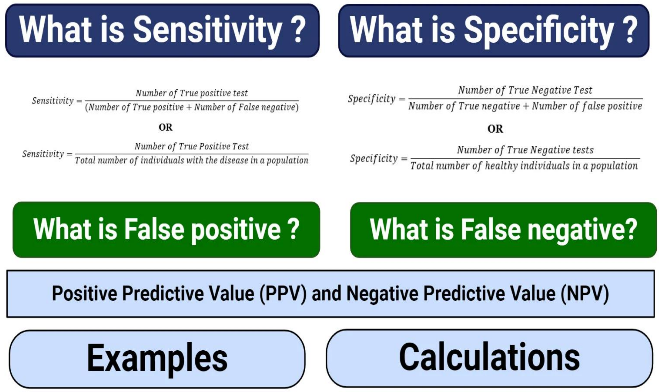
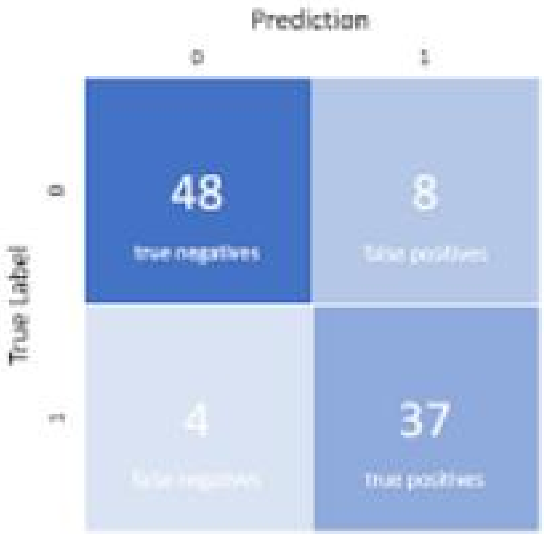
Chapter-5

Evaluation Metrics for supervised learning



		Reference Data			
		Water	Forest	Urban	Total
Classified Data	Water	21	6	0	27
	Forest	5	31	1	37
	Urban	7	2	22	31
	Total	33	39	23	95

		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative



Introduction

- Evaluation aims at selecting the most appropriate **learning schema** for a specific problem

- We evaluate its ability to generalize what it has been learned from the training set on the **new unseen instances**

- Comparison of multiple classifiers on a specific domain (e.g. to find the best algorithm for a given application task)

Absolute and Mean Square Error

– Refers to the error committed to classify an object to the desired class
– Error is defined as the difference between the **desired value** and the **predicted value**

$$Absolute\ Error = \sum_{i=1}^{N} |e_i|$$

$$Mean\ Square\ Error\ (MSE) = \frac{1}{N}\left(\sum_{i=1}^{N} e_i^{\,2}\right)$$ where e_i = desired – predicted value

Accuracy

$$Accuracy = \frac{number\ of\ correctly\ classified\ instances}{total\ number\ of\ instances} \times 100$$

- It assumes equal cost for all classes
- Misleading in unbalanced datasets
- It doesn’t differentiate between different types of errors

- Ex 1:
  - Cancer Dataset: 10000 instances, 9990 are ***normal***, 10 are ***ill ,*** If our model classified all instances as ***normal*** accuracy will be 99.9 %
  - Medical diagnosis: 95 % healthy, 5% disease.
  - e-Commerce: 99 % do not buy, 1 % buy.
  - Security: 99.999 % of citizens are not terrorists.

Binary classification Confusion Matrix

		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

- Type I error: is equivalent to a False positive.
- Type II error: is equivalent to a False negative.

- FN+TP being the total number of positives
- TN+FP being the total number of Negatives

*Is Type 1 or 2 error worse?

Binary classification Confusion Matrix

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

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

$$Success\ rate = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Error\ rate = 1 - \text{success rate}$$

TN rate= TN/TN+FP
FN rate= FN/FN+TP

Where TP= T*rue Positive Rate, FP= False Positive Rate, Accuracy=Success rate and Loss=error rate*

Sensitivity & Specificity

- Sensitivity: Measures the classifier ability to detect positive classes (its positivity)

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

- Specificity: The specificity measures how accurate is the classifier in not detecting too many false positives (it measures its negativity)

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

Example:

Dataset:

- Contains 39 instances, 10 attributes
- The class labels are "negative, positive"
- 22 positive & 17 negative instances.

Classifier used: J48 – 10 folds cross validation

Confusion Matrix:

Classified as→	Positive	Negative
Positive	22	0
Negative	17	0

Classifier Accuracy = $\frac{22}{39} \times 100 = 56.4\%$

- TP= 22
- TN= 0
- FP= 17
- FN= 0
- Sensitivity = $\frac{22}{22+0} = 1$ → this means that all positive cases are classified correctly
- Specificity = $\frac{0}{17+0} = 0$ → this means that no negative cases are classified (i.e.) the classifier classifies everything as positive

Recall & Precision

- It is used by information retrieval researches to measure accuracy of a search engine, they define the recall as (number of relevant documents retrieved) divided by (total number of relevant documents)
- **Recall** (also called **Sensitivity** in some fields) measures the proportion of actual positives which are correctly identified as such (e.g. the percentage of sick people who are identified as having the condition):

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

- **Precession** of class **Yes** in classification can defined as the number of instance classified **correctly** as class **Yes** divided by the total number of instances classified as **Yes** by the classifier

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

F-measure

- The ***F-measure*** is the harmonic-mean (average of rates) of precision and recall and takes account of both measures.

$$F\ measure = \frac{1}{\frac{1}{Recall} + \frac{1}{Precision}} = \frac{2 \times Precision \times Recall}{Precision + Recall} = \frac{2 \times TP}{2 \times TP + FP + FN}$$

- It is biased towards all cases except the true negatives

Example:

Dataset:

- Contains 39 instances, 10 attributes
- The class labels are "negative, positive"
- 22 positive & 17 negative instances.

Classifier used: J48 – 10 folds cross validation

Confusion Matrix:

Classified as →	Positive	Negative
Positive	22	0 **FN**
Negative	17	0

Classifier Accuracy = $\frac{22}{39} \times 100 = 56.4\%$

- TP= 22
- TN= 0
- FP= 17
- FN= 0
- The area under ROC curve: 0.5 in both cases, cause the TP rate = FP rate.
- Precision & Recall
  - Recall = $\frac{22}{22+0} = 1$
  - Precision = $\frac{22}{22+17} = 0.564$
- The *F-measure* = $= \frac{2 \times 22}{2 \times 22 + 17 + 0} = 0.7213$

Multiclass classification

- For **Multiclass prediction** task, the result is usually displayed in confusion matrix where there is a row and a column for each class,
  - Each matrix element shows the number of test instances for which the actual class is the row and the predicted class is the column
  - Good results correspond to large numbers down the diagonal and small values (ideally zero) in the rest of the matrix

Classified as	**a**	**b**	**c**
A	TP_{aa}	FN_{ab}	FN_{ac}
B	FP_{ab}	TN_{bb}	$\boldsymbol{FN_{bc}}$
C	FP_{ac}	$\boldsymbol{FN_{cb}}$	TN_{cc}

Multiclass classification

- For example in three classes task {a , b , c} with the confusion matrix below, if we selected a to be the class of interest then

$$\text{True positives for class } \boldsymbol{a} = \text{TP}_{\text{aa}}$$

$$\text{True Negatives for class } \boldsymbol{a} = \text{TN}_{\text{cc}} + \text{TN}_{\text{bb}}$$

$$\text{False Positives for class } \boldsymbol{a} = \text{FP}_{\text{ab}} + \text{FP}_{\text{ac}}$$

$$\text{False Negatives for class } \boldsymbol{a} = \text{FN}_{\text{ab}} + \text{FN}_{\text{ac}}$$

- Note that we don’t care about the values (FNcb & FNbc) as we are considered with evaluating how the classifier is performing with class a, so the misclassifications between the other classes is out of our interest.

Notes on Metrics

- As we can see the **True Positive rate = Recall = Sensitivity** all are measuring how good the classifier is in finding true positives.

- When **FP rate** increases, **specificity** & **precision** decreases & vice verse,

- It doesn't mean that **specificity** and **precision** are correlated,
  - For example in unbalanced datasets the precision can be very low where the specificity is high
  - Cause the number of instances in the negative class is much higher than the number of positive instances

Other evaluation metrics (NLP)

- Word error rate (WER)
- Character Error Rate(CER)
- Bilingual Evaluation Understudy (BLEU) score

Further reading

- **Analysis of variance (ANOVA):** is a statistical method that separates observed variance data into different components to use for additional tests.
- **Maximum Likelihood Estimation (MLE):** is a method that determines values for the parameters of a model. The parameter values are found such that they maximise the likelihood that the process described by the model produced the data that were actually observed.
- **Interval estimation:** is the use of sample data to calculate an interval of possible values of an unknown population parameter

What is Sensitivity ?

$$\text{Sensitivity} = \frac{\text{Number of True positive test}}{(\text{Number of True positive} + \text{Number of False negative})}$$

OR

$$\text{Sensitivity} = \frac{\text{Number of True Positive Test}}{\text{Total number of individuals with the disease in a population}}$$

What is Specificity ?

$$\text{Specificity} = \frac{\text{Number of True Negative Test}}{\text{Number of True negative} + \text{Number of false positive}}$$

OR

$$\text{Specificity} = \frac{\text{Number of True Negative tests}}{\text{Total number of healthy individuals in a population}}$$

What is False positive ?

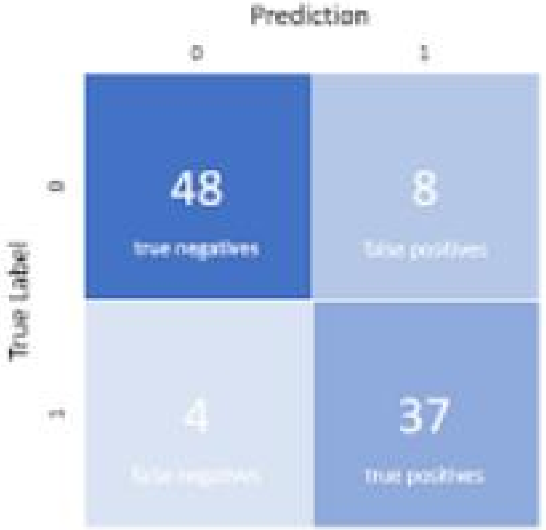
What is False negative?

Positive Predictive Value (PPV) and Negative Predictive Value (NPV)

Examples

Calculations

		Reference Data			
		Water	Forest	Urban	Total
Classified Data	Water	21	6	0	27
	Forest	5	31	1	37
	Urban	7	2	22	31
	Total	33	39	23	95



		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

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