CS 789 ADVANCED BIG DATA

EVALUATION

* Some contents are adapted from Dr. Hung Huang and Dr. Chengkai Li at UT Arlington
Evaluation for Classification

Data set

Training set

Learn on the training set

The model

Evaluate on the test set
Evaluation Metrics

- Confusion Matrix: shows performance of an algorithm, especially predictive capability.
  - rather than how fast it takes to classify, build models, or scalability.

<table>
<thead>
<tr>
<th>Actual Class</th>
<th>Predicted Class</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Class = Yes</td>
<td>Class = YES</td>
<td>True Positive</td>
</tr>
<tr>
<td>Class = No</td>
<td>Class = No</td>
<td>False Positive</td>
</tr>
<tr>
<td>Class = Yes</td>
<td>Class = No</td>
<td>True Negative</td>
</tr>
<tr>
<td>Class = No</td>
<td>Class = YES</td>
<td>False Positive</td>
</tr>
</tbody>
</table>
# Evaluation Metrics

<table>
<thead>
<tr>
<th>Predicted condition</th>
<th>Total population</th>
<th>Predicted Condition positive</th>
<th>Predicted Condition negative</th>
<th>Prevalence</th>
<th>True positive (Type I error)</th>
<th>False Negative (Type II error)</th>
<th>False positive rate (FPR), Miss rate</th>
<th>True negative</th>
<th>False positive rate (FPR), Fall-out</th>
<th>Sensitivity, Recall</th>
<th>False negative rate (TNR), Specificity (SPC)</th>
<th>True negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>True condition</td>
<td>condition positive</td>
<td>True positive</td>
<td>False Positive</td>
<td></td>
<td>false positive</td>
<td>true negative</td>
<td>false positive</td>
<td>true negative</td>
<td>false positive</td>
<td>false negative</td>
<td>true negative</td>
<td>true negative</td>
</tr>
<tr>
<td></td>
<td>condition negative</td>
<td>False Positive</td>
<td>True negative</td>
<td></td>
<td>true negative</td>
<td>false negative</td>
<td>true negative</td>
<td>true negative</td>
<td>true negative</td>
<td>true negative</td>
<td>true negative</td>
<td>true negative</td>
</tr>
</tbody>
</table>

- **Accuracy (ACC)**: $\frac{\sum \text{True positive} + \sum \text{True negative}}{\sum \text{Total population}}$
- **Positive predictive value (PPV), Precision**: $\frac{\sum \text{True positive}}{\sum \text{Test outcome positive}}$
- **False omission rate (FOR)**: $\frac{\sum \text{False negative}}{\sum \text{Test outcome negative}}$
- **Positive likelihood ratio (LR+)**: $\frac{\text{TPR}}{\text{FPR}}$
- **Diagnostic odds ratio (DOR)**: $\frac{\text{LR+}}{\text{LR-}}$
- **Negative predictive value (NPV)**: $\frac{\sum \text{True negative}}{\sum \text{Test outcome negative}}$
- **Negative likelihood ratio (LR-)**: $\frac{\text{FNR}}{\text{TNR}}$
Type I and II error

Type I error (false positive)

Type II error (false negative)

- You’re pregnant
- You’re not pregnant
Evaluation Metrics

- **Sensitivity or True Positive Rate (TPR)**
  \[ \frac{TP}{TP+FN} \]

- **Specificity or True Negative Rate (TNR)**
  \[ \frac{TN}{FP+TN} \]

- **Precision or Positive Predictive Value (PPV)**
  \[ \frac{TP}{TP+FP} \]

- **Negative Predictive Value (NPV)**
  \[ \frac{TN}{TN+FN} \]

- **Accuracy**
  \[ \frac{TP+TN}{TP+FP+TN+FN} \]
Consider a binary classification problem
- Number of Class 0 examples = 9990
- Number of Class 1 examples = 10
- If predict all as 0, accuracy is $\frac{9990}{10000}=99.9\%$

- Precision
- Recall

- Weighted Accuracy $= \frac{w_{TP}TP + w_{TN}TN}{w_{TP}TP + w_{FP}FP + w_{TN}TN + w_{FN}FN}$
ROC curve

- Receiver Operating Characteristic
  - Graphical approach for displaying the tradeoff between true positive rate (TPR) and false positive rate (FPR) of a classifier
    - $\text{TPR} = \frac{\text{positives correctly classified}}{\text{total positives}}$
    - $\text{FPR} = \frac{\text{negatives incorrectly classified}}{\text{total negatives}}$
  - TPR on y-axis and FPR on x-axis
ROC curve
ROC curve
ROC curve

- Points of interests (TP, FP)
  - (0, 0): everything is negative
  - (1, 1): everything is positive
  - (1, 0): perfect (ideal)

- Diagonal line
  - Random guessing (50%)

- Area Under Curve (AUC)
  - Measurement how good the model on the average
  - Good to compare with other methods
Evaluation

- Model Selection
  - How to evaluate the performance of a model?
  - How to obtain reliable estimates?

- Performance estimation
  - How to compare the relative performance with competing models?
Motivation

- We often have a finite set of data
  - If using the entire training data for the best model,
    - The model normally overfits the training data, where it often gives almost 100% correct classification results on training data
- Better to split the training data into disjoint subsets
- Note that test data is not used in any way to create the classifier → Cheating!
Methods of Validation

- **Holdout**
  - Use 2/3 for training and 1/3 for testing

- **Cross-validation**
  - Random subsampling
  - K-Fold Cross-validation
  - Leave-one-out

- **Stratified cross-validation**
  - Stratified 10-fold cross-validation is often the best

- **Bootstrapping**
  - Sampling with replacement
  - Oversampling vs undersampling
Holdout

- Split dataset into two groups for training and test
  - Training dataset: used to train the model
  - Test dataset: use to estimate the error rate of the model

- Drawback
  - When “unfortunate split” happens, the holdout estimate of error rate will be misleading
Random Subsampling

- Split the data set into two groups
  - Randomly selects a number of samples without replacement
    - Usually, one third for testing, the rest for training
K-Fold Cross-validation

- **K-fold Partition**
  - Partition K equal sized sub groups
  - Use K-1 groups for training and the remaining one for testing

```
Experiment 1
 Experiment 2
 Experiment 3
 Experiment 4
 Experiment 5
```

Test set
K-fold cross-validation

- Suppose that $E_i$ is the performance in the $i$-th experiment.
- The average error rate is

$$E = \frac{1}{K} \sum_{i=1}^{k} E_i$$
Leave-one-out cross-validation

- Use N-1 samples for training and the remaining sample for testing (i.e., there is only one sample for testing)
- The average error rate is

\[ E = \frac{1}{N} \sum_{i=1}^{N} E_i \]

where N is the total sample number.
How many folds?

- If a large number of folds
  - Bias to the true estimator will be small.
  - The estimator will be accurate
  - Computationally expensive

- If a small number of folds
  - Cheap computational time for experiments
  - Variance of the estimator will be small
  - Bias will be large

- 5 or 10-Fold CV is a common choice for K-fold CV
Stratified cross-validation

- When randomly selecting training or test sets, ensure that class proportions are maintained in each selected set.

1. Stratify instances by class
2. Randomly select instances from each class proportionally

Bootstrapping

- Oversampling
  - Amplifying the minor class samples so that the classes are equally distributed
Bootstrapping

- Undersampling
  - Consider less numbers of samples in the major class so that the classes are equally distributed
Thinking of experimental results

- Comparing averages?
- Only one CV experiment?
- Measuring accuracy only?
- How to make ROC curve with multiple experiments?
- Cross-validation with normalization