ECON 950 — Winter 2020
Prof. James MacKinnon

10. Performance of Classification Methods

For regression methods, it is easy to evaluate performance based on average squared error or average absolute error for the test data. Of course, this only makes sense if the disturbances are homoskedastic.

For classification methods, it is not so obvious how to evaluate performance.

One possibility is to use the **deviance** for the test data, which is essentially the average fit according to the loglikelihood function.

Suppose the test dataset contains M observations. If there are K outcomes, and the outcome for observation i is G_i, then the average deviance is

$$\frac{-2}{M}\sum_{i=1}^{M}\sum_{k=1}^{K}\mathbb{I}(G_i=k)\log\hat{p}_k(\boldsymbol{x}_i), \tag{1}$$

where $\hat{p}_k(\boldsymbol{x}_i)$ is the probability of outcome k for observation i.

In the binary case, if we used a logit model (with or without regularization), the probability is either

$$\frac{\exp(\boldsymbol{X}_i\hat{\boldsymbol{\beta}})}{1+\exp(\boldsymbol{X}_i\hat{\boldsymbol{\beta}})} \quad \text{or} \quad \frac{1}{1+\exp(\boldsymbol{X}_i\hat{\boldsymbol{\beta}})}\,. \tag{2}$$

Thus the average deviance for the test data is

$$\frac{1}{M}\sum_{i=1}^{M} -\log\bigl(1+\exp \boldsymbol{X}_i\hat{\boldsymbol{\beta}}) + \frac{1}{M}\sum_{y_i=1} \boldsymbol{X}_i\hat{\boldsymbol{\beta}}. \tag{3}$$

Many classification methods do not give probabilities, however. They just classify each observation in the test set.

Moreover, if we care about mistakes in classification, a criterion based on probabilities may not be what we want.

Consider the binary case. Let 1 denote “a positive” and 0 denote “a negative.” For example, having a disease might be a 1, and not having it might be a 0.

We may care much more about false negatives than false positives, so a criterion like deviance that weights them equally may not be appropriate.

Four possible outcomes are:

- True positive (TP)
- True negative (TN)
- False positive (FP)
- False negative (FN)

Then the **true positive rate**, or **sensitivity**, is

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}\,, \tag{4}$$

and the **true negative rate**, or **specificity**, is

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}\,. \tag{5}$$

In each case, the rate is the ratio of true positives (or negatives) to the total number of positives (or negatives) in the test set.

There are many other ratios that we could calculate, such as the **false positive rate**, or FPR, and **false negative rate**, or FNR. These satisfy

$$\mathrm{FPR} = 1 - \mathrm{TNR} \quad \text{and} \quad \mathrm{FNR} = 1 - \mathrm{TPR}. \tag{6}$$

Also of interest is the **accuracy**, or **ACC**, given by

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}}\,. \tag{7}$$

However, accuracy may be misleading, because it does not distinguish between the two types of error.

In the context of hypothesis testing, if we think of the null hypothesis as a negative, a Type I error is equivalent to a false positive, and a Type II error is equivalent to a false negative. Thus the FPR is like the size of a test, and the FNR is like one minus the power.

The four outcomes of interest (TP, FP, TN, and FN) can be organized into what is sometimes called a **confusion matrix**.

This is a 2×2 matrix with the actual outcome on the horizontal axis and the predicted outcome on the vertical axis.

Table 1. A Confusion Matrix

Prediction / Outcome	Positive	Negative
Positive	TP	FP
Negative	FN	TN

Ideally, this matrix would have non-zero elements only on the diagonal.

If we add the column and row totals, we can turn the confusion matrix into a contingency table.

Here PP and PN mean "predicted positive" and "predicted negative," and AN and AP mean "actual positive" and "actual negative."

TOT is the total number of test observations, so that

$$\text{TOT} = \text{PP} + \text{PN} = \text{AP} + \text{AN}. \tag{8}$$

Table 2. A Contingency Table

Prediction / Outcome	Positive	Negative	Total
Positive	TP	FP	PP
Negative	FN	TN	PN
Total	AP	AN	TOT

For a contingency table like this, it would be natural to test the hypothesis that the prediction method is useless by using a standard χ^2 test. In this case, it would have just one degree of freedom.

The test statistic is simply

$$\frac{(\mathrm{TP}-\mathrm{TP}_{\mathrm{E}})^2}{\mathrm{TP}_{\mathrm{E}}}+\frac{(\mathrm{FP}-\mathrm{FP}_{\mathrm{E}})^2}{\mathrm{FP}_{\mathrm{E}}}+\frac{(\mathrm{TN}-\mathrm{TN}_{\mathrm{E}})^2}{\mathrm{TN}_{\mathrm{E}}}+\frac{(\mathrm{FN}-\mathrm{FN}_{\mathrm{E}})^2}{\mathrm{FN}_{\mathrm{E}}},$$

where the “E” subscripts mean “expected.” These expected quantities are calculated using row and column totals. For example,

$$\mathrm{TP}_{\mathrm{E}}=\frac{\mathrm{PP}\times\mathrm{AP}}{\mathrm{TOT}}. \tag{9}$$

This is the fraction of the observations that are actually positive (AP/TOT) times the number that are predicted to be positive.

If, for example, 40% of the observations are actually positive, and 80 observations are predicted to be positive, we expect there to be 32 true positives and 48 false positives if the model has no predictive ability.

10.1. ROC Curves

An ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) for various classification thresholds.

The idea is that we can adjust the threshold for classifying an observation as positive or negative. This is like adjusting the critical value for a test.

With any sort of binary response model, we get a probability, say P_i, for each observation in the test set.

We can vary the cutoff probability, say P^c, for classifying an observation as positive or negative. We classify as positive when $P_i \geq P^c$.

The usual cutoff is $P^c = 0.5$. But if we set $P^c = 0$, we classify all observations as positive, and if we set $P^c = 1$, we classify them all as negative.

ROC stands for "receiver operating characteristic" because it was originally used in the context of interpreting radar (and sonar) signals.

In the testing context, imagine a test that is asymptotically $\chi^2(1)$ based on a test statistic τ using critical value τ^c.

- If we set τ^c to a very large number, say τ^c_{big}, the test will never reject, whether or not the null is true.
- If we set $\tau^c = 0$, the test will always reject.
- As we reduce τ^c from τ^c_{big} to smaller values, the test will start to reject. It ought to reject more often under the alternative than under the null.
- For any value of τ^c, there is a rejection rate under the null (FPR) and a rejection rate under the alternative (TPR).
- The ROC curve (size-power curve) puts FPR on the horizontal axis and TPR on the vertical axis.

Varying P^c in the context of classification is like varying τ^c in the context of testing.

For the test sample, we can find the FPR and TPR for each value of P^c. If the test sample is large, this should give us a fairly smooth curve.

The curve starts at $(0,0)$ and ends at $(1,1)$. It should always be above the 45° line if TPR $>$ FPR everywhere.

One way to measure the performance of a classifier is to calculate the area under the ROC curve (AUROC, or just AUC).

The maximum value of AUROC is 1. A classifier with no information would have an AUROC of 0.5.

The **Gini coefficient** is twice the area between the ROC curve and the 45° line.

Some parts of the ROC curve may be much more interesting than others.

In the size-power case, we care about power for small test sizes (say, .01 to .10) not for large ones.

In the classification case, we may care about TPR and FPR for values of P^c not too far from 0.5.

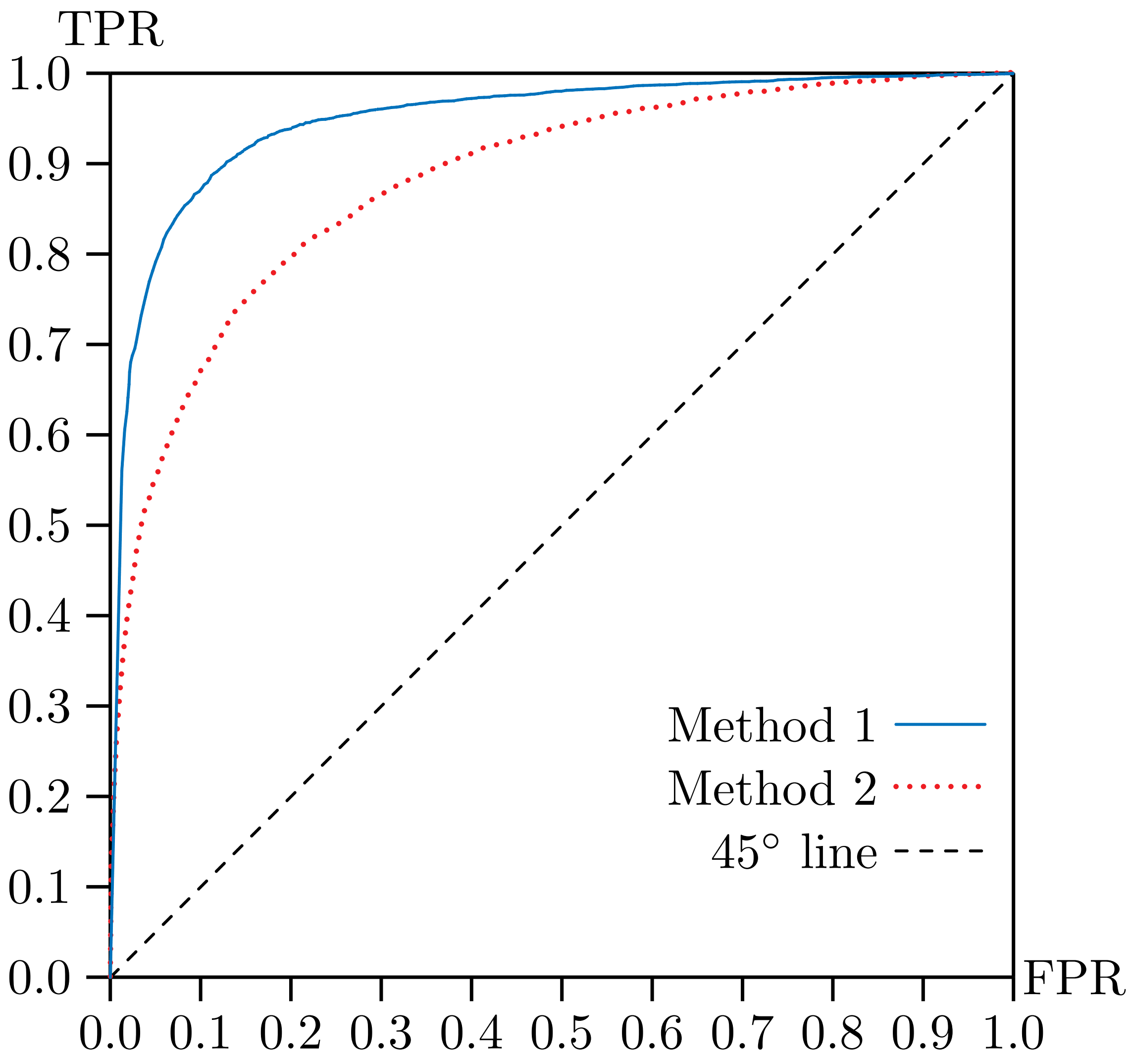

Figure 1. Two ROC Curves

People sometimes plot ROC curves differently, putting the TNR (specificity) on the horizontal axis instead of the FPR.

Since TNR $= 1 -$ FPR, this inverts the horizontal axis. The ideal point is now the upper right-hand corner instead of the upper left-hand corner. See ESL, Figure 9.6.