Bioinformatics and Biomarker Discovery

Part 1: Foundations

Limsoon Wong
3 September 2014
Themes of Bioinformatics

Bioinformatics =
Data Mgmt +
Knowledge Discovery +
Sequence Analysis +
Physical Modeling + ....

Knowledge Discovery =
Statistics + Algorithms + Databases

Applications include diagnosis, prognosis, & treatment optimization, often thru biomarker discovery
What is Knowledge Discovery?

Jonathan’s blocks
- Hexagon
- Circle
- Triangle
- Cylinder

Jessica’s blocks
- Triangle
- Square
- Cylinder
- Diamond

Whose block is this?

Jonathan’s rules: Blue or Circle
Jessica’s rules: All the rest
What is Knowledge Discovery?

Question: Can you explain how?
Key Steps of Knowledge Discovery

• Training data gathering

• Feature generation
  – k-grams, colour, texture, domain know-how, ...

• Feature selection
  – Entropy, $\chi^2$, CFS, t-test, domain know-how...

• Feature integration
  – SVM, ANN, PCL, CART, C4.5, kNN, ...
What is Accuracy?
What is Accuracy?

Accuracy = \frac{\text{No. of correct predictions}}{\text{No. of predictions}}

= \frac{TP + TN}{TP + TN + FP + FN}
Examples (Unbalanced Population)

<table>
<thead>
<tr>
<th>classifier</th>
<th>TP</th>
<th>TN</th>
<th>FP</th>
<th>FN</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>25</td>
<td>75</td>
<td>75</td>
<td>25</td>
<td>50%</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>150</td>
<td>0</td>
<td>50</td>
<td>75%</td>
</tr>
<tr>
<td>C</td>
<td>50</td>
<td>0</td>
<td>150</td>
<td>0</td>
<td>25%</td>
</tr>
<tr>
<td>D</td>
<td>30</td>
<td>100</td>
<td>50</td>
<td>20</td>
<td>65%</td>
</tr>
</tbody>
</table>

- Clearly, D is better than A
- Is B better than A, C, D?

Exercise: What is B’s Prediction strategy?
What is Sensitivity (aka Recall)?

\[
\text{Sensitivity wrt positives} = \frac{\text{No. of correct positive predictions}}{\text{No. of positives}} = \frac{\text{TP}}{\text{TP} + \text{FN}}
\]

Sometimes sensitivity wrt negatives is termed \textit{specificity}. 

<table>
<thead>
<tr>
<th></th>
<th>predicted as positive</th>
<th>predicted as negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>positive</td>
<td>TP</td>
<td>FN</td>
</tr>
<tr>
<td>negative</td>
<td>FP</td>
<td>TN</td>
</tr>
</tbody>
</table>
**What is Precision?**

<table>
<thead>
<tr>
<th></th>
<th>predicted as positive</th>
<th>predicted as negative</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>positive</strong></td>
<td>TP</td>
<td>FN</td>
</tr>
<tr>
<td><strong>negative</strong></td>
<td>FP</td>
<td>TN</td>
</tr>
</tbody>
</table>

Precision = \[
\frac{\text{No. of correct positive predictions}}{\text{No. of positives predictions}} = \frac{\text{TP}}{\text{TP + FP}}
\]
Unbalanced Population Revisited

<table>
<thead>
<tr>
<th>classifier</th>
<th>TP</th>
<th>TN</th>
<th>FP</th>
<th>FN</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>25</td>
<td>75</td>
<td>75</td>
<td>25</td>
<td>50%</td>
<td>50%</td>
<td>25%</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>150</td>
<td>0</td>
<td>50</td>
<td>75%</td>
<td>0%</td>
<td>ND</td>
</tr>
<tr>
<td>C</td>
<td>50</td>
<td>0</td>
<td>150</td>
<td>0</td>
<td>25%</td>
<td>100%</td>
<td>25%</td>
</tr>
<tr>
<td>D</td>
<td>30</td>
<td>100</td>
<td>50</td>
<td>20</td>
<td>65%</td>
<td>60%</td>
<td>38%</td>
</tr>
</tbody>
</table>

- What are the sensitivity and precision of B and C?
- Is B better than A, C, D?
Comparing Prediction Performance

- **Accuracy is the obvious measure**
  - But it conveys the right intuition only when the positive and negative populations are roughly equal in size

- **Recall and precision together form a better measure**
  - But what do you do when A has better recall than B and B has better precision than A?

So let us look at some alternate measures ....
Adj Accuracy

• Weigh by the importance of the classes

Adjusted accuracy = \alpha \ast \text{Sensitivity} + \beta \ast \text{Specificity}

where \( \alpha + \beta = 1 \)

typically, \( \alpha = \beta = 0.5 \)

<table>
<thead>
<tr>
<th>classifier</th>
<th>TP</th>
<th>TN</th>
<th>FP</th>
<th>FN</th>
<th>Accuracy</th>
<th>Adj Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>25</td>
<td>75</td>
<td>75</td>
<td>25</td>
<td>50%</td>
<td>50%</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>150</td>
<td>0</td>
<td>50</td>
<td>75%</td>
<td>50%</td>
</tr>
<tr>
<td>C</td>
<td>50</td>
<td>0</td>
<td>150</td>
<td>0</td>
<td>25%</td>
<td>50%</td>
</tr>
<tr>
<td>D</td>
<td>30</td>
<td>100</td>
<td>50</td>
<td>20</td>
<td>65%</td>
<td>63%</td>
</tr>
</tbody>
</table>

But people can’t always agree on values for \( \alpha, \beta \)
What is Cross Validation?
Construction of a Classifier

Training samples → Build Classifier → Classifier

Test instance → Apply Classifier → Prediction
Estimate Accuracy: Wrong Way

Training samples -> Build Classifier -> Classifier

Build Classifier -> Apply Classifier -> Predictions

Apply Classifier -> Estimate Accuracy -> Accuracy

Exercise: Why is this way of estimating accuracy wrong? Think of what will happen in the case of 1-NN classifier.
Estimate Accuracy: Right Way

Training samples → Build Classifier → Classifier

Testing samples → Apply Classifier → Predictions

Estimate Accuracy → Accuracy

Testing samples are NOT to be used during “Build Classifier”
Cross Validation

- Divide samples into $k$ roughly equal parts
- Each part has similar proportion of samples from different classes
- Use each part to test other parts
What is Feature Selection?
Curse of Dimensionality

- Given a sample space of $p$ dimensions/features
- It is possible that some features are irrelevant
- Irrelevant features can confuse a classifier algorithm (or the human analyst!)
- Need to find ways to separate those dimensions (aka features) that are relevant (aka signals) from those that are irrelevant (aka noise)
Signal Selection (Basic Idea)

- Choose a feature with low intra-class distance
- Choose a feature with high inter-class distance
Signal Selection (e.g., t-statistics)

The t-stats of a signal is defined as

\[ t = \frac{|\mu_1 - \mu_2|}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}} \]

where \( \sigma_i^2 \) is the variance of that signal in class \( i \), \( \mu_i \) is the mean of that signal in class \( i \), and \( n_i \) is the size of class \( i \).

Exercise: Look up other feature selection methods.
Self-fulfilling Oracle

- Construct artificial dataset with 100 samples, each with 100,000 randomly generated features and randomly assigned class labels
- Select 20 features with the best t-statistics (or other methods)
- Evaluate accuracy by cross validation using only the 20 selected features
- The resultant estimated accuracy can be ~90%
- But the true accuracy should be 50%, as the data were derived randomly

Exercise: What went wrong?
Concluding Remarks
What have we learned?

• **Methodology of data mining**
  – Feature generation, feature selection, feature integration

• **Evaluation of classifiers**
  – Accuracy, sensitivity, precision
  – Cross validation

• **Curse of dimensionality**
  – Feature selection concept
  – Self-fulfilling oracle
References


• Trevor Hastie et al., *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, 2001. Chapters 1, 7

• Lance D. Miller et al., Optimal gene expression analysis by microarrays, *Cancer Cell* 2:353--361, 2002

• David Hand et al., *Principles of Data Mining*, MIT Press, 2001

• Jinyan Li et al., Data Mining Techniques for the Practical Bioinformatician, *The Practical Bioinformatician*, Chapter 3, pages 35—70, WSPC, 2004