02 Model Diagnostics

Bias-Variance Tradeoff

Predictability measure (error on test set) = variance + bias

Evaluate Goodness of Model

For Linear Regression/ GLM

We can use hypothesis test and model selection criteria to evaluate the goodness of a LR/ GLM model.

Hypothesis Test

- Wald test
- Score test
- Likelihood ratio test

Model Selection Criteria

- AIC
- BIC
- ...

For all Type of Supervised Model

Training Set / Validation Set / Test Set

- **Training set**: used to train s model
- **Validation set**: used to evaluate the performance of trained models, their performance will be used to model selection
- **Test set**: pass the test set to your final model, the resulting performance (eg.accuracy) is the final performance.

Cross-Validation

Randomly split the whole dataset to p equal-sized subset. For each subset, use it as validation set (for model selection) and use other p-1 subset as training set to get a model. We will get p models in total, and p validation set error in total. The final performance is the average of the p validation set error.

Gains Curve

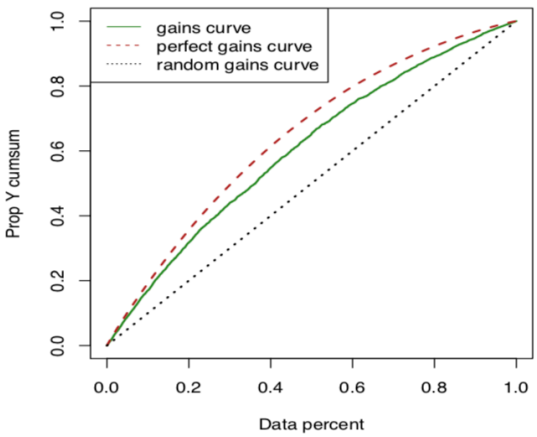
Tools to evaluate models performance in context of regression with **continuous response**.

Gain curve at about is better.

Y轴:$\frac{\sum_{i=0}^{j} Y_i}{\sum_{k=1}^{n} Y_k}$ for $j = 1...n$ in case of $(f(X_i), Y_i)$ pairs are sorted by the value of $f(X_i)$ from high to low

X轴: $\frac{j}{n}$ for $j = 1...n$

- 灵感来自于low predictions should corresponding with low response observations, high predictions should corresponding with high observations.
- 比如现在有两个model $f_1(X)$, $f_2(X)$ and a dataset of observations X_is.
- Sort $(f_1(X_i), Y_i)$ pairs by the value of $f_1(X_i)$ from high to low.
- Look how good the Y_i's are in a sorted sequence? 按照predictions sorting之后, true response是不是也被sorted了? 如果没有, 说明predictions错得厉害.
- Sorting for $f_2(X)$, 看看哪个sorting 过程能把true response Y排序得好.
- 把Y排得好的那个model为佳.



Note:

- 当observation很多时, 很难看出Y sort得好不好, 所以使用 $\frac{cumulative\ sum\ of\ Y}{total\ sum\ of\ Y}$ to represent the proportion of cumulative Y. It is better if the term increase earlier.
- So for Gain curve, **above is better.**
- Gain curve 只关注顺序. 忽略了具体值. 不能保证goodness of fit.

All has perfect gain chart, but..

Y_i	$f_1(X_i)$	$f_2(X_i)$	$f_3(X_i)$	$f_4(X_i)$
1	1	101	2.98	-1,000
2	2	102	2.99	0
3	3	103	3.00	0.1
4	4	104	3.01	50
5	5	105	3.02	50.001

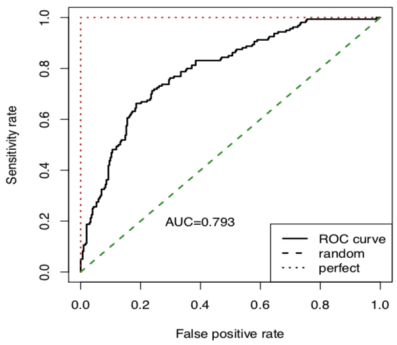
ROC Curve (Receiver Operating Characteristic Curve) & AUC

ROC curve is a tools (which is similar as gain curve) to evaluate models performance in context of **binary response** {0,1}. AUC is the area under the ROC curve

Y轴: Sensitivity rate for variant cutoff

X轴: False Positive rate for variant cutoff

- Sensitivity Rate: proportion of real 1 predicted as 1
- False negative Rate: proportion of real 1 predicted as 0
- Specificity Rate: proportion of real 0 predicted as 0
- False Positive Rate: proportion of real 0 predicted as 1



Note:

左上角的ROC is best, 对应AUC的最大值1.0