CSC 589: ROC Analysis

(Based on ROC Graphs: Notes and Practical Considerations for Data Mining Researchers by Tom Fawcett, January 2003.

Common Evaluation Measures
1. Confusion Matrix

True Class **Hypothesized Class**	**Positive**	**Negative**
Yes	True Positives (TP)	False Positives (FP)
No	False Negatives (FN)	True Negatives (TN)
Column Totals	P	N

Common Evaluation Measures
2. Accuracy, Precision, Recall, etc…

- FP Rate = FP/N (False Alarm Rate)
- Precision = TP/(TP+FP)
- Accuracy = (TP+TN)/(P+N)
- TP Rate = TP/P = Recall = Hit Rate = Sensitivity
- F-Score = Precision * Recall (though a number of other formulas are also acceptable)

Common Evaluation Measures
3. Problem with these Measures

- They describe the state of affairs at a fixed point in a larger evaluation space.
- We could get a better grasp on the performance of our learning system if we could judge its behaviour at more than a single point in that space.
- For that, we should consider the ROC Space

What does it mean to consider a larger evaluation space? (1)

- Often, classifiers (e.g., Decision Trees, Rule Learning systems) only issue decisions: true or false.
- → There is no evaluation space to speak of: we can only judge the value of the classifier's decision.
- WRONG!!! Inside these classifiers, there is a continuous measure that gets pitted against a threshold in order for a decision to be made.
- If we could get to that inner process, we could estimate the behaviour of our system in a larger space

What does it mean to consider a larger evaluation space? (2)

- But why do we care about the inner process?
- Well, the classifier's decision relies on two separate processes. 1) The modeling of the data distribution; 2) The decision based on that modeling. Let's take the case where (1) is done very reliably, but (2) is badly done. In that case, we would end up with a bad classifier even though the most difficult part of the job (1) was well done.
- It is useful to separate (1) from (2) so that (1), the most critical part of the process, can be estimated reliably. As well, if necessary, (2) can easily be modified and improved.

A Concrete Look at the issue: The Neural Network Case (1)

- In a Multiple Layered Perceptron (MLP), we expect the output unit to issue 1 if the example is positive and 0, otherwise.
- However, in practice, this is not what happens. The MLP issues a number between 0 and 1 which the user interprets to be 0 or 1.
- Usually, this is done by setting a threshold at .5 so that everything above .5 is positive and everything below .5 is negative.
- However, this may be a bad threshold. Perhaps we would be better off considering a .75 threshold or a .25 one.

A Concrete Look at the issue: The Neural Network Case (2)

- Please, note that a .75 threshold would amount to decreasing the number of false positives at the expense of false negatives. Conversely, a threshold of .25 would amount to decreasing the number of false negative, this time at the expense of false positives.
- ROC Spaces allow us to explore such thresholds on a continuous basis.
- They provide us with two advantages: 1) They can tell us what the best spot for our threshold is (given where our priority is in terms of sensitivity to one type over the other type of error) and 2) They allow us to see graphically the behaviour of our system over the whole range of possible tradeoffs.

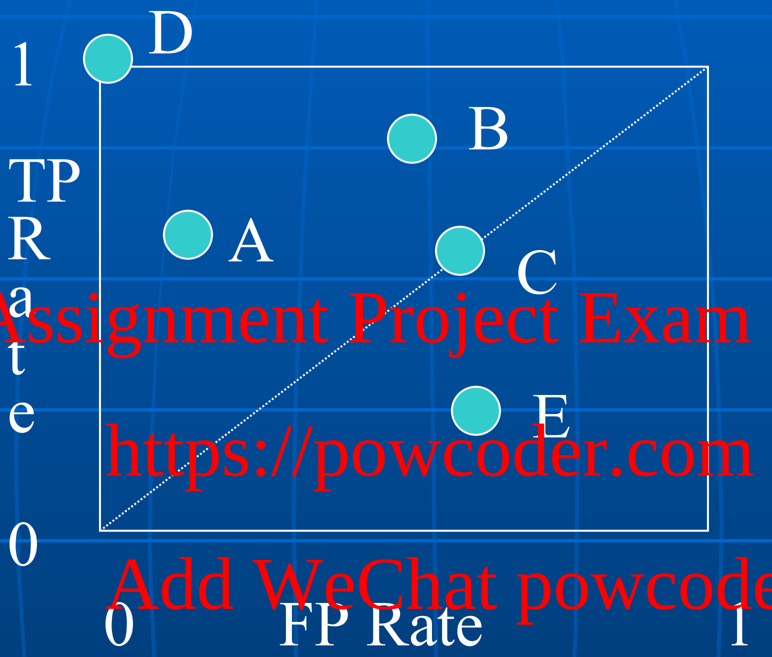
A Concrete Look at the Issue: Decision Trees

- Unlike an MLP, a Decision Tree only returns a class label. However, we can ask how this label was computed internally.
- It was computed by considering the proportion of instances of both classes at the leaf node the example fell in. The decision simply corresponds to the most prevalent class.
- Rule learners use similar statistics on rule confidences and the confidence of a rule matching an instance.
- There does, however, exist systems whose process cannot be translated into a score. For these systems, a score can be generated from an aggregation process. ➔ But is that what we really want to do???

ROC Analysis

- Now that we see how we can get a score rather than a decision from various classifiers, let's look at ROC Analysis per se.
- ***Definition:*** ROC Graphs are two-dimensional graphs in which the TP Rate is plotted on the Y Axis and the FP Rate is plotted on the X Axis. A ROC graph depicts relative tradeoffs between benefits (true positives) and costs (false positives)

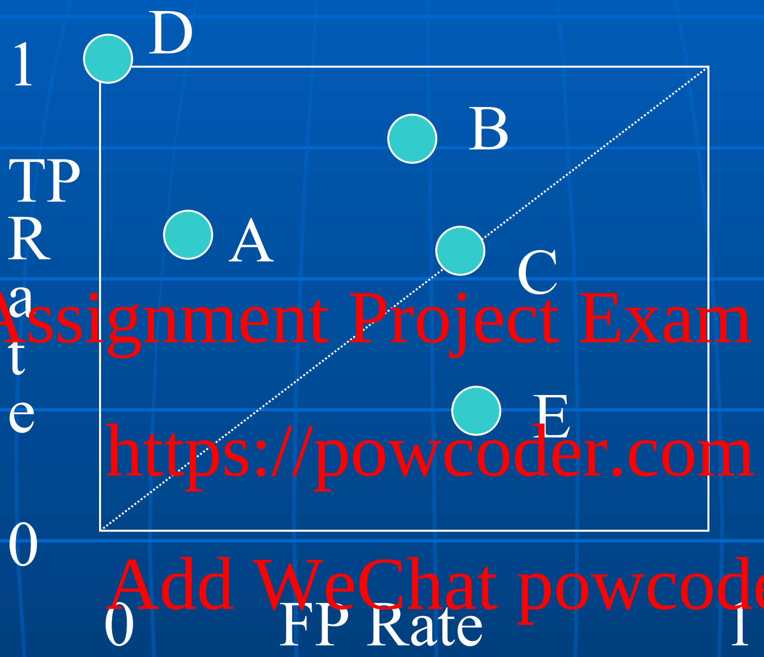
Points in a ROC Graph (1)



Interesting Points:

(0,0): Classifier that never issues a positive classification
➔ No false positive errors but no true positives results
(1,1): Classifier that never issues a negative classification
➔ No false negative errors but no true negative results
(0,1), D: Perfect classification

Points in a ROC Graph (2)



Interesting Points:

Informally, one point is better than another if it is to the Northwest of the first.

Classifiers appearing on the left handside can be thought of as more conservative. Those on the right handside are more liberal in their classification of positive examples.

The diagonal y=x corresponds to random guessing

ROC Curves (1)

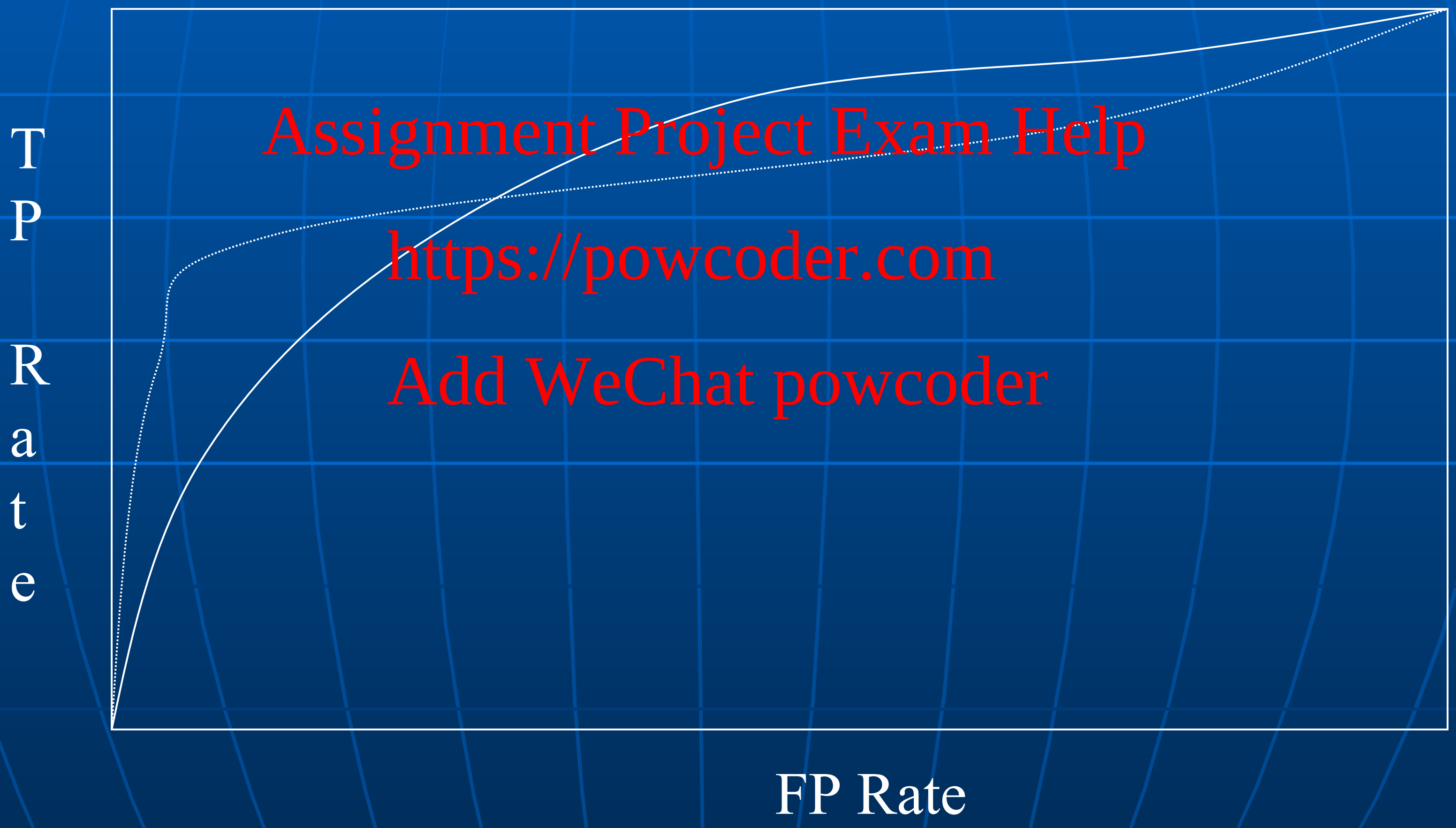
- If a classifier issues a discrete outcome, then this corresponds to a point in ROC space. If it issue a ranking or a score, then, as discussed previously, it needs a threshold to issue a discrete classification.
- In the continuous case, the threshold can be placed at various points. As a matter of fact, it can be slided from -∞ to +∞, producing different points in ROC space, which can connect to trace a curve.
- This can be done as in the algorithm described on the next slide.

ROC Curves (2)

- L= Set of test instances, f(i)= continuous outcome of classifier, min and max:= smallest and largest values returned by f, increment=the smallest difference between any two f values
- for t=min to max by increment do

```
    • FP ← 0
    • TP ← 0
    • for i ∈ L do
        ▪ if f(i)  t then
            • if i is a positive example then
                ▪ TP ← TP + 1
            • else
                ▪ FP ← FP + 1
    • Add point (FP/N, TP/P) to ROC Curve
```

An exemple of two curves in ROC space



ROC Curves: A Few remarks (2)

Inst Numb	1	2	3	4	5	6	7	8	9	10
True	p	p	p	p	p	p	n	n	n	n
Predict	y	y	y	y	y	y	y	y	n	n
Score	.99	.98	.98	.97	.95	.94	.65	.51	.48	.44

If the threshold is set at .5, the classifier will make two Errors. Yet, if it is set at .7, it will make none. This can Clearly be seen on a ROC graph.

ROC Curves: Useful Property

- ROC Graphs are insensitive to changes in class distribution. I.e., if the proportion of positive to negative instances changes in a test set, the ROC Curve will not change.
- That's because the TP Rate is calculated using only statistics about the positive class while the FP Rate is calculated using only statistics from the negative class. The two are never mixed.
- This is important in domains where the distribution of the data changes from, say, month to month or place to place (e.g., fraud detection).

Modification to the Algorithm for Creating ROC Curves

- The algorithm on slide 16 is inefficient because it slides to the next point by a constant fixed factor. We can make the algorithm more efficient by looking at the outcome of the classifier and processing it dynamically.
- As well, we can, compute some averages in various segments of the curve or remove all concavities in a ROC Curves (Please see section 5 of Fawcett's paper for a discussion of all these issues).

Area under a ROC Curve (AUC)

- The AUC is a good way to get a score for the general performance of a classifier and to compare it to that of another classifier.
- There are two statistical properties of the AUC:
  - The AUC of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance
  - The AUC is closely related to the GINI Index (used in CART): GINI + 1 = 2 * AUC
- Note that for specific issues about the performance of a classifier, the AUC is not sufficient and the ROC Curve should be looked at, but generally, the AUC is a reliable measure.

Averaging ROC Curves

- A single ROC Curve is not sufficient to make conclusions about a classifier since it corresponds to only a single trial and ignores the second question we asked on Slide 2.
- In order to avoid this problem, we need to average several ROC Curves. There are two averaging techniques:
  - Vertical Averaging
  - Threshold Averaging

Additional Topics (Section 8 of Fawcett's paper)

- The ROC Convex Hull
- Decision Problems with more than 2 classes
- Combining Classifiers
- Alternative to ROC Graphs:
  - DET Curves
  - Cost Curves
  - LC Index