Model Validation

Sonal Ghanshani

R-Sqd & MSE

$$R^2 \equiv 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}.$$

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$$

Adj R-sqd

$$\text{Adjusted } R^2 = 1 - \frac{\text{RSS}/(n-d-1)}{\text{TSS}/(n-1)}$$

- The adjusted R2 statistic is another popular approach for selecting among a set of models that contain different numbers of variables.
- The usual R2 is defined as 1 − RSS/TSS, where TSS =(yi −y)2 is the total sum of squares for the response. Since, RSS always decreases as more variables are added to the model, the R2 always increases as more variables are added.
- Maximizing the adjusted R2 is equivalent to minimizing RSS/(n−d−1). While RSS always decreases as the number of variables in the model increases, RSS/(n−d−1) may increase or decrease, due to the presence of d in the denominator.
- Unlike Cp, AIC, and BIC, for which asmall value indicates a model with a low test error, a large value of adjusted R2 indicates a model with a small test error. Maximizing the adjusted

Mallow’s Cp

$$C_p = \frac{1}{n}\left(\mathrm{RSS} + 2d\hat{\sigma}^2\right)$$

- where ˆ σ2 is an estimate of the variance of the error associated with each response measurement. Essentially, the Cp statistic adds a penalty of 2dˆσ2 to the training RSS in order to adjust for the fact that the training error tends to underestimate the test error.
- Clearly, the penalty increases as the number of predictors,d, in the model increases; this is intended to adjust for the corresponding decrease in training RSS.
- The Cp statistic tends to take on a small value for models with a low test error, so when determining which of a set of models is best, **we choose the model with the lowest Cp value.**

AIC (Akaike information criterion) & BIC (Bayesian information criterion)

$$\mathrm{AIC} = \frac{1}{n\hat{\sigma}^2}\left(\mathrm{RSS} + 2d\hat{\sigma}^2\right)$$

$$\mathrm{BIC} = \frac{1}{n}\left(\mathrm{RSS} + \log(n)d\hat{\sigma}^2\right)$$

- Like Cp, the BIC will tend to take on a small value for a model with a low test error, and so generally **we select the model that has the lowest BIC value**.
- BIC replaces the 2dˆ σ2 used by Cp with a log(n)dˆ σ2 term, where n is the number of observations. Since log n > 2 for any n > 7, the BIC statistic generally places a heavier penalty on models with many variables, and hence results in the selection of smaller models than Cp.

Confusion Matrix

		Predicted class	
		P	N
Actual Class	P	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

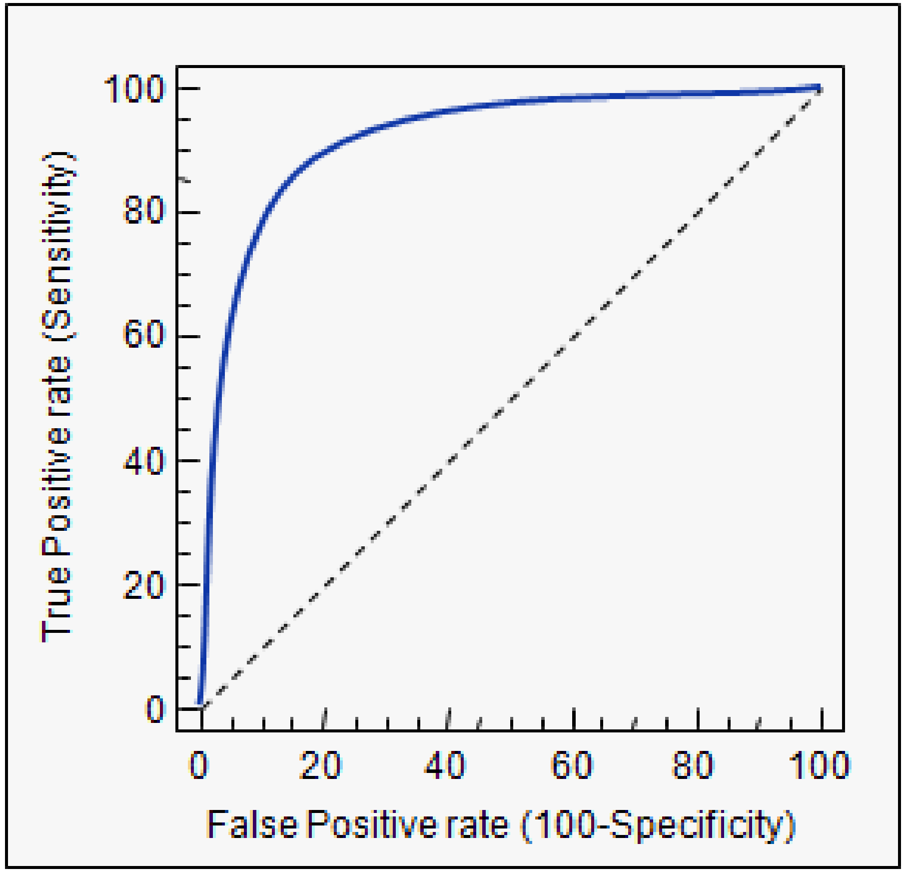
Sensitivity, recall, hit rate, or true positive rate (TPR)

$$\mathrm{TPR} = \frac{\mathrm{TP}}{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

Specificity or true negative rate (TNR)

$$\mathrm{TNR} = \frac{\mathrm{TN}}{N} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

ROC



True Positive rate (Sensitivity)
100
80
60
40
20
0
0
20
40
60
80
100
False Positive rate (100-Specificity)