Bioinformatics 2 - Lecture 2

Guido Sanguinetti

School of Informatics
University of Edinburgh

January 19, 2011

Basics of probability theory

- 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 infinite experimental trials
- Alternatively, we can take probability to be our *belief* that a certain value will be taken

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

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

- *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

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

- 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* $cov(x, y) = \langle (x - \langle x \rangle)(y - \langle y \rangle) \rangle$ and 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

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

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

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

- Using Bayes' theorem, one obtains also

$$p(x|y) = p(x)$$

Continuous states

- If the state space Ω is continuous 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

$$p\colon \Omega \to \mathbb{R} \quad \text{s.t.} \quad p(x) > 0 \forall x \quad \text{and} \quad \int_\Omega p(x)dx = 1$$

- 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 $p_i \quad i = 1 \ldots, N$. Formally

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

  Notice the p_i values can be thought as parameters of the distribution

- 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

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

  Z is a normalisation constant which is expressed in terms of the *Beta* function.

Basic distributions

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

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

  Parameters are the mean vector $\boldsymbol{\mu}$ and covariance matrix Σ (symmetric and positive definite)

- Gamma distribution: distribution over positive real numbers, density

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

  with shape parameter k and scale parameter θ.

Parameters?

- 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 $\hat{\theta}$ by maximum likelihood

$$\hat{\theta} = \text{argmax} \left[\prod_i p(x_i|\theta) \right]$$

- 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

$$\mathcal{L} = \prod_{i=1}^{N} p(x_i|\mathbf{p}) = \prod_{j=1}^{D} p_j^{n(x=j)}$$

- Maximum likelihood (bear in mind the constraint $\sum_i p_i = 1$) leads to

$$p_j = \frac{n(x=j)}{N}$$

- Placing a Dirichlet prior over $\mathbf{p}$ we obtain a posterior

$$p(\mathbf{p}|x_1, \ldots, x_N, \boldsymbol{\alpha}) \propto \prod_{j=1}^{D} p_j^{\alpha_j + n(x=j) - 1}$$

which is again a Dirichlet distribution with *pseudocounts* $\boldsymbol{\alpha}$

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 (take notes)
- Exercise: conjugate priors for the univariate normal

The most basic problem

- We are given samples of a certain quantity from two conditions
- E.g., we have measurements of a protein expression x in two cohorts (treated and control)
- We want to know whether the differences observed between the two populations are statistically significant, so x may be used as a biomarker
- How can we do that since we know nothing of the distribution of x?

t-tests and p-values

- We know how to tell whether two Gaussian distributed samples are different
- Use the paired t-test

$$t = \frac{\hat{\mu_1} - \hat{\mu_2}}{\sqrt{\frac{1}{N}\left(\sigma_1^2 + \sigma_2^2\right)}}$$

- The value of t measures how distant the two samples are; it is distributed according to a Student t distribution with $N - 1$ degrees of freedom
- Looking up on a table we can get the p-value, *the probability that a value greater or equal to t would have been obtained at random*

In practice

- It may be useful to transform the data in a way that it becomes approximately Gaussian (e.g. take log of positive numbers)
- Tests of Gaussianity exist (e.g. Kolmogorov-Smirnov)
- One of the main practical uses of testing is for experimental design, e.g. telling the experimentalist how many more samples are needed to make mean differences statistically significant
- Notice that t grows as $\sqrt{N}$ so increasing N we get more and more statistically significant results

Problem formulation

- Data in biology often very high dimensional with very few samples
- E.g., in a cancer study, we could have $\sim$40 subjects and 10000 features (genes) each
- Find a suitable 2D projection of the data that highlights structure
- The projection also identifies the most relevant features

Principal Component Analysis

- A plausible assumption is that the interesting directions are the ones with the greatest variation
- The empirical covariance of a data set $\mathbf{x}_i$ with mean $\hat{\mu}$ is

$$\hat{\Sigma} = \frac{1}{N-1} \sum_{i=1}^{N} (\mathbf{x} - \hat{\mu}) (\mathbf{x} - \hat{\mu})^T$$

- The directions that maximise the projected variance satisfy

$$\hat{\Sigma} U = \Lambda U$$

with Λ a diagonal matrix containing the largest eigenvalues of the empirical covariance

An algorithm for PCA

- Compute empirical mean as $\hat{\mu} = \frac{1}{N} \sum \mathbf{x}_i$
- Compute empirical covariance $\hat{\Sigma}$
- Compute first two eigenvectors $\mathbf{u}_{1,2}$ of $\hat{\Sigma}$
- Compute projected 2D data set as $\hat{x}_i = (\mathbf{x}_i^T \mathbf{u}_1, \mathbf{x}_i^T \mathbf{u}_2)$
- PCA was introduced for the analysis of microarray data in Alter et al, PNAS 97(18) (2000)

Extensions to PCA

- Linear/ global/ Gaussian structure of PCA potentially problematic
- Many extensions proposed in ML
- Kernel PCA maps data in higher dimensional space through non-linear map and then applies PCA (Scholkopf et al, Neural Computation 10(5), 1998)
- Other methods use local structure, e.g. Locally Linear Embeddings (Roweis and Saul, Science 290, 2000), Maximum Variance Unfolding (Weinberger and Saul, CVPR 2004)

Problem formulation

- Given expression data, we want to identify subgroups
- Lack knowledge of number of groups
- Need a greedy, agglomerative procedure

Pearson's correlation

- A frequently used measure of similarity is *Pearson's correlation*
- Given two vectors $\mathbf{x}, \mathbf{y}$, we view them as two zero-mean random variables
- The variance of each vector is its length (as we assume zero mean),$\sigma_{\mathbf{x}} = \sqrt{\mathbf{x}^T\mathbf{x}}$
- The correlation coefficient is then

$$r = \frac{\mathbf{x}^T\mathbf{y}}{\sigma_{\mathbf{x}}\sigma_{\mathbf{y}}} = \cos(\phi)$$

the cosine of the angle between the vectors (see wikipedia page for a good review)

HC algorithm

- Compute pairwise Pearson's correlations
- Find highest correlation pair; merge them by computing mean
- Compute Pearson's of new item with other items
- Repeat previous steps until only one item left

Reflections on HC

- The HC algorithm gives us a tree representation of the data (dendrogram)
- We can keep track of the correlation in the branch length
- Complexity of HC?
- You can cut the tree at a desired level of correlation/ number of clusters

Next lecture

- Next week's lecture is a guest lecture on microarray technology
- Next core lecture will be in two weeks' time
- We will be looking at networks
- Please look at Shannon et al, Genome Research 13:2498-2504 (2003), available from www.cytoscape.org