THE UNIVERSITY of EDINBURGH | usher institute

Model evaluation

Steven Kerr

Supported by

THE UNIVERSITY of EDINBURGH | DDI Data-Driven Innovation

Model evaluation

- It is important to have a measure of how good our model is.
- The complete set of information about the model's performance in a dataset is contained in the full set of actual values and predicted values.
- However, that is usually too much information.
- Instead, we use metrics that summarise performance.
- Let y_i be actual values, and $\hat{y}_i$ predictions.

Mean errors

- Mean squared error:

$$\frac{1}{n}\Sigma_i(y_i - \hat{y}_i)^2$$

- Root mean squared error:

$$\sqrt{\frac{1}{n}\Sigma_i(y_i - \hat{y}_i)^2}$$

- Mean absolute error:

$$\frac{1}{n}\Sigma_{\mathrm{i}}|y_i - \hat{y}_i|$$

R^2 - coefficient of determination

- Let $\overline{y} = \frac{1}{n}\Sigma_i\, y_i$.
- Residual sum of squares:

$$RSS := \Sigma_i (y_i - \hat{y}_i)^2$$

- Explained sum of squares:

$$ESS := \Sigma_i (\hat{y}_i - \overline{y})^2$$

- Total sum of squares:

$$TSS := \Sigma_i (y_i - \overline{y})^2$$

R^2 - coefficient of determination

- In an OLS model $y_i = \beta \cdot x_i + \epsilon_i$, it is possible to show that

$$TSS = ESS + RSS$$

- This motivates the following definition:

$$R^2 := \frac{ESS}{TSS}$$

- R^2 is the 'proportion of total variation in the dependent variable that is explained by the model'.

- For nonlinear models, TSS is in general not equal to $ESS + RSS$. The interpretation breaks down, and this R^2 is not suitable.

Confusion matrix

- Consider a binary classification problem.

Actual \ Predicted	1	0
1	True positive (TP)	False negative (FN)
0	False positive (FP)	True negative (TN)

- This is called a confusion matrix.

- There are many quantities that can be derived from the confusion matrix.

Confusion matrix

- True positive rate:

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

The sample probability that a positive is correctly identified as such.

- False positive rate:

$$FPR = \frac{FP}{FP + TN}$$

The sample probability that a negative is incorrectly predicted to be positive.

Confusion matrix

- True negative rate:

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

The sample probability that a negative is correctly identified as such.

- False negative rate:

$$FNR = \frac{FN}{TP + FN}$$

The sample probability that a positive is incorrectly predicted to be negative.

Confusion matrix

- Just to confuse you even more:
  - TPR is also called sensitivity or recall.
  - TNR is also called specificity or selectivity.

- We prefer to use the terminology of true/false positive/negative.

Confusion matrix

- Confusion matrices can be extended to the multi-class case.

Actual \ Predicted	Apple	Banana	Orange
Apple	10	1	2
Banana	0	12	1
Orange	3	0	8

- TPR, FPR, etc can be calculated for each class. Just turn it into a binary confusion matrix by bundling together all the other categories.

Actual \ Predicted	Apple	Other
Apple	10	3
Other	3	21

C statistic

- Consider again a binary classification problem.
- The C statistic is the sample probability that a randomly chosen positive will have a higher predicted probability of event than a randomly chosen negative.

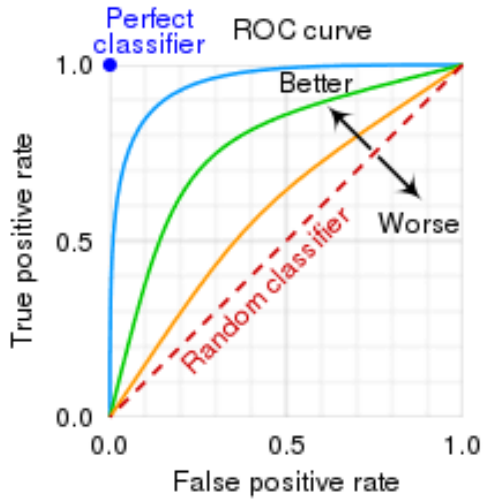
Receiver operating characteristic (ROC) curve

- We can implement a threshold value for binary classification. For example,
  - If $p_i \geq 0.5$, classify as 1.
  - Else, classify as 0.

- The threshold value can be varied, and we can calculate performance metrics at each value.

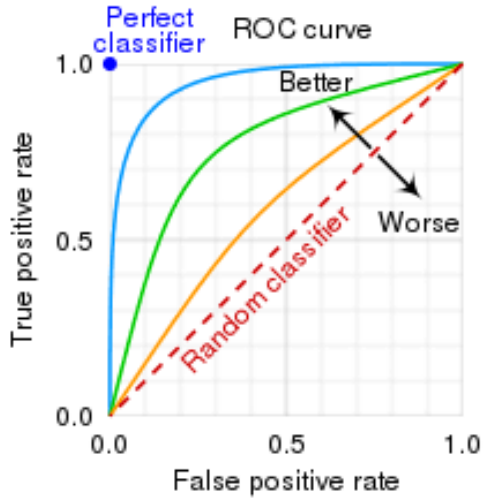
- The ROC curve is a plot of TPR against FPR as the threshold is varied from 0 to 1.

Receiver operating characteristic (ROC) curve



- Top left corner is $FPR = 0$, $\mathrm{T}PR = 1$. This is a perfect classifier.
- Bottom right corner is $FPR = 1$, $\mathrm{T}PR = 0$. This is the worst possible classifier.

Receiver operating characteristic (ROC) curve



- If threshold value is 0, everything will be classified as positive. $FPR = 1, \mathrm{T}PR = 1.$
- If threshold value is 1, everything will be classified as negative. $FPR = 0, \mathrm{T}PR = 0.$
- In between is the interesting region. How close do we get to the top left corner (perfect classifier)?

Receiver operating characteristic (ROC) curve

- The area under the curve (AUC) roughly speaking, a measure of how close the classifier gets to being perfect as the threshold is varied.

- It is possible to show that AUC is equal to the C statistic.

Information criteria

- Consider a maximum likelihood model with k fitted parameters.
- Let $\hat{L}$ be the maximised value of the likelihood.
- The *Akaike information criterion* (AIC) is:

$$-2 \ln \hat{L} + 2k$$

- The *Bayesian information criterion* (BIC) is:

$$-2 \ln \hat{L} + k \ln n$$

- AIC and BIC are typically used in model selection. Models with lower AIC/BIC are preferred.

Information criteria

- AIC is derived from information theory. It estimates the *information loss* when using our model to represent reality.
- BIC is derived from Bayesian statistics. Minimising BIC is equivalent to maximising our *posterior probability* that the model is correct given the data.
- It is possible to create a model that predicts perfectly on a given dataset if we include a large number of parameters. However, this risks *overfitting*.
- Both AIC and BIC penalise additional parameters through the second term.

sparklyr

- In sparklyr, use the following command to calculate model performance metrics:

  ml_evaluate(model, data)

- The metrics that are available vary depending on the type of model that was fitted.

- To extract the predictions from a model:
  ml_predict(model, data)

sparklyr

- These functions can be applied to the result of ml_predict to calculate performance metrics:

  ml_binary_classification_eval(predictions, label_col, prediction_col, metric_name)

  ml_multiclass_classification_eval(predictions, label_col, prediction_col, metric_name)

  ml_regression_evaluator(predictions, label_col, prediction_col, metric_name)