Logistic Regression with R

1. How to fit a logistic model in R
2. How to interpret the coefficients?
   Increasing X1 by one unit —-> change the odds by exp(beta1)
3. What's null deviance and residual deviance
4. How to do model selection with AIC/BIC
5. How to do model selection with Lasso

Evaluate Classification Accuracy

Confusion Matrix and ROC Curve

		Predicted Class	
		No	Yes
Observed Class	No	TN	FP
	Yes	FN	TP

TN	True Negative
FP	False Positive
FN	False Negative
TP	True Positive

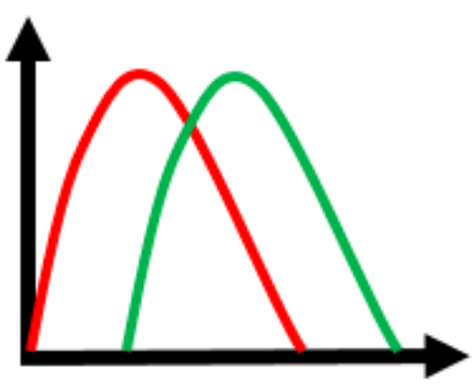
Model Performance

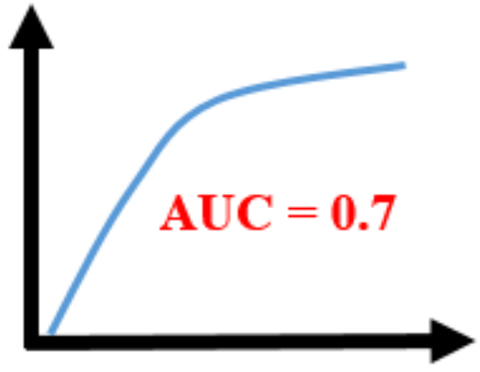
Accuracy	= (TN+TP)/(TN+FP+FI
Precision	= TP/(FP+TP)
Sensitivity	= TP/(TP+FN)
Specificity	= TN/(TN+FP)

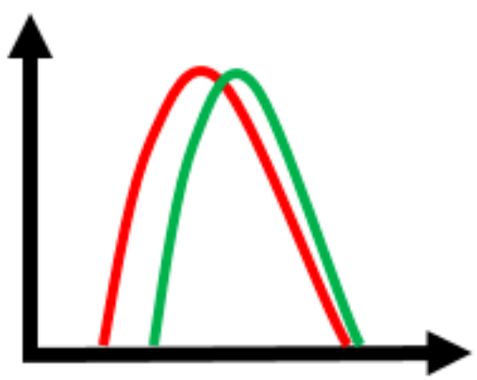
Evaluate Classification Accuracy

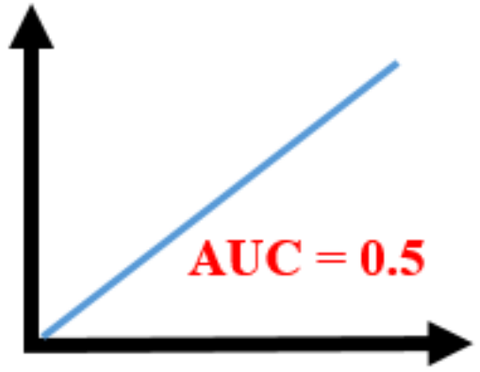
		True condition			
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$
Predicted condition	Predicted condition positive	**True positive**, Power	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$
		True positive rate (TPR), Recall, Sensitivity, probability of detection = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$
		False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$	F_1 score = $\frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$

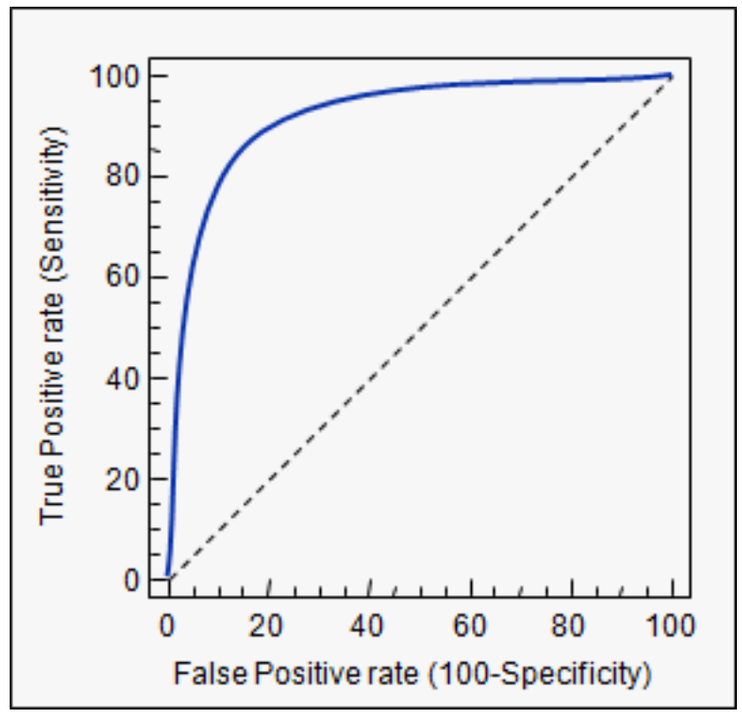
AUC and ROC




AUC = 0.7




AUC = 0.5


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

AUC and Mann-Whitney U Statistic

Mann-Whitney U-stat or Wilcoxon rank sum Stat

1. Assign numeric ranks to all the observations (put the observations from both groups to one set), beginning with 1 for the smallest value. Where there are groups of tied values, assign a rank equal to the midpoint of unadjusted rankings. E.g., the ranks of $(3, 5, 5, 5, 5, 8)$ are $(1, 3.5, 3.5, 3.5, 3.5, 6)$ (the unadjusted rank would be $(1, 2, 3, 4, 5, 6)$).
2. Now, add up the ranks for the observations which came from sample 1. The sum of ranks in sample 2 is now determinate, since the sum of all the ranks equals $N(N+1)/2$ where N is the total number of observations.
3. U is then given by:[4]

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2}$$

AUC = $U_1/(n_1 n_2)$

More on Logistic Regression

- Convergence issue with logistic regression when data are well-separated
- Multinomial logistic regression
- Move beyond linear decision boundary: add quadratic terms to logistic regression
- Retrospect sampling (both LDA and Logistic can handle this)

We usually assume that data (x_i, y_i)'s are collected as iid samples from a population of interest. However, in some applications, we may have data collected retrospectively. For example, in a study on cancer, instead of taking a random sample of 100 people from the whole population (then the number of cancer patients will be too small), we may draw 50 samples from the cancer group and 50 from the control group.

Let Z indicate whether an individual is sampled or not. Using retrospective samples, we are estimating $P(Y = 1|Z = 1, X = x)$, while what we care is $P(Y = 1|X = x)$. Assume the logit of the latter follows a linear model, that is,

$$P(Y = 1|X = x) = \frac{\exp(\alpha + x^t\beta)}{1 + \exp(\alpha + x^t\beta)}, \tag{2}$$

where we isolate the intercept α from other coefficients β. Then the question is whether we can estimate α or β in (2) based on a retrospective sample.

It turns out that for logistic models, the coefficients estimated from a retrospective sample is roughly the same as the one from an ordinary random sample. Again, let Z indicate whether an individual is sampled or not and denote the sample proportions (in each class) by

$$r_0 = \mathbb{P}(Z = 1|Y = 0), \quad r_1 = \mathbb{P}(Z = 1|Y = 1).$$

Then

$$\begin{aligned}
\mathbb{P}(Y = 1|Z = 1, X = x) &= \frac{\mathbb{P}(Z = 1|Y = 1, x)\mathbb{P}(Y = 1|x)}{\mathbb{P}(Z = 1, Y = 1|x) + \mathbb{P}(Z = 1, Y = 0|x)} \\
&= \frac{r_0 \exp(\alpha + x^t\beta)}{r_1 + r_0 \exp(\alpha + x^t\beta)} \\
&= \frac{\exp(\alpha^* + x^t\beta)}{1 + \exp(\alpha^* + x^t\beta)}.
\end{aligned}$$

Although the data have been sampled retrospectively, the logistic model continues to apply with the same coefficients β but a different intercept.