UNIVERSITY OF CALIFORNIA, DAVIS
DEPARTMENT OF COMPUTER SCIENCE

ECS 171: Homework Set 3

Instructor: Ilias Tagkopoulos
TAs: Ameen Eetemadi, Nicholas Joodi, Cheng-En Tan
{eetemadi, npjoodi, cetan}@ucdavis.edu

Homework is due on November 28, 2017

General Instructions: The homework package should be submitted electronically through Canvas. Each submission should be a zip file that includes the following: (a) a report in pdf format ("report_HW3.pdf") that includes your answers to all questions, plots, figures and any instructions to run your code, (b) the python code files. Please note: (a) do not include any other files, for instance files that we have provided such as datasets, (b) each function should be written in a separate file, with the appropriate remarks in the code so it is generally understand-able (what it does, how it does it), (c) do not use any toolbox unless is it explicitly allowed in the homework description. Late submissions have 20% penalty per day.

1 REATING A VIRTUAL CELL: PREDICTING PHENOTYPIC AND ENVIRONMENTAL CHARACTERISTICS [80PT]

In this exercise, you will use a set of 223 transcriptional profiling samples from the gram-negative bacterium *Escherichia coli*. *E. coli* is the most well-studied organism with great importance to human health and biotechnology. This meta-dataset has been created by curating several published datasets and annotating the entries with meta-data. It contains 4502 features, the first 6 corresponding to gene ID, strain, medium, environmental and genetic perturbation, and information about the growth rate. The last 4496 entries correspond to the expression of all genes in the bacterium. The dataset can be downloaded from Smartsite (under the HW3 folder). The

main file is "ecs171.dataset.txt" (a xls version is also available). the "ecs171.readme.txt" and "ecs171.genes.txt" files has the definition of the features and gene names, respectively. Perform and report (code and results) the following:

1. Create a predictor of the bacterial growth attribute by using only the expression of the genes as attributes. Not all genes are informative for this task, so use a regularized regression technique (lasso, elastic net, ridge) and explain what it does (we have not covered the specifics of each method, so you have to do some reading). Which one is the optimal constrained parameter value (usually denoted by λ)? Report the number of features that have non-zero coefficients and the 10-fold cross-validation generalization error of the technique. [20pt]

2. Extend your predictor to report the confidence interval of the prediction by using the bootstrapping method. Clearly state the methodology and your assumptions. [10pt]

3. What is the predicted growth for a bacterium whose genes are expressed exactly at the mean expression value? That is, for any given gene, its expression is equal to the gene mean expression across all the samples. [5pt]

4. Create four separate SVM classifiers to categorize the strain type, medium type, environmental and gene perturbation, given all the gene transcriptional profiles. The classifier should select as features a small subset of the genes, either by performing feature selection (wrapper method) or by using only the non-zero weighted features from the regularized regression technique of the first aim. For each classifier (4 total) report the number of features and the classification performance though 10-fold cross-validation by plotting the ROC and PR curves and reporting the AUC/AUPRC values.[20pt]

5. Create one composite SVM classifier to simultaneously predict medium and environmental perturbations and report the 10-fold cross-validation AUC/AUPRC value. Does this classifier perform better or worse that the two individual classifiers together for these predictions? That is, are we better off building one composite or two separate classifiers to simultaneously predict these two features? What is the baseline prediction performance (null hypothesis)? [15pt]

6. Perform Principal Component Analysis, keeping only the 3 Principal Components (PCs) as features for the SVM classifier (no other features except of those three). Report the 10-fold cross-validation AUC/AUPRC value and plot the ROC/PR curves on the same plot as before. Do the PCs retain most of the classification performance while reducing the dimensionality? [10pt]

GOOD LUCK!