CM2604 Machine Learning

Evaluation and Ethics

Week 10 | Prasan Yapa

ROBERT GORDON UNIVERSITY ABERDEEN
INFORMATICS INSTITUTE OF TECHNOLOGY

Content

- Model evaluation
- Model selection
- Performance metrics
- Classification evaluation metrics
- Regression evaluation metrics
- Clustering evaluation metrics
- Trade-offs

Model Evaluation

- Model evaluation is a process of assessing the model's performance on a chosen evaluation setup
- Model selection is the process of choosing the best classifier for a given task
- It is done by comparing various model candidates on chosen evaluation metrics
- Choosing the correct evaluation schema, whether a simple train test split or a complex cross-validation strategy

How to evaluate machine learning models?

- Step 1: *Choose a proper validation strategy*
- Step 2: *Choose the right evaluation metric*
- Step 3: *Keep track of your experiment results*
- Step 4: *Compare experiments and pick a winner*

Model selection in machine learning

- Resampling methods
  - simple techniques of rearranging data samples to inspect if the model performs well on data samples
  - resampling helps us understand if the model will generalize well
- Random split
  - used to randomly sample a percentage of data into training, testing, and preferably validation sets
  - random splitting will prevent a biased sampling of data: test set is used for model evaluation

Model selection in machine learning

- Time-Based Split
  - There are some types of data where random splits are not possible
  - If we have to train a model for weather forecasting, we cannot randomly divide the data into training and testing sets.
- K-Fold Cross-Validation
  - randomly shuffling the dataset and then splitting it into k groups
  - model is tested on the test group and the process continues for k groups
- Stratified K-Fold
  - unlike in k-fold cross-validation, the values of the target variable is taken into consideration in stratified k-fold.

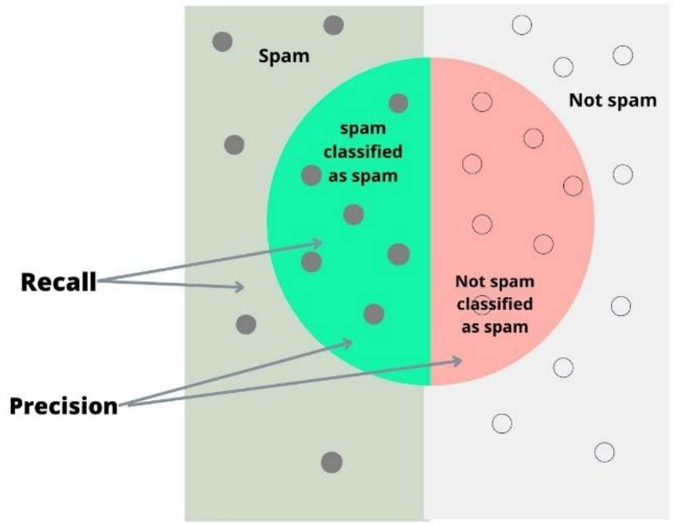
How to choose performance metrics

- Right choice of an evaluation metric is crucial and often depends upon the problem
- A clear understanding of a wide range of metrics can help the evaluator to chance upon an appropriate match
- Types
  - Classification metrics
  - Regression metrics
  - Clustering metrics

Classification evaluation metrics

- For every classification model prediction, a matrix called the confusion matrix can be constructed

	Actual 0	Actual 1
Predicted 0	True Negatives (TN)	False Negatives (FN)
Predicted 1	False Positives (FP)	True Positives (TP)



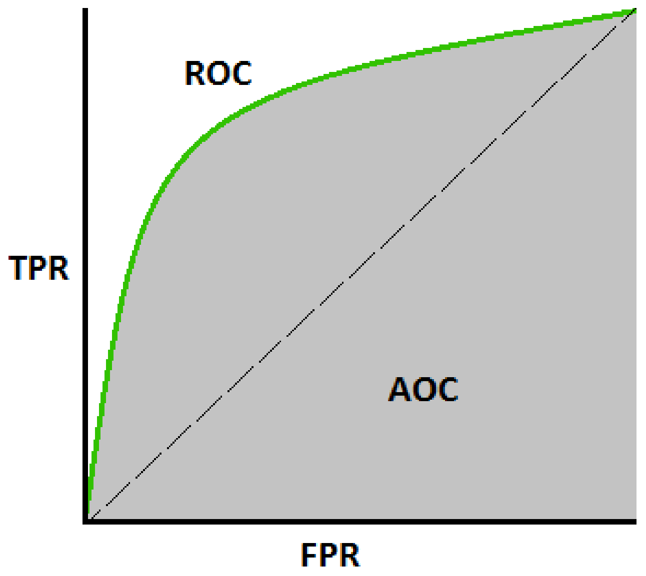
  - TN: Number of negative cases correctly classified
  - TP: Number of positive cases correctly classified
  - FN: Number of positive cases incorrectly classified as negative
  - FP: Number of negative cases correctly classified as positive

Classification evaluation metrics

- $Precision = \frac{True\ Positive}{True\ Positive+False\ Positive}$ -> Exactness
- $Recall = \frac{True\ Positive}{True\ Positive+False\ Negative}$ -> Fraction of positives
- $F1 = 2 * \frac{Precision*Recall}{Precision+Recall}$ -> Harmonic mean of P and R
- $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$
    - number of test cases correctly classified divided by the total number of test cases

Classification evaluation metrics

- AUC-ROC
  - ROC curve is a plot of true positive rate (recall) against false positive rate (TN / (TN+FP))
  - AUC-ROC stands for Area Under the Receiver Operating Characteristics and the higher the area
  - If the curve is somewhere near the 50% diagonal line, it suggests that the model randomly predicts the output variable



AUC - ROC Curve [Image 2] (Image courtesy: My Photoshopped Collection)

Classification evaluation metrics

- Log loss
  - Log loss is a very effective classification metric and is equivalent to -1* log (likelihood function) where the likelihood function suggests how likely the model thinks the observed set of outcomes was
- Gain & Lift
  - Lift charts measure the improvement that a model brings in compared to random predictions.
  - Gain and lift charts evaluate the model on portions of the whole population
- K-S chart
  - The K-S chart determines the degree of separation between positive class distribution and the negative class distribution
  - he higher the difference, the better is the model at separating the positive and negative cases

Regression evaluation metrics

- Regression models provide a continuous output variable, unlike classification models that have discrete output variables
- Mean Squared Error
  - Calculates the difference between the actual value and the predicted value (error)
- Root Mean Squared Error
  - Helps to bring down the scale of the errors closer to the actual values, making it more interpretable
- Mean Absolute Error
  - Mean of the absolute error values (actuals – predictions)

Clustering evaluation metrics

- Clustering algorithms predict groups of datapoints and hence, distance-based metrics are most effective
- Dunn Index
  - Focuses on identifying clusters that have low variance
- Silhouette Coefficient
  - Tracks how every point in one cluster is close to every point in the other clusters in the range of -1 to +1
- Elbow method
  - Determine the number of clusters by plotting the number of clusters on the x-axis against the percentage of variance explained on the y-axis

Trade-offs in model selection

- Bias vs Variance
  - A model with high bias will oversimplify by not paying much attention to the training points
  - Bias occurs when a model is strictly ruled by assumptions
  - This leads to underfitting when the actual values are non-linearly related to the independent variables
  - A model with high variance will restrict itself to the training data by not generalizing for test points
  - Variance is high when a model focuses on the training set too much

Ethics in Machine Learning

Prasan Yapa

Key concerns

- Data
  - A good ML system needs lots of data. But where are we going to get this data?
  - Is it alright if you steal the someone's private data?
- Algorithms
  - What if a patented algorithm, in the right hands can help millions?
  - Can one's own sense of right and wrong be used to reverse engineer the algorithm to benefit others?
- Results
  - If you get the same practice questions in the exam, is your score on the exam a good measure of how much you learnt?
  - Or is it a measure of how much you were able to memorize?

Questions