Machine learning applications in genomics: practical issues & challenges

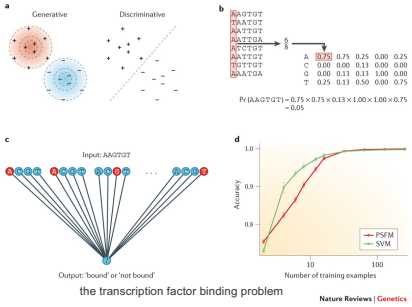
Yuzhen Ye

School of Informatics and Computing, Indiana University

Reference

- Machine learning applications in genetics and genomics
  Nature Reviews Genetics 16, 321–332 (2015) doi:10.1038/nrg3920
- Topics
  - Generative vs discriminative models
  - Incorporating prior knowledge
  - Handling heterogeneous data
  - Feature selection
  - Imbalanced class sizes
  - Performance measure
  - Handling missing data
  - Modelling dependence among examples (genes)

Generative and discriminative models



Generative and discriminative models are different in their interpretability and prediction accuracy.

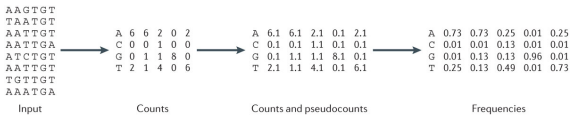
the transcription factor binding problem

Nature Reviews | Genetics

Generative vs discriminative models

- The generative modelling approach offers several compelling benefits.
  - The generative description of the data implies that the model parameters have well-defined semantics relative to the generative process.
    - As shown in the **transcription factor binding problem**, the model not only predicts the locations to which a given transcription factor binds but also explains why the transcription factor binds there.
    - If we compare two different potential binding sites, we can see that the model prefers one site over another and also that the reason is, for example, the preference for an adenine rather than a thymine at position 7 of the motif.
  - Generative models are frequently stated in terms of probabilities, and the probabilistic framework provides a principled way to handle problems like missing data.
    - For example, it is still possible for a PSFM to make a prediction for a binding site where one or more of the bound residues is unknown. This is accomplished by probabilistically averaging over the missing bases.
    - The output of the probabilistic framework has well-defined, probabilistic semantics, and this can be helpful when making downstream decisions about how much to trust a given prediction.
- The primary benefit of the discriminative modelling approach is that it probably achieves better performance than the generative modelling approach with infinite training data
  - In practice, analogous generative and discriminative approaches often converge to the same solution
  - generative approaches can sometimes perform better with limited training data.
  - when the amount of labelled training data is reasonably large, the discriminative approach will tend to find a better solution

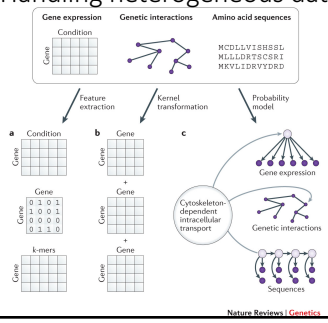
Incorporating prior knowledge



Nature Reviews | Genetics

A simple, principled method for putting a probabilistic prior on a position-specific frequency matrix involves augmenting the observed nucleotide counts with pseudocounts and then computing frequencies with respect to the sum. The magnitude of the pseudocount corresponds to the weight assigned to the prior.

Handling heterogeneous data



These diverse data types can be
1) encoded into fixed-length features,
2) represented using pairwise similarities (that is, kernels) or
3) directly accommodated by a probability model (e.g., Bayesian network).

Nature Reviews | Genetics

Feature selection

- In practice, it is important to distinguish among three distinct motivations for carrying out feature selection.
  - we want to identify a very small set of features that yield the best possible classifier. For example, we may want to produce an inexpensive way to identify a disease phenotype on the basis of the measured expression levels of a handful of genes. Such a classifier, if it is accurate enough, might form the basis of an inexpensive clinical assay.
  - we may want to use the classifier to understand the underlying biology. In this case, we want the feature selection procedure to identify only the genes with expression levels that are actually relevant to the task at hand in the hope that the corresponding functional annotations or biological pathways might provide insights into the aetiology of disease.
  - We often simply want to train the most accurate possible classifier. In this case, we hope that the feature selection enables the classifier to identify and eliminate noisy or redundant features. Researchers are often disappointed to find that feature selection cannot optimally perform more than one of these three tasks simultaneously.
- Feature selection is especially important in the third case because the analysis of high-dimensional data sets, including genomic, epigenomic, proteomic or metabolomic data sets, suffers from the curse of dimensionality — the general observation that many types of analysis become more difficult as the number of input dimensions (that is, data measurements) grows very large.

Feature selection approaches

- Three main categories: Wrappers, Filters, Embedded methods

Wrappers

- **Evaluate** feature sets; select a subset of features that gives the best accuracy
- Exhaustive search -> Exponential problem (M features, 2^M subsets)
- Search strategy
  - Sequential forward selection (evaluates M(M+1)/2 instead of 2^M feature sets)
  - Recursive backward elimination
  - Simulated annealing
  - ….

Filters

- Replace evaluation of models with quick-to-compute statistics
- Examples of filtering criterion
  - Mutual information with target variable
  - Correlation with the target variable
  - chi-square statistic

Imbedded methods

- The classifier performs feature selection as part of the learning procedure
- Example: the logistic LASSO

$$f(\boldsymbol{x}) = \frac{1}{1 + e^{-(\boldsymbol{w}^T \boldsymbol{x})}} = P(Y = 1|\boldsymbol{x})$$

With Error Function:

$$E = \underbrace{-\sum_{i=1}^{N} \{y_i \log f(\boldsymbol{x}_i) + (1 - y_i) \log(1 - f(\boldsymbol{x}_i))\}}_{\text{Cross-entropy error}} + \underbrace{\lambda \sum_{j=1}^{d} |w_j|}_{\text{Regularizing term}}$$

Imbalanced class sizes: Enhancer prediction problem

- A common stumbling block in many applications of machine learning to genomics is the large imbalance (or label skew) in the relative sizes of the groups being classified.
  - Enhancer prediction problem: Starting with a set of 641 known enhancers, the genome can be broken up into 1,000-bp segments and each segment assigned a label ('enhancer' or 'not enhancer') on the basis of whether it overlaps with a known enhancer. This procedure produces 1,711 positive examples and around 3,000,000 negative examples — 2,000 times as many negative examples as positive examples.
  - Assume a classifier achieved an overall accuracy (that is, the percentage of predictions that were correct) of 99.9%. The accuracy seems good, but it is not an appropriate measure because a null classifier that simply predicts everything to be non-enhancers achieves nearly the same accuracy.

Enhancer prediction problem (Cont.)

- It is more appropriate to separately evaluate sensitivity (i.e., the fraction of enhancers detected) and precision (i.e., the percentage of predicted enhancers that are truly enhancers). The balanced classifier described above has a high precision (>99.9%) but a very low sensitivity of 0.5%.
- The behavior of the classifier can be improved by using all of the enhancers for training and then picking a random set of 49,000 non-enhancer positions as negative training examples. However, balancing the classes in this way results in the classifier learning to reproduce this artificially balanced ratio. The resulting classifier achieves much higher sensitivity (81%) but very poor precision (40%); thus, this classifier is not useful for finding enhancers that can be validated experimentally (too many false positives)
- It is possible to trade off sensitivity and precision while retaining the training power of a balanced training set by placing weights on the training examples. For example, using the balanced training set and weighting each negative example 36 times more than a positive example during training resulted in a sensitivity of 53% with a precision of 95%.

Performance measure

- Precision: true positives/(true positives + false positives)
- Recall (or Sensitivity): true positives/(true positives + false negatives)
- F1 score: the harmonic mean of precision and recall

  F1 = 2 * (precision * recall) / (precision + recall)

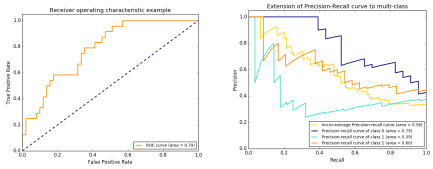


Image credit: scikit-learn online documents

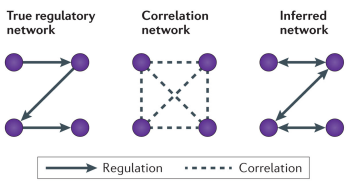
Performance measure

- In general, the most appropriate performance measure depends on the intended application of the classifier.
  - For problems such as identifying which tissue a given cell comes from, it may be equally important to identify rare and abundant tissues, and so the overall number of correct predictions may be the most informative measure of performance.
  - In other problems, such as enhancer detection, predictions in one class may be more important than predictions in another. For example, if positive predictions will be published (and/or experimentally validated), the most appropriate measure may be the sensitivity among a set of predictions with a predetermined precision (for example, 95%).
- A wide variety of performance measures are used in practice, including the F_1 measure, the receiver operating characteristic (ROC) curve and the precision-recall curve, among others. Machine learning classifiers perform best when they are optimized for a realistic performance measure.

Handling missing data

- Missing data values
  - missing at random or for reasons that are unrelated to the task at hand
  - values that, when absent, provide information about the task at hand (e.g., patients become too sick)
- Different ways to deal with missing data
  - Impute the missing values
    - replacing all of the missing values with zero
    - Or use a more sophisticated strategy (e.g., Troyanskaya *et al* used the correlations between data values to impute missing microarray values)
  - Include in the model information about the 'missingness' of each data point.
    - For example, Kircher *et al.* aimed to predict the deleteriousness of mutations based on functional genomic data. For each feature, the authors added a Boolean feature that indicated whether the corresponding feature value was present. The missing values themselves were then replaced with zeroes. An advantage of this approach is that it is applicable regardless of whether the absence of a data point is significant — if it is not, the model will learn to ignore the absence indicator.
  - Use probability models to explicitly model missing data by considering all the potential missing values.
    - Missing data points are handled by summing over all possibilities for that random variable in the model. This approach, called marginalization, represents the case in which a particular variable is unobserved. However, marginalization is only appropriate when data points are missing for reasons that are unrelated to the task at hand. When the presence or absence of a data point is likely to be correlated with the values themselves, incorporating presence or absence explicitly into the model is more appropriate.

Modelling dependence among examples



Methods that infer each relationship in a network separately, such as by computing the correlation between each pair, can be confounded by indirect relationships. Methods that infer the network as a whole can identify only direct relationships. Inferring the direction of causality inherent in networks is generally more challenging than inferring the network structure

Nature Reviews | Genetics

Moving forward

- On the one hand, machine learning methods, which are most effective in the analysis of large, complex data sets, are likely to become ever more important to genomics as more large data sets become available through international collaborative projects, such as the 1000 Genomes Project, the 100,000 Genomes Project, ENCODE, the Roadmap Epigenomics Project and the US National Institutes of Health's 4D Nucleome Initiative.
- On the other hand, even in the presence of massive amounts of data, machine learning techniques are not generally useful when applied in an arbitrary manner. In practice, achieving good performance from a machine learning method usually requires theoretical and practical knowledge of both machine learning methodology and the particular research application area.
- As new technologies for generating large genomic and proteomic data sets emerge — pushing beyond DNA sequencing to mass spectrometry, flow cytometry and high-resolution imaging methods — demand will increase not only for new machine learning methods but also for experts that are capable of applying and adapting them to big data sets. In this sense, both machine learning itself and scientists proficient in these applications are likely to become increasingly important to advancing genetics and genomics.

Nature Reviews Genetics 16, 321–332 (2015) doi:10.1038/nrg3920