PoliSci 4782 Political Analysis II

Model Evaluation

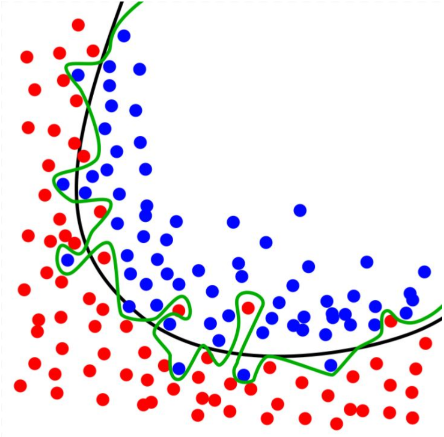
Seamus Wagner

The Ohio State University

What is a Good Model?

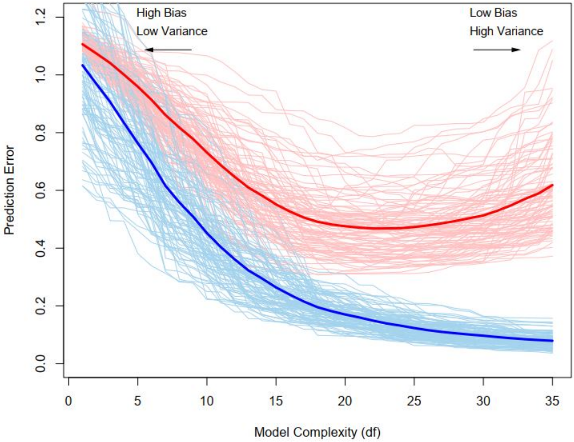
- Within samples, good models go with higher R^2, lower root-mean-square-error, higher log likelihood scores, lower deviance, lower AIC/BIC, etc.
- But how about beyond samples?
- Is a model with the highest likelihood score and lowest AIC/BIC in a sample always the best in the population?
- Good models capture those persistent structures, not idiosyncratic features in samples (which are in fact stochastic patterns in population).

Over-fitting



The black curve accurately captures the fundamental structure, whereas the green one pays unnecessary attention to sample idiosyncracies and therefore over-fits.

In-sample (Blue) vs. Out-of-sample (Red)



Trevor Hastie et al. 2009. *The Elements of Statistical Learning*, pp.220.

The Core Idea

- Out-of-sample prediction is the true test of model performance.
- If we only have a sample, break our sample into a training set and a test set (normally 20-30%) for training.
- Search for good models with the training set, use the chosen model to predict out-of-sample onto the test set, and evaluate model fit.

Cross Validation

- A process of randomly choosing k observations as the test set and doing out-of-sample testing.
- One round may not be enough, so people often do multiple rounds to assess the validity of the model.
- There are different partition schemes to do cross validation:
  - leave-one-out
  - k-fold
  - ...

Leave-one-out Cross Validation

1. Let $(y_k; \boldsymbol{X}_k)$ be the kth observation in the dataset
2. Temporarily remove observation k from the dataset
3. Train on the remaining N-1 observations
4. Predict observation k using the training data and save the error ϵ_k
5. Repeat for all observations in the dataset
6. Report the cross-validation error $\Delta = \frac{n_k}{n} \sum_{k=1}^{K} \frac{(y-\hat{y})^2}{n_k}$ (lower is better)[1]

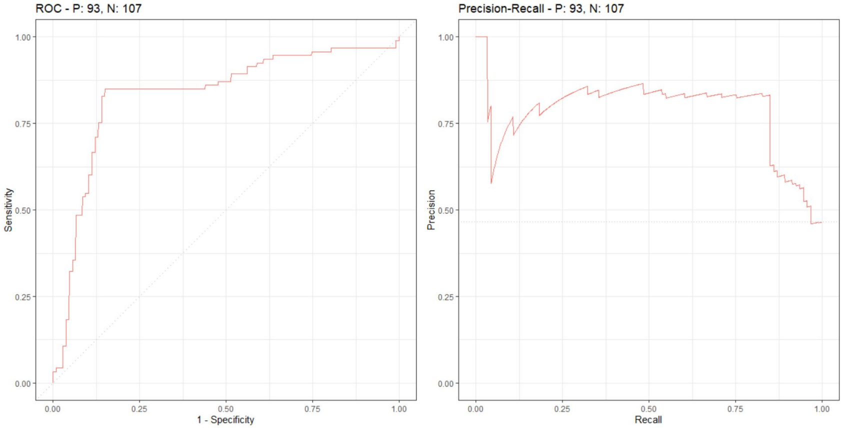
[1] n for no. of observations; n_k for no. of observations in each training set, K for no. of rounds

K-fold Cross Validation

1. Randomly divide the dataset into k partitions
2. For each partition: train on all observations not in this partition, and then test with this partition
3. Repeat for all partitions
4. Report the CV error $\Delta = \frac{n_k}{n} \sum_{k=1}^{K} \frac{(y - \hat{y})^2}{n_k}$ (lower is better)

Visualizing Model Performance

ROC (receiver-operating characteristic) and precision-recall plots (for binary classification):

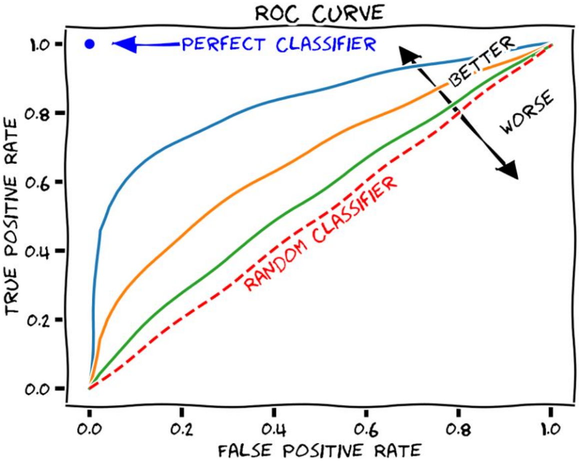


ROC Curves

- The true positive fraction (TPF): no. of cases with $y = \hat{y} = 1$ / no. of cases with $\hat{y} = 1$
- The false positive fraction (FPF): no. of cases with $\hat{y} = 1$ but $y = 0$ / no. of cases with $\hat{y} = 1$
- Sensitivity is equivalent to TPF
- Specificity is equivalent to FPF
- ROC plots specificity (horizontal) against sensitivity (vertical)
- The closer to the top-left, the better

ROC Curves

The blue dominates the yellow and the green

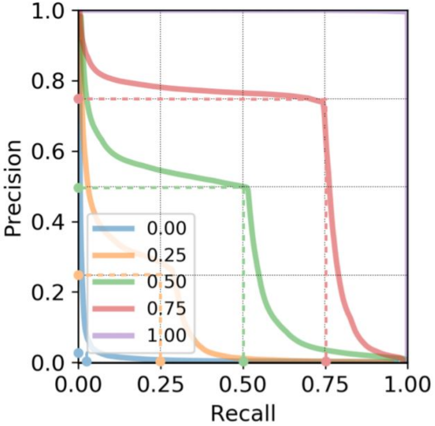


Precision-recall Curves

- True positive (TP): no. of cases with $y = \hat{y} = 1$
- False negative (FN): no. of cases with $\hat{y} = 0$ but actually $y = 1$
- Precision is equivalent to TPF
- Recall is the fraction of TPs over the sum of TPs and FNs
- Precision-recall plots recall (horizontal) against precision (vertical)
- The closer to the top-right, the better

PR Curves

The red dominates the green, the yellow, and the blue.



Simon L, Webster R, Rabin J. (2019)

Problem with Standard Error Estimation

Think about SE

In math, all the standard errors of a model are derived from a variance-covariance matrix ($[\boldsymbol{X'X}]^{-1}$) with n equal to the number of data points:

$$\Sigma = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \cdots & \sigma_{1n}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 & \cdots & \sigma_{2n}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1}^2 & \sigma_{n2}^2 & \cdots & \sigma_{nn}^2 \end{pmatrix}$$

Independence and Homoskedasticity

The ideal scenario when the model is accurately specified and the model errors (σ^2) are the same across observations and no additional correlations conditional on all the explanatory variables, which means $\Sigma = \sigma^2[\boldsymbol{X'X}]^{-1}$:

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}$$

Independence and Heteroskedasticity

Now the model has a mis-specification problem such that regression errors (σ^2) vary across observations:

$$\Sigma = \begin{pmatrix} \sigma_{11}^2 & 0 & \dots & 0 \\ 0 & \sigma_{22}^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_{nn}^2 \end{pmatrix}$$

Autocorrelation and Heteroskedasticity

Now the problem is even worse, not just with non-constant errors but also having correlations with each other:

$$\Sigma = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \dots & \sigma_{13}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 & \dots & \sigma_{23}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1}^2 & \sigma_{n2}^2 & \dots & \sigma_{nn}^2 \end{pmatrix}$$

Robust Standard Error

- When heteroskedasticity is present, $Var(\beta) \neq \sigma^2[\boldsymbol{X'X}]^{-1}$
- Instead, Huber, Eicker, and White developed a solution by replacing σ^2 by regression errors, which is called robust (or sandwich) standard error:
  - $Var(\hat{\beta}) = [\boldsymbol{X'X}]^{-1}[\sum_{i=1}^{N} \hat{\epsilon_i^2} XX'][\boldsymbol{X'X}]^{-1}$
- But robust SE works only when the model is somewhat accurately specified; otherwise, $\hat{\epsilon_i^2}$ is far from the truth.
- So instead using robust SE is a panacea, better use it as a test on model specification.
- When original SEs and robust SEs lead to contradicting conclusions, our model is not right.

Coming Up

- Lab on cross validation and robust standard error computation
- Theories and techniques of dealing with missing data in the next week