How do you tell when two samples come from different distributions?

Chandi Bhandari, Rahul Kumar, Brian Robinson, and Simon Stolarczyk

November 23, 2015

Motivating Scenario

From Gretton et. al.

In bioinformatics, it is of interest to co mpare microarray data from identical tissue types as measured by different laboratories, to detect whether the data may be analysed jointly, or whether differences in experimental procedure have caused systematic differences in the data distributions.

Basic Question

Given two distributions p and q, how do we test whether they are different on the basis of samples drawn from each of them?

$X = (X^1, ..., X^m)$ drawn from p

$Y = (Y^1, ..., Y^n)$ drawn from q

Basic Plotting

Useful for lower dimensional data, but how do we visualize the difference when $p = N_d(\mu, I)$ when $d >> 3$?

Permuation Testing

See book info as well as
http://stats.stackexchange.com/questions/59774/test-whether-variables-follow-the-same-distribution

Kolmogorov-Smirnov Test

Motivating Fact

Expectations over all continuous functions can distinguish probability distributions:

$$p = q \text{ iff. } E_p[f(x)] = E_q[f(y)] \;\; \forall f \in C(X)$$

Mean Maximum Discrepancy

For a set of functions $\mathcal{F}$ define

$$\mathrm{MMD}[\mathcal{F}, p, q] = \sup_{f \in F}(E_p[f(x)] - E_q[f(y)])$$

MMD Estimator

$$MMD_b[\mathcal{F}, p, q] = \sup_{f \in F}\left(\frac{1}{m}\sum_{i=1}^{m} f(x_i) - \frac{1}{n}\sum_{i=1}^{n} f(y_i)\right)$$

How to choose $\mathcal{F}$

We need something computationally feasible. We want our space $\mathcal{F}$ to be a Hilbert Space with the nice property that taking the expectation of any function is the same as the inner product with some special function

$$E_x f = \langle f, \mu_p \rangle_{\mathcal{H}}$$

and we want

$$k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$$

Estimator

$$
\begin{aligned}
MMD_u^2[\mathcal{F}, X, Y] = & \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j \neq i}^{m} k(x_i, x_j) \\
& + \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} k(y_i, y_j) \\
& - \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} k(x_i, y_j)
\end{aligned}
$$

Linear Estimator

$$MMD_l^2[\mathcal{F}, X, Y] = \frac{2}{m} \sum_{i=1}^{m/2} h((x_{2i-1}, y_{2i-1}), (x_{2i}, y_{2i}))$$

where

$z \sim (x, y), h_(z_i, z_j) := k(x_i, x_j) + k(y_i, y_j) - k(x_i, y_j) - k(x_j, y_i)$

Applying the tests to artificially generated data

Apply the test to experimental data

- Medical data
- The distributions for connections on a graph.
- other from Dr. Fu