Advanced Genome Bioinformatics

Naive-Bayes Assignment

Modeling tumor types with a Naive-Bayes model

Introduction

Tumors originate from genetic lesions, which affect multiple pathways that contribute to create and sustain the cancerous phenotype. Some of these alterations are related to the activation of specific aggresive phenotypes, and are often reflected in gene expression changes. Additionally, alterations in the expression of genes may be triggered by environmental pressures, such as hypoxia, inflammation or metabolic stress, and may facilitate tumor spread and metastasis or lead to stem cell like properties like self-renewal. Accordingly, the development of gene expression signatures has been instrumental for the molecular characterization of tumors and have improved their classification. In this assignment we aim to study the patterns of gene expression in multiple tumor types to investigate whether they can be used as predictive signatures.

We have calculated a normalized expression value per gene in a large number of patient samples separated by tumor type, using data from The Cancer Genome Atlas (TCGA) project. The TCGA project is a comprehensive and coordinated effort to accelerate the understanding of the molecular basis of cancer through the application of genome analysis technologies, including large-scale genome, exome and RNA sequencing. TCGA is a joint effort of the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI). You can find more information here.

In particular, we have obtained expression data for all genes for 8 different tumor types (using the TCGA nomenclature):

1. BRCA: Breast Invasive Carcinoma
2. COAD: Colon adenocarcinoma
3. HNSC: Head and neck squamous cell carcinoma
4. KIRC: Kidney renal clear cell carcinoma
5. LUAD: Lung adenocarcinoma
6. LUSC: Lung squamous cell carcinoma
7. PRAD: Prostate adenocarcinoma
8. THCA: Thyroid carcinoma

These are tab-delimited files. Each file contains a first raw with the list of samples. Each sample is labeled with an alphanumeric identifier appended by `N` or `T` depending of whether the sample is a normal or a tumor sample, respectively. Subsequent lines contain a gene identifier and then an expression value Z-score per sample, in the same order as the sample identifiers of the same line. (The first row has one column less):

```
A0B3N   A0CEN   A0DLT   A0E0T   ...
NUMB|8650       0.726962953358609     -0.819445899559172     -2.52666637657443      -4.34178657444036      ...
QKI|9444       -0.225271888363044     -0.46746965408111       4.98501111091547      -2.45043798356706      ...
MYC|4609        0.95324816129648       0.317154963268593     2.4732208853073        3.17861940608758      ...
...
```

Genes are labeled by their HGNC gene name and the Entrez ID as provided by NCBI (see http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3013746/ for further details).

Expression values are given as z-scores, which is a way to standardize the values from a distribution:

$$z = \frac{x - \mu}{\sigma}$$

μ = Mean
σ = Standard Deviation

In this case the we have used a robust Z-score, which is calculated using the median instead of the mean, and using the median absolut deviation (MAD) instead of the standard deviation:

```
Robust Z-score = ( x - median )/(1.486 * MAD)
```

where the MAD is the median value of the differences in absolute value between the median and the values:

$$\mathrm{MAD} = \mathrm{median}_i \left(\left| X_i - \mathrm{median}_j (X_j) \right| \right),$$

For each gene, the distribution used for calculating the median and MAD values is the distribution of expression values only in the normal samples. Thus in the tumor samples, **the z-score represents the deviation of expression with respect to the normal expression values**. We will be only interested in the expression values in tumors.

Objectives

The objective of this assignment is to build a Naive Bayes model to predict the tumor type from the gene expression values of a sample.

1. Separate the datasets into balanced sets for training and testing. A subset of tumor samples will be used for training and a different subset of tumor samples for testing. Make sure that for each tumor type the same number of samples are used for training and for testing. The class values will be the tumor types.
2. The attributes that describe each example (sample) will be the gene expression values. We will discretize the gene expression z-scores into three possible values: **up**, **down** and **no-change**, according to whether the z-score is greater than 2 (z>2), smaller than -2 (z<-2), or between these two values (-2inf, `NA`.
3. For each class (tumor type), and for each attribute (gene), you will have to measure the likelihoods of each value in each class. For instance, for gene `g`, we will measure:

   ```
   P(g_up|brca), P(g_down|brca), P(g_no-change|brca),
   P(g_up|coad), P(g_down|coad), P(g_no-change|coad),
   ...
   ```

4. Using as attributes the **discretized expression** of genes, and using **Mutual Information** on the training set, determine which attributes are the most informative to separate between the tumor types. The Mututal Information provides a single value per gene, which gives a sense of how well the discretized gene expression (attribute values) are associated to the tumor types (class values). Recall that for a given set of classification values `S` and an attribute `A` (gene), the Mutual Information is defined as:

   ```
   MI(S,A) = H(S) - H(S|A)
   ```

   where the relative entropy is calculated as (see the course slides):

   ```
   H(S|A) =    P(brca, g_up) log ( P(brca, g_up) / P(g_up) )  +  P(brca, g_down) log ( P(brca, g_down) / P(g_down) )  + ...
            +  P(coad, g_up) log ( P(coad, g_up) / P(g_up) )  +  P(coad, g_down) log ( P(coad, g_down) / P(g_down) )  + ...
            +  P(hnsc, g_up) log ( P(hnsc, g_up) / P(g_up) )  +  ...
   ```

5. Using the best predictive attributes, build a Naive Bayes model with the training set and use it to predict the tumor type on the testing set. The output of the program should be the resulting classification for each test case using the Naive Bayes classifier, together with a score and the real label. Remember that the scores can be transformed into a probability. The output should be of the form:

   ```
   score    prediction   label   patient

   -20.04   brca         brca    A1RHT
   -30.03   coad         brca    A203T
   -21.32   thca         coad    A39GT
   ...
   ```

6. Consider the use of pseudocounts and discuss whether they are necessary or not.
7. Determine the accuracy of the model by computing the coincidences and the discrepancies between the predicted tumor types and the actual type (label). You can calculate the following quantities:

   - TP (true positives): number of tumor type labels that we predict correctly.
   - FP (false positives): number of predicted tumor types that do not agree with the actual type.
   - FN (false negatives): number of real donor sites incorrectly predicted (i.e., missed). Note that in this case, FP = FN.

8. Discuss the choice of a score (or probability) cut-off to select your predictions and thereby reduce the number of false positives. Can you find an optimal cut-off?
9. Discuss whether this is a good classifier or not. Can you propose a way to improve the classifier?

Important

Since we are going to multiply probabilities, we will obtain in general very small numbers. This can become a problem as computers have a limitation in the number of decimals they can handle. A solution to this problem is to consider the logarithm of the probabilities. The products become sums and the maximization procedure to select the best hypothesis remain the same.

References

Teschendorff AE, Miremadi A, Pinder SE, Ellis IO, Caldas C. An immune response gene expression module identifies a good prognosis subtype in estrogen receptor negative breast cancer. Genome Biol. 2007;8(8):R157.

Taherian-Fard A, Srihari S, Ragan MA. Breast cancer classification: linking molecular mechanisms to disease prognosis. Brief Bioinform. 2015 May;16(3):461-74.

Swanton C, Caldas C. Molecular classification of solid tumours: towards pathway-driven therapeutics. Br J Cancer. 2009 May 19;100(10):1517-22. doi: 10.1038/sj.bjc.6605031.
