Bioinformatics: Supervised Machine Learning

Swakkhar Shatabda

Department of Computer Science and Engineering
BRAC University

References

A problem of classification

RNA-Seq Data - Classification

Imagine you have **RNA-seq** of a collection of **labeled normal** lung and **lung cancer** tissues. Given a new sample of RNA-seq from the lung with unknown diagnosis, will you be able to predict based on the existing labeled samples and the expression data whether the new sample is normal or tumor? This is a sample classification problem, and it could be solved using **unsupervised** and **supervised** learning approaches.

- **Unsupervised learning** is clustering or dimension reduction. You can use hierarchical clustering, k-means, fuzzy k-means after MDS, or PCA. After clustering and projection the data to lower dimensions, you examine the labels of the known samples. Then you can assign label to the unknown sample based on its distance to the known samples.

Supervised Classification Problem

- Supervised learning considers the labels with known samples and tries to identify features that can separate the samples by the label.
- Cross validation is conducted to evaluate the performance of different approaches and avoid over fitting.

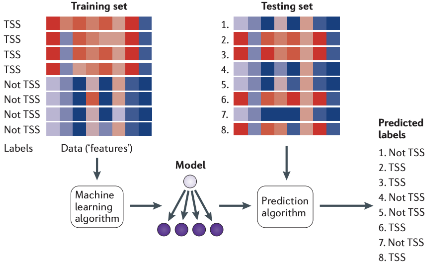
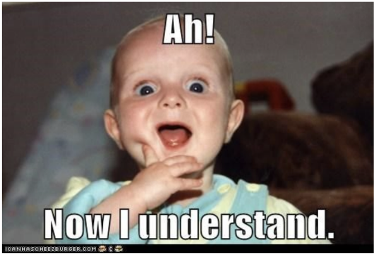


Figure 1 | **A canonical example of a machine learning application.** A training set of DNA sequences is provided as input to a learning procedure, along with binary labels indicating whether each sequence is centred on a transcription start site (TSS) or not. The learning algorithm produces a model that can then be subsequently used, in conjunction with a prediction algorithm, to assign predicted labels (such as 'TSS' or 'not TSS') to unlabelled test sequences. In the figure, the red–blue gradient might represent, for example, the scores of various motif models (one per column) against the DNA sequence.

https://www.nature.com/articles/nrg3920

Steps in Supervised Learning Based Classification

Select a Problem → Gather Data → Generate Features → Select Features → Design Classifier → Test your System → Deploy App.



Step 1: Problem Formulation

- We have molecules DNA, RNA, Proteins and Drugs
- Predict functions?
  - DNA - Origin of Replication, Promoters, Enhancers, Recombination hot-spots, methylation spots, protein binding spots, etc.
  - RNA - location, coding vs non-coding, editing, types, etc
  - Proteins - Inflammatory vs anti-inflammatory, toxicity, anti-viral, bacteriphage, dna binding, location etc.
- Predict structures?
- Predict interactions: drug-target, protein-protein, etc.
- In each of these problems, the input is a sequence and output is a class or properties (binary/ multiclass/ multi-label)
- Sometimes from the pipelines of genomic technology data, we formulate novel problems, often we select already defined problems.

Step 2: Data Collection

- We need validated sources.
- The results are often validated *in vitro* and then published in the literature.
- People are there who curate those from literature and submit to databases.
- From these, online databases, we can collect sequences and labels.
- There are two problems:
  - **Imbalance:** Often we have only positive data, we rarely have negative data or the other way.
    - Use undersampling or oversampling
  - **Redundancy:** Often there are homologous sequences. We need to remove them.
- For the problems that are already defined in the literature often we have benchmark datasets, we can use them, or curate data to enhance them.

Step 3: Feature Extraction

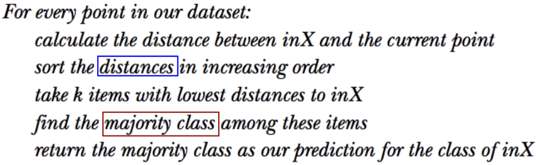
- Features are properties of sequences or molecules that play a role in classification
- We need to convert everything to a number.
- Different feature types are possible:
  - k-mer composition (sequence based)
  - Physico-chemical properties
  - Structural Properties
  - Embeddings
- We like features that are cheap to generate and human understandable
- Often we will require scaling of the features.
- There are several tools to generate / extract features.

Step 4: Feature Selection

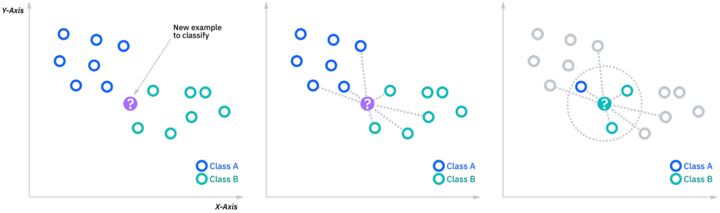
- Often, the number of features become large, the models sufffer due to that. We need to select effective features.
- Generally methods are of two types:
  - Filter Method
  - Wrapper Method
- We can also deploy PCA or similar dimensionality reduction methods.
- One caution: supervised feature selection should be only based on the train dataset.

Step 5: Classification Algorithm

- There are many algorithms, we try to use the simplest that explains the data best!
- Lets try KNN.



```
For every point in our dataset:
    calculate the distance between inX and the current point
    sort the distances in increasing order
    take k items with lowest distances to inX
    find the majority class among these items
    return the majority class as our prediction for the class of inX
```



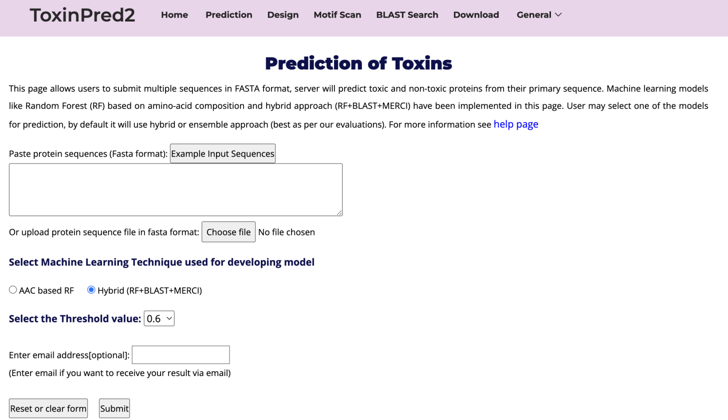
Step 6: Test Your System

- For testing we need to fix two things: metrics and sampling
- Metrics:
  - Accuracy/Error
  - F1, MCC
  - auROC, auPR
- In sampling, we use different techniques
  - Train-Test Split
  - Train-Val-Test Split
  - Cross-Validation

https://colab.research.google.com/drive/12t82nozqlTUDK6tS5mrhP4RSoD8bguP_?usp=sharing

Step 7: Deployment

https://webs.iiitd.edu.in/raghava/toxinpred2/batch.html



Prediction of Toxins

This page allows users to submit multiple sequences in FASTA format, server will predict toxic and non-toxic proteins from their primary sequence. Machine learning models like Random Forest (RF) based on amino-acid composition and hybrid approach (RF+BLAST+MERCI) have been implemented in this page. User may select one of the models for prediction, by default it will use hybrid or ensemble approach (best as per our evaluations). For more information see help page

Paste protein sequences (Fasta format): Example Input Sequences

Or upload protein sequence file in fasta format: Choose file No file chosen

Select Machine Learning Technique used for developing model

○ AAC based RF ● Hybrid (RF+BLAST+MERCI)

Select the Threshold value: 0.6

Enter email address[optional]:

(Enter email if you want to receive your result via email)

Reset or clear form Submit