Lu.Getz.Miska.Nature.June.2005.PDT.mRNA

Module name: Lu.Getz.Miska.Nature.June.2005.PDT.mRNA
Description: PNN Prediction using mRNA
Author: Gad Getz (Broad Institute), gp-help@broad.mit.edu

Summary

The following description of the analysis is from the supplementary material (http://www.broad.mit.edu/mpr/publications/projects/microRNA/Supplementary_Notes.pdf) of the paper (1):

A two-class PNN [7] prediction was calculated based on the following class posterior probability:

$$P(c \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid c)P(c)}{\sum_{c'} P(\mathbf{x} \mid c')P(c')} = \frac{\frac{P(c)}{n_c}\sum_{i:\bar{y}_i \in c} \exp\left(-\mathrm{D}(\mathbf{x},\mathbf{y}_i)^2 / 2\sigma^2\right)}{\sum_{c'}\left[\frac{P(c')}{n_{c'}}\sum_{i:\bar{y}_i \in c'} \exp\left(-\mathrm{D}(\mathbf{x},\mathbf{y}_i)^2 / 2\sigma^2\right)\right]},$$

where **x** is the predicted sample and *c* is the class for which the posterior probability is calculated. The training set samples are $\mathbf{y}_i$, n_c is the number of samples of class *c* in the training set, and D(**x**,$\mathbf{y}_i$) is the distance between the predicted sample and training sample *i*. In our case, the sum in the denominator (of c') is over two class values, since we predict a sample either to belong or not to belong to a specific tissue-type. Note that the first step is derived using Bayes rule which allows to incorporate a prior probability for each class, $P(c)$. We used a uniform prior over all 11 tissue-types which translated to 1/11 for being in a certain type and 10/11 for not being in that type. We did not use the tissue-type frequencies in the training set since they likely do not represent the frequencies of different tumors in the general population.

Multi-class prediction using PNN was achieved by breaking down the question into multiple one vs. the rest (OVR) predictions. To perform PNN OVR two-class classification, we built a model based on the training set. This model has two parameters: the number of features used, and σ (the standard deviation of the Gaussian kernel which is used to calculate the contribution of each training sample to the classification). The optimal parameters (for each OVR classifier) were selected using a leave-one-out cross-validation procedure from all possible parameter-pairs in which the number of features ranges from 2 to 30 in steps of 2 and σ takes the values from 1 to 4 times the median nearest neighbor distance, in steps of 0.5 (a total number of 105 combinations). The best model was determined by (i) the fewest number of leave-one-out errors on the training set, which include both false-positive and false-negative errors with the same weight, and (ii) among all conditions with the same error rate, the parameters that gave rise to the maximal mean log-likelihood of the training set were selected. The mean log-likelihood is defined as $L[\{\mathbf{x}_i\};M] = \frac{1}{\#\text{of training examples}}\sum_i \log(P_M(\mathrm{c}_i \mid \mathbf{x}_i))$ where c_i is the true class of sample $\mathbf{x}_i$ and the probability is evaluated using the model *M*. The top n features were selected using the variance-thresholded t-test score in a balanced manner; n/2 features with the top positive scores and n/2 features with most negative scores. The cosine distance measure was used; D(**x**,$\mathbf{y}_i$)=1-cosine(**x**,$\mathbf{y}_i$).

References:

GenePattern

- Lu, Getz, Miska, et al. “MicroRNA Expression Profiles Classify Human Cancers,” Nature 435, 834-838 (9 June 2005)