Artificial Intelligence and Machine Learning

Model Evaluation

Fundamentals

- The most important part of the design of an evaluation experiment for a predictive model is to ensure that the data used to evaluate the model is not the same as the data used to train the model.
- The purpose of the evaluation:
  - To determine which model is the most suitable for the task
  - To estimate how the model will perform
  - To convince users that the model will meet their needs

Standard Approach

- Measuring misclassification rate on a hold-out test set.
- For binary prediction problems, there are 4 possible outcomes:
  - True positive (TP)
  - True negative (TN)
  - False positive (FP)
  - False negative (FN)
- Misclassification accuracy
  = (FP + FN)/(TP + TN + FP + FN)

Evaluating the Regression Model

- Residual (Error)

$$e_i = y_i - \hat{y}_i$$

where y_i is the prediction result, and $\hat{y}_i$ is the observed result.

- Residual plots: non-random pattern may indicate that there is some left over correlation.
- Standard error

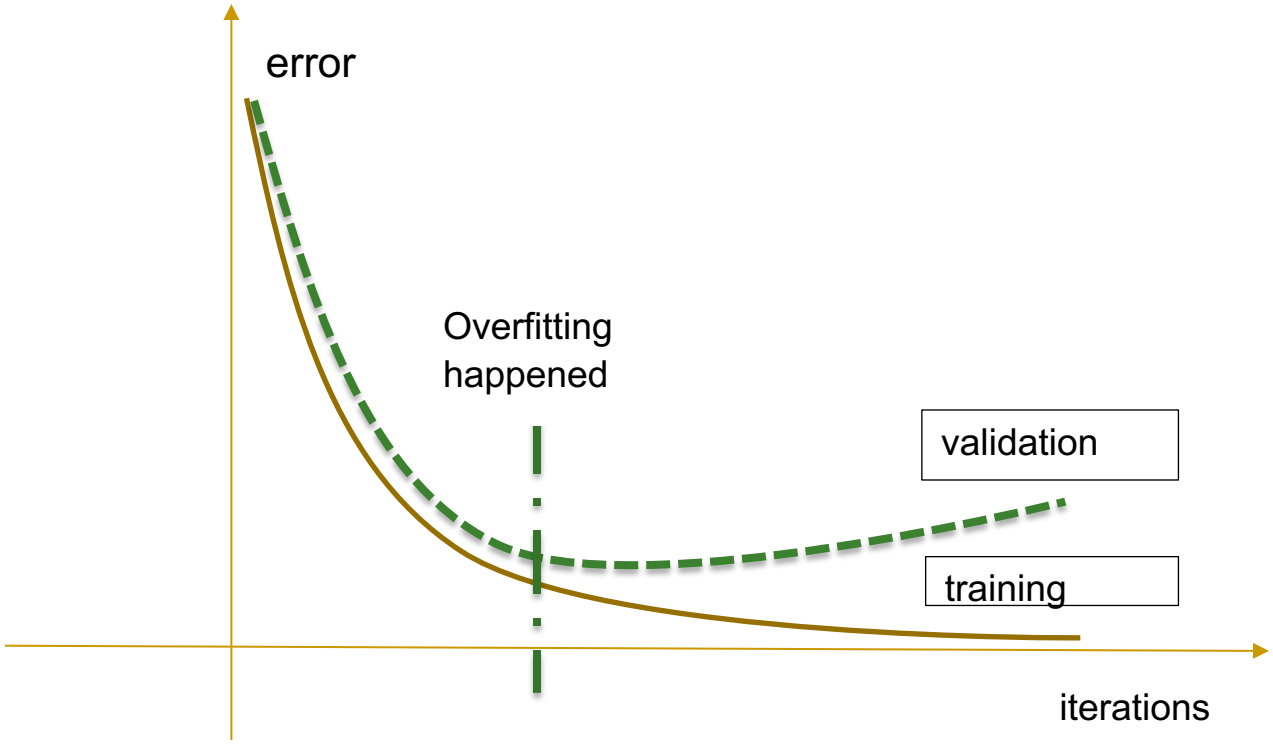
$$s_e = \sqrt{\frac{1}{N-2}\sum_{i=1}^{N} e_i^2}$$

- The standard error is related to the size of the average error that the model produces. We can compare this error to the sample mean of y or with the standard deviation of y to gain some perspective on the accuracy of the model.

Designing Evaluation Experiments

- Hold-out Sampling
  - Divide the full data set into training, validation, and test sets
  - Validation set can be used to avoid overfitting in iterative machine learning algorithms
- K-Fold cross validation
  - Divide the full data set into k sub sets
  - Train K models for evaluation purposes, each one use one of the subsets as test set and the remaining as training set
  - Extreme case: Leave-one-out Cross Validation
- Out-of-time Sampling

Using Validation Dataset to Avoid Overfitting Problem



Performance Measures – Categorical Targets

- Confusion matrix

		Prediction	
		TRUE	FALSE
Target	TRUE	TP	FN
	FALSE	FP	TN

- Rates:

$$TPR = \frac{TP}{(TP + FN)}, TNR = \frac{TN}{(TN + FP)}$$
$$FPR = \frac{FP}{(TN + FP)}, FNR = \frac{FN}{(TP + FN)}$$

- Precision = $\frac{TP}{(TP+FP)}$ (percent of found positives are actually positives)
- Recall = $\frac{TP}{(TP+FN)}$ (percent of true positives are found to be positive)
- Average Class Accuracy = $\frac{1}{|levels(t)|} \sum_{i \in levels(t)} Recall_i$
- F1 measure = $2 \times \frac{(precision \times recall)}{(precision + recall)}$

Performance Measures

- It is not always correct to treat all outcomes equal, sometimes it is useful to take into account the cost of the different outcomes when evaluating models.
- Profit Matrix

		Prediction	
		TRUE	FALSE
Target	TRUE	TP_profit	FN_profit
	FALSE	FP_profit	TN_profit

- Profit = $TP \times TP_{profit} + FN \times FN_{profit} + FP \times FP_{profit} + TN \times TN_{profit}$

Performance Measures – Prediction Scores

- Make the classification prediction model return a score between 0 and 1, then threshold the score

$$\text{output}(\text{score}, \text{threshold}) = \begin{cases} positive, if\ score \geq threshold \\ negative, otherwise \end{cases}$$

- The basis of most of performance measurements is to measure how well the distributions of scores produced by the model for different target levels are separated.

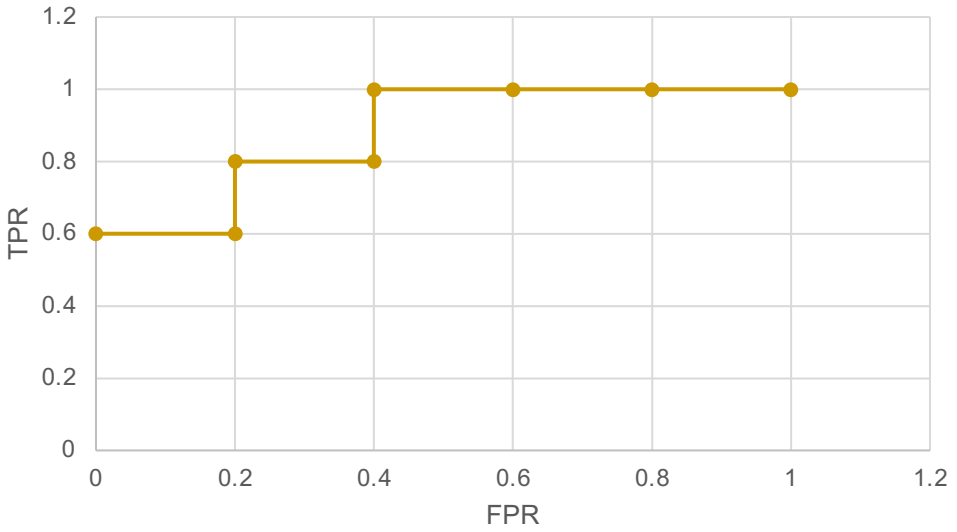
Receiver Operating Characteristic Curve

- TPR and TNR are intrinsically tied to the threshold used to convert prediction scores into target levels.
- This threshold can be changed, however, which leads to different predictions and a different confusion matrix.
- As threshold decreases, TPR increases and FPR also increases.
- Capturing this tradeoff (between a hit rate and a false alarm rate) is the basis of the ROC curve.

ROC Example

ID	target	score
1	TRUE	0.9
2	TRUE	0.85
3	TRUE	0.8
4	FALSE	0.7
5	TRUE	0.65
6	FALSE	0.6
7	TRUE	0.5
8	FALSE	0.4
9	FALSE	0.35
10	FALSE	0.3

cut-off score	FPR	TPR
0.75	0	0.6
0.68	0.2	0.6
0.63	0.2	0.8
0.55	0.4	0.8
0.45	0.4	1
0.38	0.6	1
0.32	0.8	1
0.25	1	1



ROC Index

- The receiver operating characteristic index is based on the receiver operating characteristic curve.
- The ROC index measures the area underneath an ROC curve.
- The ROC index of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance. (Signal Detection Theory)
- ROC index can be used to compare classifiers: the bigger the ROC index the better. (see next slide)
- ROC index = $\sum_{2}^{|T|} \frac{(FPR(T[i]-FPR(T[i-1]))\times(TPR(T[i])+TRP(T[i-1))}{2}$

Evaluating Model using ROC Index

- We can use the traditional academic point system to evaluate a model using ROC index.
  - 0.90 – 1.00 ➔ excellent (A)
  - 0.80 – 0.90 ➔ good (B)
  - 0.70 – 0.80 ➔ fair (C)
  - 0.60 – 0.70 ➔ poor (D)
  - 0.50 – 0.60 ➔ fail (F)
  - < 0.50 ➔ worse than a random model

Performance Measures – Multinomial Targets

- Confusion Matrix for a multinomial prediction problem with K target levels:

		Prediction			Recall
		Level_1	…	Level_K	
Target	Level_1				
	…				
	Level_K				
Precision					

- Precision(i) = $\frac{TP(i)}{TP(i)+FP(i)}$
- Recall(i) = $\frac{TP(i)}{TP(i)+FN(i)}$

Performance Measure – Continuous Targets

- Sum of squared errors

$$\frac{1}{2}\sum_{i=1}^{N}(t_i - \boldsymbol{M}(d_i))^2$$

- Mean squared error

$$\frac{\sum_{i=1}^{N}(t_i - \boldsymbol{M}(d_i))^2}{n}$$

- Root mean squared error

$$\sqrt{\frac{\sum_{i=1}^{N}(t_i - \boldsymbol{M}(d_i))^2}{n}}$$

- Mean absolute error

$$\frac{\sum_{i=1}^{N} abs(t_i - \boldsymbol{M}(d_i))}{n}$$

- Domain independent measures of error

$$R^2 = 1 - \frac{sum\ of\ squared\ errors}{total\ sum\ of\ squares}$$

$$total\ sum\ of\ squares = \frac{1}{2}\sum_{i=1}^{N}(t_i - \bar{t})^2$$

Model Evaluation After Deployment

- Purpose: to capture a signal that indicates model decay has occurred
- Signal source:
  - The performance of the model measured using appropriate performance measures
    - Data may not be available
  - The distribution of the outputs of a model
    - Stability Index
  - The distributions of the descriptive features in query instances presented to the model
    - May have too many data

Stability Index

$$stability\ index = \sum_{k \in levels(t)} \left(\left(\frac{|A_{t=k}|}{|A|} - \frac{|B_{t=k}|}{|B|} \right) \times log_e \left(\frac{|A_{t=k}|}{|A|} \Big/ \frac{|B_{t=k}|}{|B|} \right) \right)$$

- In general,
  - Stability index < 0.1, OK.
  - Stability index is between 0.1 and 0.25, monitor.
  - Stability index > 0.25, significant concept drift has occurred, and corrective action is required.