Class Imbalance & Multiclass

Mike Bowles, PhD
Patricia Hoffman, PhD

The Class Imbalance Problem (Sec. 5.7, p. 204)

- So far we have treated the two classes equally. We have assumed the same loss for both types of misclassification, used 50% as the cutoff and always assigned the label of the majority class.

- This is appropriate if the following three conditions are met

1) We suffer the same cost for both types of errors

2) We are interested in the probability of 0.5 only

3) The ratio of the two classes in our training data will match that in the population to which we will apply the model

The Class Imbalance Problem (Sec. 5.7, p. 204)

- If any one of these three conditions is not true, it may be desirable to "turn up" or "turn down" the number of observations being classified as the positive class.

- This can be done in a number of ways depending on the classifier.

- Methods for doing this include choosing a probability different from 0.5, using a threshold on some continuous confidence output or under/over-sampling.

Recall and Precision (page 297)

	PREDICTED CLASS		
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

Recall = $\frac{a}{a+b} = \frac{TP}{TP+FN}$ **=** $\frac{Correctly\ Predicted\ Positives}{Actual\ Positives}$ **(Sensitivity)**

Precision = $\frac{a}{a+c} = \frac{TP}{TP+FP}$ **=** $\frac{Correctly\ Predicted\ Positives}{All\ Predicted\ Positives}$ **(Precision)**

Before we just used accuracy = $\frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN}$

The *F* Measure (page 297)

- ***F*** **combines recall and precision into one number**

- ***F*** **=** $\frac{2rp}{r+p} = \frac{2TP}{2TP+FP+FN}$

- **It equals the harmonic mean of recall and precision**

$$\frac{2rp}{r+p} = \frac{2}{1/r+1/p}$$

- **Your book calls it the** F_1 **measure because it weights both recall and precision equally**

- **See http://en.wikipedia.org/wiki/Information_retrieval**

The ROC Curve (Sec 5.7.2, p. 298)

- **ROC stands for Receiver Operating Characteristic**

- **Since we can "turn up" or "turn down" the number of observations being classified as the positive class, we can have many different values of true positive rate (TPR) and false positive rate (FPR) for the same classifier.**

TPR= $\frac{TP}{TP+FN}$ **FPR=** $\frac{FP}{FP+TN}$

- **The ROC curve plots TPR on the y-axis and FPR on the x-axis**

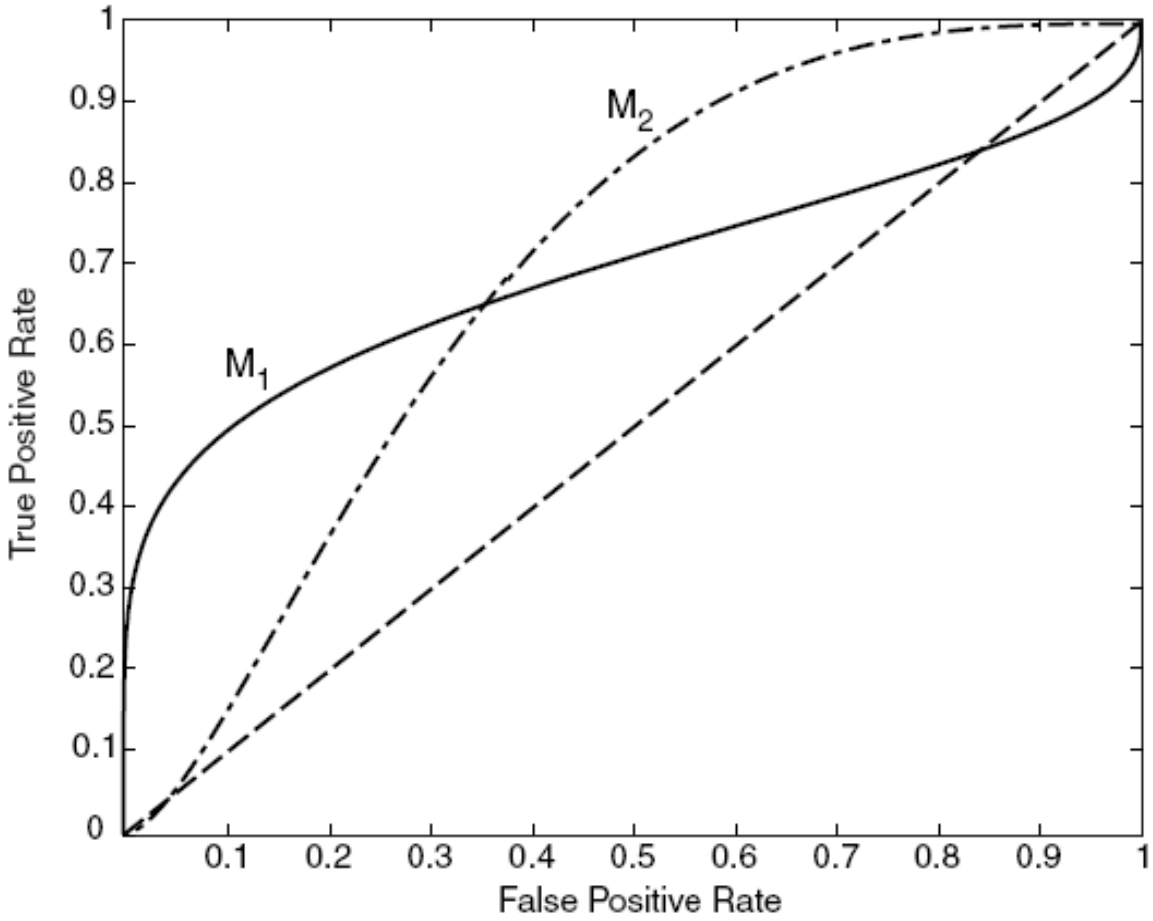

Figure 5.41. ROC curves for two different classifiers.

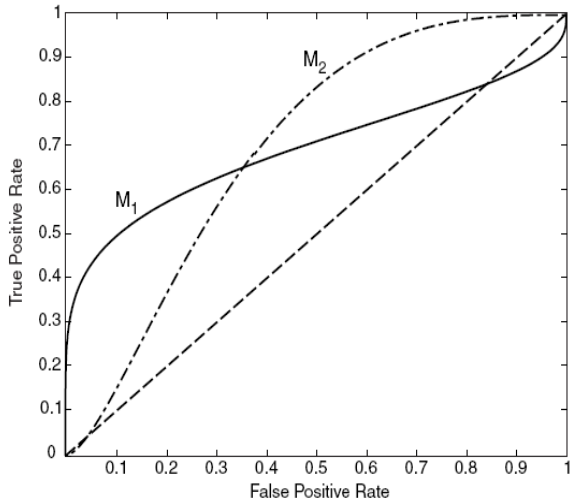

Figure 5.41. ROC curves for two different classifiers.

$$TPR = TP/(TP+FN)$$
$$FPR = FP/(TN+FP)$$

(TPR = 1, FPR = 0) The ideal model

(TPR = 0, FPR = 0)
Model Predicts every instance to be a negative class

(TPR = 1, FPR = 1)
Model predicts every instance to be a positive class

Dotted Line = Random Guesses

The ROC Curve (Sec 5.7.2, p. 298)

- The ROC curve plots TPR on the y-axis and FPR on the x-axis
- The diagonal represents random guessing
- A good classifier lies near the upper left
- ROC curves are useful for comparing 2 classifiers
- The better classifier will lie on top more often
- The Area Under the Curve (AUC) is often used a metric

This is textbook question #17 part (a) on page 322. It is part of your homework so we will not do all of it in class. We will just do the curve for M_1.

You are asked to evaluate the performance of two classification models, M_1 and M_2. The test set you have chosen contains 26 binary attributes, labeled as A through Z.

Table 5.5 shows the posterior probabilities obtained by applying the models to the test set. (Only the posterior probabilities for the positive class are shown). As this is a two-class problem, $P(-) = 1 - P(+)$ and $P(-|A, \ldots, Z) = 1 - P(+|A, \ldots, Z)$. Assume that we are mostly interested in detecting instances from the positive class.

Instance	True Class	$P(+\|A, \ldots, Z, M_1)$	$P(+\|A, \ldots, Z, M_2)$
1	+	0.73	0.61
2	+	0.69	0.03
3	−	0.44	0.68
4	−	0.55	0.31
5	+	0.67	0.45
6	+	0.47	0.09
7	−	0.08	0.38
8	−	0.15	0.05
9	+	0.45	0.01
10	−	0.35	0.04

(a) Plot the ROC curve for both M_1 and M_2. (You should plot them on the same graph.) Which model do you think is better? Explain your reasons.

In class exercise #41:

This is textbook question #17 part (b) on page 322.

(b) For model M_1, suppose you choose the cutoff threshold to be t = 0.5. In other words, any test instance whose posterior probability is greater than t will be classified as a positive example. Compute the precision, recall, and F-measure for the model at this threshold value.

(c) Repeat the analysis from (b) using the same cutoff threshold on model M_2. Compare the F-measure results for both models. Which model is better? Are the results consistent with what you expect from the ROC curve?

(d) Repeat part (c) for model M_1 using the threshold t=0.1 Which threshold do you prefer, t=0.5 or t=0.1? Are the results consistent with what you expect from the ROC curve?

Multiclass Problem

The Target Y = {y1, y2, … , yk} has multiple values – Iris Data is Example

- One-Against-Rest (1-r)
  - K binary classifiers, one for each yi in Y
  - yi is the positive example – rest are negative examples

- One-Against-One (1-1)
  - K(K-1)/2 binary classifiers
  - each classifier distinguishes between a pair of classes (yi, yj)
  - instances that do not belong to either yi or yj are ignored

Test Instances Classified
  - combine predictions
  - from binary classifiers
  - voting scheme or probability estimate

Error Correcting Output Coding - ECOC

- Robust Method for multiclass problem

- For each class, yi
  - unique bit string of length n known as codeword
  - result of n binary classifiers predicts each bit of codeword

- Hamming Distance between two codewords
  - number of bits that differ

- Predicted class of test instance
  - codeword closest Hamming Distance

- If minimum Hamming distance
  - between pair of codewords is d
  - any (d-1)/2 errors corrected using nearest codeword