Evaluation of Binary Classifiers

MIE 1624
University of Toronto
St. George Campus
Winter 2018

Binary Classification

- Observed response y – two possible values + and -
- Define relationship b/t h(x) and y
- Use the decision rule:
- E.g. Sentiments analysis: $\hat{y} = \begin{cases} +, & score \geq 0 \\ -, & score < 0 \end{cases}$

Evaluation of Binary Classifiers

- **Binary Classifier:** **algorithm** that categorizes the elements of a given set into two disjoint pre-defined groups.
  - ➢The two categories are considered dichotomous and the elements of the given set are labeled “positive” or “negative”.
- **Classification**: the **output** of a classifier on a given set
  - ➢ i.e. the number of “positives” & the number of “negatives”.
- **Prevalence**: how often a classification category occurs in the population
- **Example:** in sentiment analysis, Twitter data is divided (classified) into “positive” and “negative” tweets.

Positives

- **True positives (TP):** the elements in the given set (e.g., tweets) that are **"positive"** and are correctly identified by the classifier as "**positive**".

- **False negatives (FN):** the elements that are **"positive",** but are incorrectly classified as **"negative".**

- **Condition Positive** (CP): TP + FN

  - All can be arranged into a 2×2 **confusion matrix** (classification results on the vertical axis and the true category on the horizontal axis).

Negatives

- **True negatives(TN):** the items that are **"negative"** and correctly identified as such by the algorithm.

- **False positives(FP):** the items that are **"negative"** and incorrectly classified as **"positive"**

- **Condition Negative** (CP): TN + FP.

Accuracy

- The percentage of **correctly** classified instances among the total number of cases examined:

$$(TN + TP)/(TP + FP + FN + TN)$$

- TP true positives
- TN true negatives
- FP false positives
- FN false negatives

True positive	False Negative (Type II error)
False Positive (Type I error)	True negative

Sensitivity / Recall / True Positive Rate (TPR)

- Proportion of elements (e.g., tweets) that were classified as **positive,** and are indeed **positive,** of all the elements that are in fact positive

- **Meaning** of **high** sensitivity: fewer **actual positives** go undetected
  - epidemiology: fewer patients go undetected
  - factory quality control: fewer faulty products go to the market.

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

Specificity / True Negative Rate (TNR)

- Proportion of elements (e.g., tweets) that were classified **Negative,** and are indeed **Negative,** of all the elements that are in fact **Negative**

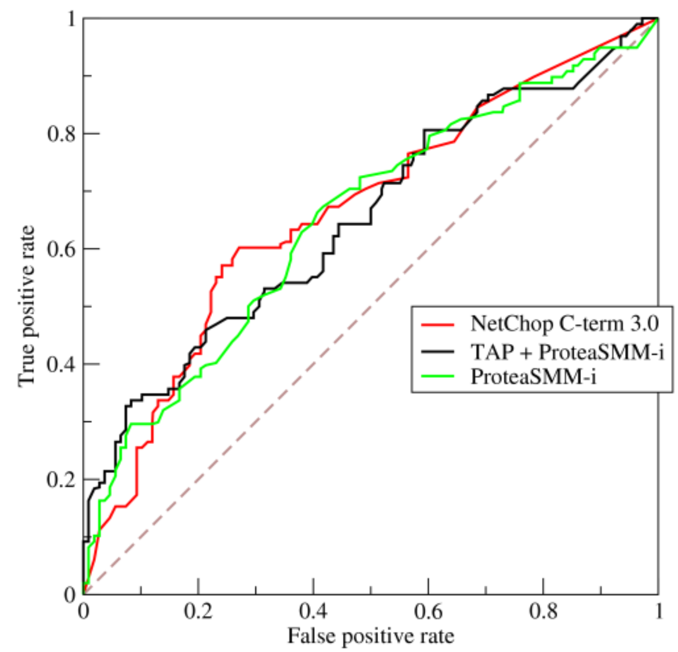
- **Meaning** of **high** specificity: fewer positive cases are mislabeled.
  - in epidemiology: fewer healthy people are labeled as sick
  - Factory quality control: fewer good products are thrown away

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

- **Note**: *sensitivity* and *specificity* are independent: i.e., is possible to achieve 100% in both.

Receiver Operating Characteristic (ROC)

The relationship between sensitivity and specificity can be visualized using the ROC curve.



*https://en.wikipedia.org/wiki/Receiver_operating_characteristic

Positive and Negative Predictive Values

- ***positive*** classification result , how well does that ***predict*** an actual positive value?
  - **Positive Predictive Value (PPV),** a.k.a. **Precision**: the proportion of true positives out of all positive results.

$$precision = \frac{TP}{TP + FP}$$

- ***negative*** classification result, how well does that ***predict*** an actual negative value?
  - **Negative Predictive Value (NPV)** the proportion of true negatives out of all negative results.

$$\frac{TN}{TN + FN}$$

<table>
<tr><td></td><td></td><td colspan="2">True condition</td><td></td><td colspan="2"></td></tr>
<tr><td></td><td>Total population</td><td>Condition positive</td><td>Condition negative</td><td>Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$</td><td colspan="2">Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$</td></tr>
<tr><td rowspan="2">Predicted condition</td><td>Predicted condition positive</td><td>True positive, Power</td><td>False positive, Type I error</td><td>Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$</td><td colspan="2">False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$</td></tr>
<tr><td>Predicted condition negative</td><td>False negative, Type II error</td><td>True negative</td><td>False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$</td><td colspan="2">Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$</td></tr>
<tr><td></td><td></td><td>True positive rate (TPR), Recall, Sensitivity, probability of detection $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$</td><td>False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$</td><td>Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$</td><td rowspan="2">Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$</td><td rowspan="2">$F_1$ score = $\frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$</td></tr>
<tr><td></td><td></td><td>False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$</td><td>True negative rate (TNR), Specificity (SPC) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$</td><td>Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$</td></tr>
</table>

Confusion Matrix

*https://en.wikipedia.org/wiki/Confusion_matrix

Confusion Matrix

- The numbers in the Confusion Matrix can be totaled into **grand totals** and **marginal totals**
- The entire table, true positives, false negatives, true negatives, and false positives adds up to 100% of the set.
- The number of true positives and false positives add up to 100% of the test positives (likewise for negatives).
- The number of true positives and false negatives add up to 100% of the condition positives (likewise for negatives).
- Further statistics can be obtained by taking ratios, ration of these ratios, or more complicated functions.

F-measure / F-score

- Combines **precision** and **recall** into a single score.

- The score can be interpreted as a **weighted** average of the precision and recall

  - The traditional or balanced F-score, a.k.a. the F1-score is the harmonic mean of precision and recall
  - **F = 1 is considered as the best, 0 is the worse**

Note: F-measures do not take the **negatives** into account

F-measure / F-score

$$F_1 = 2 \cdot \frac{1}{\frac{1}{recall} + \frac{1}{precision}}$$

$$= \frac{2 \times precision \times recall}{precison + recall} = \frac{2TP}{(TP+FP)(TP+FN)} / (\frac{1}{TP+FN} + \frac{1}{TP+FP})$$

$$= \frac{2TP}{(TP + FP)(TP + FN)} * \frac{(TP + FP)(TP + FN)}{TP + FP + TP + FN}$$

The best case, we set FN = 0, FP = 0, then

$$= \frac{2TP}{1} * \frac{1}{TP + 0 + TP + 0} = 1$$

$\boldsymbol{* \; TP + FP + TP + FN \neq 1}$

G-Measure

- Combines **precision** and **recall** into a single score

- The G-Measure is the geometric mean of precision and recall:

$$G = \sqrt{precision \cdot recall}$$

$$= \sqrt{\frac{TP}{TP + FP} \cdot \frac{TP}{TP + FN}}$$

https://en.wikipedia.org/wiki/F1_score

Drawbacks

• Accuracy, Precision/Recall, Sensitivity/Specificity, F-measure etc. suffer from the following problems:

- ➢The performance results are summarized into one or two numbers
  - -> important information is lost.
- ➢Do not always apply to **multi-class domains**.
- ➢Do not aggregate well when the performance of the classifier is considered over **multiple domains**.

Comparison Metrics

- Metrics Characteristics:

  - Prevalence: dependence / independence

    E.g. Sensitivity is a prevalence-independent statistics

  - Domain-Dependent Preference

    E.g. Sensitivity and specificity -> bio-medical domains

    Precision and recall -> computer scientists

High-dimensional Data Analysis
- Classifier Evaluation

- Projection approaches (visualization):
    - A easier way to assess classifier performance results.
    - Multiple views of classifier performance

1. Classifiers on all the domains

2. Generate performance matrices
    e.g., confusion matrix

3. Graph a projection and its distance measure

- Note: the previous performance measures are one class of projections

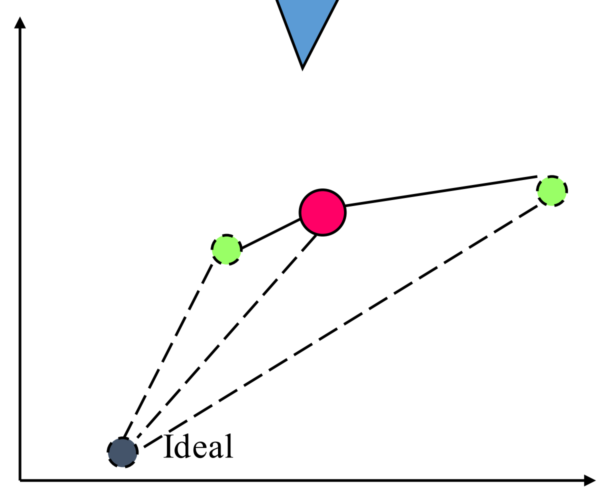
Confusion matrices for a single classifier on three domains

True class →	Pos	Neg
Yes	82	17
No	12	114

True class →	Pos	Neg
Yes	15	5
No	25	231

True class →	Pos	Neg
Yes	99	6
No	1	94

82	17	12	114	15	5	25	231	99	6	1	94



Evaluate the algorithm

- Test the whole dataset w/o training
  - Performance of the algorithm on the other dataset can not be predicted
  - Hint: Control Experiment
- Solutions?

Split Dataset

- Training Dataset e.g. 60%
- Test Dataset e.g. 40%
  - Estimate the performance of the algorithm on unseen data

- Classifier is ran on the training dataset, model created
- Evaluate the model on test dataset, calculate performance measures

  - Note: different splits result in different evaluation on algorithm (even the same percentage)
  - **Model Variance**

Cross Validation

- Reduce the variance of performance scores
- Ensure data instance is trained and tested for the same times
  - K-fold cross validation, where k # of splits to make in the dataset
- E.g. K = 10. Split the dataset into 10 parts. Algorithm runs 10 times.
  - Each run, algorithm is trained on 90% of the dataset, tested on 10%

- Note: Cross Validation only estimates the performance on the **same dataset**

- Issues in Inferring Votes With Sentiment Analysis

1. Metaphors and ironies might be misinterpreted into the opposite sentiment label
2. The subtleties of political language are missed or misinterpreted