Course Project - CS534 Machine Learning
Spring 2019

Objectives

The course project provides an opportunity to apply principals learned in class and to gain practical experience analyzing real-world data. You will learn how to deal with issues like data preprocessing and normalization, how to formulate and execute validation and model selection protocols, and how to test hypotheses about learning algorithms.

Predicting clinical outcomes of breast cancer patients

This project will investigate predicting the survival of patients diagnosed with breast cancer using the METABRIC (Molecular Taxonomy of Breast Cancer International Consortium) dataset [1, 2]. METABRIC contains clinical and demographic data, DNA sequencing, copy-number, and gene expression data for over 2000 patients.

In this project you will construct two different types of models:

1. **Low-dimensional clinical data model**. This model will be constructed from features within the clinical data file (data_clinical_sample.txt).
2. **High-dimensional molecular model**. This model will be constructed from some combination of features within the genomic data files (data_mutations_mskcc.txt, data_CNA.txt, data_expression_median.txt).

These two scenarios present different challenges. The clinical data contains many entries that are highly predictive, and that have been refined by decades of research and medical practice. This data also requires significantly more pre-processing to handle encoding of categorical variables and missing data. The genomic data is more readily analyzed, but contains a much larger number of features and making learning more difficult. This genomic data may contain untapped predictive information that could improve how we classify breast cancers. A complete description of these files and their contents is presented on the following pages.

You will measure two types of accuracy in this project - one related to to classification and the other related to regression:

1. **10-year survival.** Measure the Receiver-Operating Characteristic Area-Under-Curve (AUC) of binary classification of patients surviving longer than 10 years.
2. **Rank-ordering of events.** Measure the concordance between predicted risks of patients, and the rank-ordering of death events using Concordance Index (CI) [3].

For 10-year survival, you can either generate a binary classifier directly, or you can transform and calibrate the outputs of a regression model. For rank-ordering you can directly fit a regression model (using Cox likelihood or other), use the output of your binary classifier, or take a multi-class approach that assigns patients into > 2 ordered risk groups.

Requirements

Your project must include the following elements:

1. **"Stretch algorithms"** Each person on the team must implement some advanced method (subject to approval). *Paper(s) should be the primary reference for this implementation.*
2. **Off-the shelf algorithms** The advanced algorithms should be compared to some off-the-shelf algorithm available in libraries like scikit-learn. A rationale should be provided for choice of algorithm.
3. **Model selection and validation protocol** The team must draft a detailed protocol for selecting model hyperparameters and for reporting accuracy.

Ideas

What you do beyond the requirements is up to you. Some ideas for topics to investigate:

1. **Regularization.** Constraining models to avoid overfitting the training data.
2. **Causal inference.** Understanding the relationship between treatment and outcomes.
3. **Model interpretation.** Explaining what features are important in your model.
4. **Data augmentation.** Increasing the effective size of your dataset through GANs or other methods.
5. **Incorporating prior knowlege** Using biological knowledge and databases to code mutations.

Project teams

You will work in teams of 4 people to carry out the design, implementation, experiments, and writeup. Work should be distributed fairly and equitably. Team members will evaluate each other's contributions as part of the grading process (see **Grading** below).

Grading and deadlines

Project Plan (25%) - due 2/18 A two-page (max) description of topics you plan to investigate, the methods and tools that you plan to use, and a timeline of project milestones.

Presentation (25%) A 20-30 minute team presentation of your project and findings. These will be scheduled for the last regular class sessions and if necessary the final exam period.

Report (50%) - due 5/1. A written report including the following sections: Introduction, Methods, Validation and Model Selection Plan, Results, Discussion, and Bibliography. Use the principals and concepts to reason about the problem and performance of the algorithms. Contrast the performance of the clinical and genomic models, and compare the stretch algorithms to each other and the off-the-shelf algorithm. Providing insights into the results is key. Software should be submitted along with the report in electronic format.

Team reviews - due 5/1 Each person will submit a team review form for their teammates. The % score that each person receives will be used to weight their final project grade. *T*he instructor is available to mediate disputes.

Academic misconduct You are free to discuss the projects with anyone and can consult any resource. Do not plagiarize in your project report. Your stretch algorithm implementation should also be primarily done by referencing a paper, although some consultation with other codebases is acceptable (cutting & pasting is not however). The guiding principal of understanding your work applies.

References

[1] Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y, Gräf S, Ha G, Haffari G, Bashashati A, Russell R, McKinney S, METABRIC Group, Langerød A, Green A, Provenzano E, Wishart G, Pinder S, Watson P, Markowetz F, Murphy L, Ellis I, Purushotham A, Børresen-Dale AL, Brenton JD, Tavaré S, Caldas C, and Aparicio S. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, (486), June 2012.

[2] Pereira B, Chin SF, Rueda OM, Vollan HK, Provenzano E, Bardwell HA, Pugh M, Jones L, Russell R, Sammut SJ, Tsui DW, Liu B, Dawson SJ, Abraham J, Northen H, Peden JF, Mukherjee A, Turashvili G, Green AR, McKinney S, Oloumi A, Shah S, Rosenfeld N, Murphy L, Bentley DR, Ellis IO, Purushotham A, Pinder SE, Børresen-Dale AL, Earl HM, Pharoah PD, Ross MT, Aparicio S, and Caldas C. The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. *Nat. Comm.*., (11479), May 2016.

[3] Harrell FE, Califf RM, Pryor DB, Lee KL, and Rosati RA. Evaluating the yield of medical tests. *Journal of the American Medical Association*, 247(18), May 1982.

Data descriptions

data_clinical_patient.txt

PATIENT_ID - A unique identifier for each patient / sample.

LYMPH_NODES_EXAMINED_POSITIVE - The number of lymph nodes found to contain cancer though pathologic examination.

NPI - The Nottingham prognostic index. This is a score based on the size of the lesion, the number of positive nodes, and the tumor grade.

CELLULARITY - The cellular density of tumor assessed by histology.

CHEMOTHERAPY - Whether the patient received chemotherapy or not.

COHORT - Which cohort the patient originated from.

ER_IHC - Estrogen Receptor status assessed by immunohistochemical analysis of tissue. If positive, the patient may/will receive hormone therapy.

HER2_SNP6 - HER2 copy-number status assessed by SNP array. If gain, the patient may/will receive hormone therapy.

HORMONE_THERAPY - Did the patient receive hormone therapy.

INFERRED_MENOPAUSAL_STATE - Menopausal state inferred (by age?).

AGE_AT_DIAGNOSIS - Age in years at the time of diagnosis.

OS_MONTHS - Overall survival or time last observed alive in months. **This contains information you are trying to predict - do not include in features**.

OS_STATUS - Vital status at time *OS_MONTHS*. **This contains information you are trying to predict - do not include in features**.

CLAUDIN_SUBTYPE - Genomic classification by PAM50 method.

THREEGENE - Genomic classification by ER/PR/HER2 status.

VITAL_STATUS - Supplements *OS_MONTHS* with cause of death. For 'Living', patient was alive at last followup. For 'Died of Disease', patient died as a result of their breast cancer. For 'Died of Other Causes', patient died from another cause (but possibly disease had influence). For the purposes of this project, you can treat 'Died of Other Causes' as 'Living'. **This contains information you are trying to predict - do not include in features**

LATERALITY - Which side of the body the cancer originated in (using patient's left/right).

RADIO_THERAPY - Did the patient receive radiation therapy.

HISTOLOGICAL_SUBTYPE - The histologic classification determined by histology.

BREAST_SURGERY - What type of treatment the patient received.

data_mutations_mskcc.txt

Each row in this file describes a mutation in the DNA coding region of a gene observed in a single patient. If using this data, you will have to transform this into a 2D array of patients / features. The simplest encoding approach would just be a binary indicator of whether the mutation is present or not. More sophisticated approaches may encode mutation type, exon, and information from public databases.

#Sequenced_Samples - contains a list of the IDs of sequenced samples (some samples were not sequenced). This is needed to determine whether a sample has no mutations, or simply wasn't sampled.

Hugo_Symbol - the standardized HUGO Gene Nomenclature Committee name for the gene where the mutation was observed.

Consequence - the impact of the mutation. Single-base substitutions may be 'missense' and alter the amino acid, or may be 'synonmous' and not alter the amino acid. Other mutations may entirely alter the reading frame or truncate the gene.

Variant_Classification - classification of mutation type. For more on possible mutation types, see the NIH Genetics Home Reference.

data_cna_mskcc.txt

This is a simple tab-delimited text file describing the number of copies of each gene in each patient sample. Each row here is a gene, and each column (3 and beyond) is a patient ID. A value of -2 indicates deletion of both copies of the gene from the DNA, a -1 indicates deletion of one copy, a 0 indicates 2 copies (normal), a 1 indicates a "low-level amplification" where more than 2 copies are present, and a 2 indicates a "high-level amplification" where many more than 2 copies are present.

data_expression_median.txt

This is a simple tab-delimited text file describing the observed abundance of mRNA for each gene in each patient sample. Each row here is a gene, and each column (3 and beyond) is a patient ID. These values have been normalized in some way, but you should investigate and apply your own normalization.