frequencies maximizes the total probability of the training set, given the model.
- In general, given a model with parameters and a set of data D, the maximum likelihood estimate (MLE) for θ is the value which maximizes $P(D \mid \theta)$.

Model comparison problem

- An occasionally dishonest casino uses two kinds of dice, of which 99% are fair, but 1% are loaded, so that a 6 appears 50% of the time.
- We pick up a dice and roll [6, 6, 6]. This looks like a loaded dice, is it? This is an example of a model comparison problem.
- I.e., our hypothesis D_{loaded} is that the dice is loaded. The other alternative is D_{fair}. Which model fits the observed data better? We want to calculate:

$$P(D_{loaded} \mid [6, 6, 6])$$

Prior and posterior probability

- $P(D_{loaded} \mid [6, 6, 6])$ is the posterior probability that the dice is loaded, given the observed data.

apriórna pravdepodobnosť

- Note that the prior probability of this hypothesis is 1/100 – prior because it is our best guess about the dice before having seen any information about the it.
- The likelihood of the hypothesis D_{loaded}:

$$P([6,6,6] \mid D_{loaded}) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8}.$$

- Posterior probability – using Bayes' theorem

Aposteriórna pravdepodobnosť

$$P(X \mid Y) = \frac{P(Y \mid X)\, P(X)}{P(Y)}$$

Comparing models using Bayes' theorem

- We set $X = D_{loaded}$ and $Y = [6,6,6]$, thus obtaining

$$P(D_{loaded} \mid [6,6,6]) = \frac{P([6,6,6] \mid D_{loaded})\, P(D_{loaded})}{P([6,6,6])}$$

- The probability $P(D_{loaded})$ of picking a loaded die is 0.01.
- The probability $P([6, 6, 6] \mid D_{loaded})$ of rolling three sixes using a loaded die is $0.5^3 = 0.125$.
- The total probability $P([6, 6, 6])$ of three sixes is

$$P([6, 6, 6] \mid D_{loaded})\, P(D_{loaded}) + P([6, 6, 6] \mid D_{fair})\, P(D_{fair}).$$

- Now,

$$P(D_{loaded} \mid [6,6,6]) = \frac{(0.5^3)(0.01)}{(0.5^3)(0.01) + \left(\frac{1}{6}\right)^3 (0.99)} \approx 0.214$$

- Thus, the die is probably **fair**.

Biological example

- Lets assume that extracellular (*ext*) proteins have a slightly different composition than intercellular (*int*) ones. We want to use this to judge whether a new protein sequence $x_1,\ldots, x_n$ is *ext* or *int*.
- To obtain training data, classify all proteins in SWISS-PROT into *ext*, *int* and unclassifiable ones.
- Determine the frequencies f_a^{ext} and f_a^{int} of each amino acid *a* in *ext* and *int* proteins, respectively.
- To be able to apply Bayes' theorem, we need to determine the priors p^{int} and p^{ext}, i.e. the probability that a new (unexamined) sequence is extracellular or intercellular, respectively.

Biological example - cont.

- We have: $P(x \mid ext) = \prod_{i=1}^{n} q_{x_i}^{ext}$ and $P(x \mid int) = \prod_{i=1}^{n} q_{x_i}^{int}$
- If we assume that any sequence is either extracellular or intercellular, then we have

$$P(x) = p^{ext} P(x \mid ext) + p^{int} P(x \mid int).$$

- By Bayes' theorem, we obtain

$$P(ext \mid x) = \frac{P(ext)P(x \mid ext)}{P(x)} = \frac{p^{ext} \prod_i q_{x_i}^{ext}}{p^{ext} \prod_i q_{x_i}^{ext} + p^{int} \prod_i q_{x_i}^{int}}$$

the posterior probability that a sequence is extracellular.

- (In reality, many transmembrane proteins have both intra- and extracellular components and more complex models such as HMMs are appropriate.)

Probability vs. likelihood

pravdepodobnosť vs. vierohodnosť

- If we consider $P(X \mid Y)$ as a function of X, then this is called a probability.
- If we consider $P(X \mid Y)$ as a function of Y, then this is called a likelihood.