Classifiers!!!

BCH394P/364C Systems Biology / Bioinformatics
Edward Marcotte, Univ of Texas at Austin

Clustering = task of grouping a set of objects in such a way that objects in the same group (a cluster) are more similar (in some sense) to each other than to those in other groups (clusters).

VS.

Classification = task of categorizing a new observation, on the basis of a training set of data with observations (or instances) whose categories are known

Adapted from Wikipedia
Remember, for clustering, we had a matrix of data...

\[ M \text{ samples} \]

\[ N \text{ genes} \]

<table>
<thead>
<tr>
<th>Gene 1, sample 1</th>
<th>...</th>
<th>Gene 1, sample j</th>
<th>...</th>
<th>Gene 1, sample M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene 2, sample 1</td>
<td>...</td>
<td>Gene 2, sample j</td>
<td>...</td>
<td>Gene 2, sample M</td>
</tr>
<tr>
<td>Gene 3, sample 1</td>
<td>...</td>
<td>Gene 3, sample j</td>
<td>...</td>
<td>Gene 3, sample M</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Gene i, sample 1</td>
<td>...</td>
<td>Gene i, sample j</td>
<td>...</td>
<td>Gene i, sample M</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Gene N, sample 1</td>
<td>...</td>
<td>Gene N, sample j</td>
<td>...</td>
<td>Gene N, sample M</td>
</tr>
</tbody>
</table>

For yeast, \( N \sim 6,000 \)
For human, \( N \sim 22,000 \)

\( i.e., \) a matrix of \( N \times M \) numbers

We discussed gene expression profiles. Here’s another example of gene features.

\[ M \text{ samples} \]

\[ N \text{ genes} \]

<table>
<thead>
<tr>
<th>Gene 1, sample 1</th>
<th>...</th>
<th>Gene 1, sample j</th>
<th>...</th>
<th>Gene 1, sample M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene 2, sample 1</td>
<td>...</td>
<td>Gene 2, sample j</td>
<td>...</td>
<td>Gene 2, sample M</td>
</tr>
<tr>
<td>Gene 3, sample 1</td>
<td>...</td>
<td>Gene 3, sample j</td>
<td>...</td>
<td>Gene 3, sample M</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Gene i, sample 1</td>
<td>...</td>
<td>Gene i, sample j</td>
<td>...</td>
<td>Gene i, sample M</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Gene N, sample 1</td>
<td>...</td>
<td>Gene N, sample j</td>
<td>...</td>
<td>Gene N, sample M</td>
</tr>
</tbody>
</table>

For yeast, \( N \sim 6,000 \)
For human, \( N \sim 22,000 \)

**Gene expression profiles:** each entry indicates an mRNA’s abundance in a different condition

**Phylogenetic profiles:** each entry indicates whether the gene has homologs in a different organism
This is useful because biological systems tend to be modular and often inherited intact across evolution.

(e.g. you tend to have a flagellum or not)

Many such features are possible...

\[ i.e., \text{a matrix of } N \times M \text{ numbers} \]
We also needed a measure of the similarity between feature vectors. Here are a few (of many) common distance measures used in clustering.

<table>
<thead>
<tr>
<th>Names</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Euclidean distance</td>
<td>$|a - b|_2 = \sqrt{\sum_i (a_i - b_i)^2}$</td>
</tr>
<tr>
<td>Manhattan distance</td>
<td>$|a - b|_1 = \sum_i</td>
</tr>
<tr>
<td>cosine similarity</td>
<td>$\frac{a \cdot b}{|a| |b|}$</td>
</tr>
</tbody>
</table>

Wikipedia

We also needed a measure of the similarity between feature vectors. Here are a few (of many) common distance measures used in clustering.

<table>
<thead>
<tr>
<th>Names</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Euclidean distance</td>
<td>$|a - b|_2 = \sqrt{\sum_i (a_i - b_i)^2}$</td>
</tr>
<tr>
<td>Manhattan distance</td>
<td>$|a - b|_1 = \sum_i</td>
</tr>
<tr>
<td>cosine similarity</td>
<td>$\frac{a \cdot b}{|a| |b|}$</td>
</tr>
</tbody>
</table>

Wikipedia
Clustering refresher: SOM

Clustering refresher: \textit{k}-means
Clustering refresher: *k*-means

One of the simplest classifiers uses the same notion of decision boundaries.
One of the simplest classifiers uses this notion of decision boundaries.

Rather than first clustering, calculate the centroid (mean) of objects with each label.

New observations are classified as belonging to the group whose mean is nearest.

="minimum distance classifier"


One of the simplest classifiers uses this notion of decision boundaries.

For example....

Enzyme-based histochemical analyses were introduced in the 1960s to demonstrate that some leukemias were periodic acid-Schiff positive, whereas others were myeloperoxidase positive...

This provided the first basis for classification of acute leukemias into those arising from lymphoid precursors (acute lymphoblastic leukemia, ALL), or from myeloid precursors (acute myeloid leukemia, AML).”

“Distinguishing ALL from AML is critical for successful treatment...

chemotherapy regimens for ALL generally contain corticosteroids, vincristine, methotrexate, and L-asparaginase, whereas most AML regimens rely on a backbone of daunorubicin and cytarabine (8).

Although remissions can be achieved using ALL therapy for AML (and vice versa), cure rates are markedly diminished, and unwarranted toxicities are encountered.”
Let’s look at a specific historic example:

Take labeled samples, find genes whose abundances separate the samples...

Let’s look at a specific historic example:

Calculate weighted average of indicator genes to assign class of an unknown
PS = (V_{win} - V_{lose}) / (V_{win} + V_{lose}), where \( V_{win} \) and \( V_{lose} \) are the vote totals for the winning and losing classes.

**Fig. 3.** (A) Prediction strengths. The scatterplots show the prediction strengths (PSs) for the samples in cross-validation (left) and on the independent sample (right). Median PS is denoted by a horizontal line. Predictions with PS < 0.3 are considered as uncertain. (B) Genes distinguishing ALL from AML. The SD genes most highly correlated with the ALL-AML class distinction are shown. Each row corresponds to a gene, with the columns corresponding to expression levels in different samples. Expression levels for each gene are normalized across the samples such that the mean is 0 and the SD is 1. Expression levels greater than the mean are shaded in red, and those below the mean are shaded in blue. The scale indicates SDs above or below the mean. The top panel shows genes highly expressed in ALL, the bottom panel shows genes more highly expressed in AML. Although these genes as a group appear correlated with class, no single gene is uniformly expressed across the class.
Cross-validation

Withhold a sample, build a predictor based only on the remaining samples, and predict the class of the withheld sample.

Repeat this process for each sample, then calculate the cumulative or average error rate.

X-fold cross-validation
e.g. 3-fold or 10-fold

Can also withhold 1/X (e.g. 1/3 or 1/10) of sample, build a predictor based only on the remaining samples, and predict the class of the withheld samples.

Repeat this process X times for each withheld fraction of the sample, then calculate the cumulative or average error rate.
**Independent data**

Withhold an entire dataset, build a predictor based only on the remaining samples (the training data).

Test the trained classifier on the independent test data to give a fully independent measure of performance.

You already know how to measure how well these algorithms work (way back in our discussion of gene finding!)...
Recall = \frac{TP}{TP + FN} \quad (= \text{sensitivity})

Precision = \frac{TP}{TP + FP} \quad \text{also called positive predictive value (PPV)}

Sort the data by their classifier score, then step from best to worst and plot the performance:

**Precision-recall curve**

You already know how to measure how well these algorithms work (way back in our discussion of gene finding!)

Another good option:

Sort the data by their classifier score, then step from best to worst and plot the performance:

**ROC curve** (receiver operator characteristic)

1 - Specificity = \frac{FP}{FP + TN} \quad \text{also called False Positive Rate (FPR)}
ROC curve, as you go from stronger to weaker predictions

Thanks to Dariya Sydykova (UT Austin), for her excellent visualizations, available here: https://github.com/dariyasynova/open_projects/tree/master/ROC_animation

ROC curve, as you go from stronger to weaker classifiers

Thanks to Dariya Sydykova (UT Austin), for her excellent visualizations, available here: https://github.com/dariyasynova/open_projects/tree/master/ROC_animation
ROC versus Recall/Precision

The 2 measures are related and both useful. They differ strongly in performance as proportions of positive and negative classes change.

Thanks to Dariya Sydykova (UT Austin), for her excellent visualizations, available here: https://github.com/dariyasydykova/open_projects/tree/master/ROC_animation

ROC versus Recall/Precision

- R/P depends **strongly** on relative rates of the 2 classes
- ROC performance is **independent** of their relative rates

(It may be important or not for your particular problem...)

Thanks to Dariya Sydykova (UT Austin), for her excellent visualizations, available here: https://github.com/dariyasydykova/open_projects/tree/master/ROC_animation
Back to our minimum distance classifier...

Would it work well for this data?

Back to our minimum distance classifier...

How about this data? What might?
Back to our minimum distance classifier...

How about this data? What might?

This is a great case for something called a **k-nearest neighbors classifier**:

For each new object, calculate the *k* closest data points. Let them vote on the label of the new object.
Back to leukemias. There was a follow-up study in 2010:

- Tested clinical use of mRNA expression profiling to subtype leukemias into myeloid/lymphoid
- Meta-analysis of 11 labs, 3 continents, 3,334 patients
- Stage 1 (2,096 patients):
  92.2% classification accuracy for 18 leukemia classes (99.7% median specificity)
- Stage 2 (1,152 patients):
  95.6% median sensitivity and 99.8% median specificity for 14 subtypes of acute leukemia
- Microarrays outperformed routine diagnostics in 29 (57%) of 51 discrepant cases

Conclusion: “Gene expression profiling is a robust technology for the diagnosis of hematologic malignancies with high accuracy”

In practice, if you want to explore classifiers, I also strongly recommend always testing these classifiers:

- Random forests
- Support vector machines (SVM)

These two are surprisingly often the best for many biological classification problems. Weka can do both of them.
The two slide overview of Random forest classifiers:
(1) Construct many decision trees from random subsets of your features. Because the features vary across trees, trees tend to be weak but uncorrelated. 
(2) All the trees “vote” on the answer, majority wins.
The one slide overview of **Support vector machines:**

1. **Goal:** make a linear classifier, choosing a decision boundary that *maximizes* the *distance margin* between classes

   ![Diagram](https://quantdare.com/svm-versus-a-monkey/)

2. But what if the boundary is non-linear? Use **kernels** to implicitly map the data to higher dimension where a linear decision can be made

   ![Diagram](https://quantdare.com/svm-versus-a-monkey/)

   **Maximum margin hyperplane**

In practice, if you want to explore classifiers, I **strongly** recommend the Weka package:

http://www.cs.waikato.ac.nz/ml/weka/

It’s free, and easy to install, use, & troubleshoot. It lets you quickly test many alternative (well-vetted) classifiers, all in a proper cross-validated/precision-recall framework.

Here’s a nice step-by-step intro for biologists:


There’s also a great book to walk you through the entire process. Highly recommended!!!
In Python, you can also use the scikit-learn library: 
https://scikit-learn.org/stable/
Like Weka, it’s free, easy to install & use, and very powerful

I recommend combining it with the Pandas library for data analysis to make it easy to work with big, tabular datasets: 
https://pandas.pydata.org/