What is Data Mining?

Jonathan’s blocks
Jessica’s blocks
Whose block is this?

Jonathan’s rules: Blue or Circle
Jessica’s rules: All the rest

The Steps of Data Mining

- 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, ...

Some classifiers / machine learning methods

What is Accuracy?

<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>

Accuracy = \[
\frac{\text{No. of correct predictions}}{\text{No. of predictions}} = \frac{TP + TN}{TP + TN + FP + FN}
\]
Examples (Balanced 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>25</td>
<td>25</td>
<td>25</td>
<td>50%</td>
</tr>
<tr>
<td>B</td>
<td>50</td>
<td>25</td>
<td>25</td>
<td>0</td>
<td>75%</td>
</tr>
<tr>
<td>C</td>
<td>25</td>
<td>50</td>
<td>0</td>
<td>25</td>
<td>75%</td>
</tr>
<tr>
<td>D</td>
<td>37</td>
<td>37</td>
<td>13</td>
<td>13</td>
<td>74%</td>
</tr>
</tbody>
</table>

- Clearly, B, C, D are all better than A
- Is B better than C, D?
- Is C better than B, D?
- Is D better than B, C?

Accuracy may not tell the whole story

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>25</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?
- Is C better than B, D?
- Is D better than B, C?

High accuracy is meaningless if population is unbalanced

What is Sensitivity (aka Recall)?

\[
\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}
\]

Sometimes sensitivity wrt negatives is termed specificity

Exercise: Write down the formula for specificity

What is Precision?

\[
\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}
\]

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>25</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></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>50</td>
<td>0</td>
<td>150</td>
<td>0</td>
<td>25%</td>
<td></td>
<td></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?

Abstract Model of a Classifier

- Given a test sample S
- Compute scores \( p(S), n(S) \)
- Predict S as negative if \( p(S) < t \times n(s) \)
- Predict S as positive if \( p(S) \geq t \times n(s) \)

\( t \) is the decision threshold of the classifier

changing \( t \) affects the recall and precision, and hence accuracy, of the classifier
Recall that...
- Predict $S$ as negative if $p(S) < t \times n(s)$
- Predict $S$ as positive if $p(S) \geq t \times n(s)$

Precision-Recall Trade-off
- A predicts better than B if A has better recall and precision than B
- There is a trade-off between recall and precision
- In some apps, once you reach satisfactory precision, you optimize for recall
- In some apps, once you reach satisfactory recall, you optimize for precision

Exercise: Why is there a trade-off between recall and precision?

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 ....

F-Measure (Used in Info Extraction)
- Take the harmonic mean of recall and precision
  $$F = \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}}$$

Does not accord with intuition: C predicts everything as +ve, but still rated better than A

Adjusted Accuracy
- Weigh by the importance of the classes
  $$\text{Adjusted accuracy} = \alpha \times \text{Sensitivity} + \beta \times \text{Specificity}$$
  where $\alpha + \beta = 1$
  typically, $\alpha = \beta = 0.5$

But people can't always agree on values for $\alpha$, $\beta$
What is Cross Validation?

Construction of a Classifier

Estimate Accuracy: Wrong Way

Recall ...

...the abstract model of a classifier

- Given a test sample S
- Compute scores p(S), n(S)
- Predict S as negative if p(S) < t * n(S)
- Predict S as positive if p(S) ≥ t * n(S)

\[ t \] is the decision threshold of the classifier

K-Nearest Neighbour Classifier (k-NN)

- Given a sample S, find the k observations S_i in the known data that are “closest” to it, and average their responses
- Assume S is well approximated by its neighbours

\[
p(S) = \sum_{S_i \in N_k(S) \cap D^P} 1 \quad n(S) = \sum_{S_i \in N_k(S) \cap D^N}
\]

where \( N_k(S) \) is the neighbourhood of S defined by the k nearest samples to it.

Assume distance between samples is Euclidean distance for now

Illustration of kNN (k=8)
For sure k-NN (k = 1) has 100% accuracy in the “accuracy estimation” procedure above. But does this accuracy generalize to new test instances?

Exercise: Why does 1-NN has 100% accuracy under this scenario?

No fixed ratio between training and testing samples; but typically 2:1 ratio
Proportion of instances of different classes in testing samples should be similar to proportion in training samples
What if there are insufficient samples to reserve 1/3 for testing?
Ans: Cross validation

Choose k so that
- the k-fold cross validation accuracy does not change much from k-1 fold
- each part within the k-fold cross validation has similar accuracy
k = 5 or 10 are popular choices for k

Bias affects all measurements the same way, pushing them in the same direction. Chance errors change from measurement to measurement, sometimes up and sometimes down.

<table>
<thead>
<tr>
<th>Size of training set</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small</td>
<td>High</td>
</tr>
<tr>
<td>Medium</td>
<td>Medium</td>
</tr>
<tr>
<td>Large</td>
<td>Low</td>
</tr>
</tbody>
</table>
Bias-Variance Decomposition

- Suppose classifiers $C_j$ and $C_k$ were trained on different sets $S_j$ and $S_k$ of 1000 samples each.
- Then $C_j$ and $C_k$ might have different accuracy.
- What is the expected accuracy of a classifier $C$ trained this way?

Let $Y = f(X)$ be what $C$ is trying to predict.

The expected squared error at a test instance $x$, averaging over all such training samples, is

$$E[(C(x) - Y)^2] = E[(C(x) - E[C(x)])^2 + (E[C(x)] - f(x))^2 - 2E(C(x) - E[C(x)])(E[C(x)] - f(x))].$$

$E[C(x)] - f(x)$

$E[C(x)] - f(x)$

Variance:

how much our estimate $C(x)$ will vary across the different training sets.

Bias:

how far is our average prediction $E[C(x)]$ from the truth.

Proof of Bias-Variance Decomposition

- $E[(C(x) - f(x))^2] = E[(C(x) - E[C(x)])^2 + (E[C(x)] - f(x))^2 - 2E(C(x) - E[C(x)])(E[C(x)] - f(x))].$

Bias-Variance Trade-Off

- In k-fold cross validation,
  - small $k$ tends to underestimate accuracy (i.e., large bias downwards).
  - Large $k$ has smaller bias, but can have high variance.

Curse of Dimensionality

- How much of each dimension is needed to cover a proportion $r$ of total sample space?
- Calculate by $e_d(r) = r^{1/d}$.
- So, to cover 10% of a 15-D space, need 85% of each dimension.

Exercise: Why $e_d(r) = r^{1/d}$.

Recall kNN ...
Consequence of the Curse

• Suppose the number of samples given to us in the total sample space is fixed

• Let the dimension increase

• Then the distance of the k nearest neighbours of any point increases

• Then the k nearest neighbours are less and less useful for prediction, and can confuse the k-NN classifier

What is Feature Selection?

Tackling the Curse

• Given a sample space of p dimensions

• It is possible that some dimensions are irrelevant

• 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

Exercise: Name 2 well-known signal selection statistics

Signal Selection (e.g., t-statistics)

The t-stat of a signal is defined as

$t = \frac{|\mu_i - \mu_j|}{\sqrt{(\sigma_i^2/n_i) + (\sigma_j^2/n_j)}}$

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.

Self-fulfilling Oracle

• Construct artificial dataset with 100 samples, each with 100,000 randomly generated features and randomly assigned class labels

• Evaluate accuracy by cross validation using the 20 selected features

• The resulting accuracy can be ~90%

• Select 20 features with the best t-statistics (or other methods)

• But the true accuracy should be 50%, as the data were derived randomly
What Went Wrong?

- The 20 features were selected from whole dataset
- Information in the held-out testing samples has thus been “leaked” to the training process
- The correct way is to re-select the 20 features at each fold; better still, use a totally new set of samples for testing

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

Any Questions?

Acknowledgements

- The first two slides were shown to me 10+ years ago by Tan Ah Hwee

References

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