Bioinformatics 2 - Lecture 1A

Guido Sanguinetti

School of Informatics
University of Edinburgh

January 16, 2012

Definitions

- Random variables: results of non exactly reproducible experiments
- Either intrinsically random (e.g. quantum mechanics) or the system is incompletely known, cannot be controlled precisely
- The probability p_i of an experiment taking a certain value i is the frequency with which that value is taken in the limit of innite experimental trials
- Alternatively, we can take probability to be our belief that a certain value will be taken

More definitions

- Let x be a random variable, the set of possible values of x is the *sample space* Ω
- Let x and y be two random variables, $p(x = i, y = j)$ is the *joint probability* of x taking value i and y taking value j (with i and j in the respective sample spaces. Often just written $p(x, y)$ to indicate the function (as opposed to its evaluation over the outcomes i and j)
- $p(x|y)$ is the conditional probability, i.e. the probability of x if you know y has a certain value

Rules

- *Normalisation*: the sum of the probabilities of all possible experimental outcomes must be 1, $\sum_{x\in\Omega} p(x) = 1$
- *Sum rule*: the marginal probability $p(x)$ is given by summing the joint $p(x,y)$ over all possible values of y,

$$p(x) = \sum_{y\in\Omega} p(x,y)$$

- *Product rule*: the joint is the product of the conditional and the marginal, $p(x,y) = p(x|y)p(y)$
- *Bayes rule*: the posterior is the ratio of the joint and the marginal

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

- *Problem!* Computing the marginal is often computationally intensive

Distributions and expectations

- A probability distribution is a rule associating a number $0 \leq p(x) \leq 1$ to each state $x \in \Omega$, such that $\sum_{x\in\Omega} p(x) = 1$
- For finite state space can be given by a table, in general is given by a functional form
- Probability distributions (over numerical objects) are useful to compute expectations of functions

$$\langle f \rangle = \sum_{x\in\Omega} f(x)p(x)$$

- Important expectations are the *mean* $\langle x \rangle$ and *variance* $var(x) = \langle (x - \langle x \rangle)^2 \rangle$. For more variables, also the *covariance* $cov(x, y) = \langle (x - \langle x \rangle)(y - \langle y \rangle) \rangle$ or its scaled relative the *correlation* $corr(x, y) = cov(x, y)/\sqrt{var(x)var(y)}$

Computing expectations

- If you know analytically the probability distribution and can compute the sums (integrals), no problem
- If you know the distribution but cannot compute the sums (integrals), enter the magical realm of approximate inference (fun but out of scope)
- If you know nothing bur have N_S samples, then use a sample approximation
- Approximate the probability of an outcome with the *frequency* in the sample

$$\langle f(x) \rangle \simeq \sum_x \frac{n_x}{N_S} f(x) = \frac{1}{N_S} \sum_{i=1}^{N_S} f(x_i)$$

(prove the last equality)

Independence

- Two random variables x and y are *independent* if their joint probability factorises in terms of marginals

$$p(x, y) = p(x)p(y)$$

- Using the product rule, this is equivalent to the conditional being equal to the marginal

$$p(x, y) = p(x)p(y) \Leftrightarrow p(x|y) = p(x)$$

- Exercise: if two variables are independent, then their correlation is zero. **NOT TRUE** viceversa (no correlation does not imply independence)

Continuous states

- If the state space Ω is continuous some of the previous definitions must be modified
- The general case is mathematically difficult; we restrict ourselves to $\Omega = \mathbb{R}^n$ and to distributions which admit a *density*, a function

$$p : \Omega \rightarrow \mathbb{R} \quad \text{s.t.} \quad p(x) \geq 0 \forall x \quad \text{and} \quad \int_{\Omega} p(x) dx = 1$$

- It can be shown that the rules of probability distributions hold also for probability densities
- Notice that $p(x)$ is NOT the probability of the random variable being in state x (that is always zero for bounded densities); probabilities are only defined as integrals over subsets of Ω

Basic distributions

- Discrete distribution: a random variable can take N distinct values with probability $p_i = 1, \ldots, N$. Formally

$$p(x = i) = \prod_j p_j^{\delta_{ij}}$$

δ_{ij} is the Kronecker delta and the p_is form a vector of parameters.

- Dirichlet distribution: a distribution over vectors of continuous variables $(p_1, \ldots, p_N)$ s.t. $\sum_i p_i = 1$. Its density is given by

$$f(p_1, \ldots, p_N | \alpha_1, \ldots, \alpha_N) = \frac{1}{Z} \prod_i p_i^{\alpha_i - 1}$$

Z is a normalisation constant, αs are parameters

Basic distributions

- Multivariate normal: distribution over vectors $\mathbf{x}$, density

$$p\left(\mathbf{x}|\boldsymbol{\mu},\Sigma\right)=\frac{1}{\sqrt{2\pi|\Sigma|}}\exp\left[-\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}\right)^{T}\Sigma^{-1}\left(\mathbf{x}-\boldsymbol{\mu}\right)\right]$$

  How many parameters does a multivariate normal have?

- Gamma distribution: distribution over positive real numbers, density

$$p\left(x|k,\theta\right)=\frac{x^{k}-1\exp(-x/\theta)}{\theta^{k}\Gamma(\theta)}$$

  with shape parameter k and scale parameter θ

Parameters?

- Many distributions are written as conditional probabilities *given* the parameters
- Often the values of the parameters are not known
- Given observations, we can estimate them; e.g., we pick θ by maximum likelihood

$$\hat{\theta} = \mathrm{argmax} \left[\prod p(x - i|\theta) \right]$$

- Or one could place a prior distribution over the parameters
- Posteriors are computed via Bayes theorem

Exercise: fitting a discrete distribution

- We have independent observations $x_1, \ldots, x_N$ each taking one of D possible values, giving a likelihood

$$\mathcal{L} = \prod_{i=1}^{N} p\left(x_i | \mathbf{p}\right)$$

- Compute the Maximum Likelihood estimate of $\mathbf{p}$. What is the intuitive meaning of the result? What happens if one of the D values is not represented in your sample?
- Alternatively, place a Dirichlet prior with parameters α over $\mathbf{p}$ and compute the posterior distribution. What is the meaning of the prior parameters?

Conjugate priors

- The Bayesian way has advantages in that it quantifies uncertainty and regularizes naturally
- BUT computing the normalisation in Bayes theorem is very hard
- The case when it is possible is when the prior and the posterior are of the same form (*conjugate*)
- Example: discrete and Dirichlet (exercise before)
- Exercise: conjugate priors for the univariate normal