Methods of Cross Validation

Week 5

Dr. Nouman M Durrani

Acknowledgement to all authors whose materials have been used

Basic Concepts: Training and Testing

- **Model:** A **model** is a mathematical formula with a number of parameters that need to be learned from the data
  - Fitting a model to the data is a process known as model training
  - Testing data is the unseen data for which predictions have to be made
    - Test data is used only to assess performance of model
  - Training data's output is available to model

Basic Concepts: Training and Testing

- Success: instance (record) class is predicted correctly
- Error: instance class is predicted incorrectly
- Error rate: a percentage of errors made over the whole set of instances (records) used for testing
- Predictive Accuracy: a percentage of well classified data in the testing data set.

REMEMBER: We must know the classification (class attribute values) of all instances (records) used in the test procedure.

Training and Testing

Example:

- Testing Rules (testing record #1) = record #1.class –Succ
- Testing Rules (testing record #2) not= record #2.class –Error
- Testing Rules (testing record #3) = record #3.class –Succ
- Testing Rules (testing record #4) = instance #4.class –Succ
- Testing Rules (testing record #5) not= record #5.class –Error

Error rate: 2 errors: #2 and #5
Error rate = 2/5=40%
Predictive Accuracy: 3/5 = 60%

Confusion Matrix

- A confusion matrix is a table that is used to describe the performance of a classification model on a set of test data for which the true values are known.
- It helps in visualization the performance of an algorithm.

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

Confusion Matrix

- *TN* is the number of correct predictions that an instance is negative
- *TP* is the number of correct predictions that an instance is positive
- *FN* is the number of incorrect predictions that an instance is negative
- *FP* is the number of incorrect predictions that an instance is positive

Confusion Matrix

- Confusion Matrix for the following results is:

ID	Actual	Predicted
1	1	1
2	0	0
3	1	1
4	1	0
5	0	1

		Predicted	
Actual		Negative	Positive
	Negative	True Negative 1	False Positive 1
	Positive	False Negative 1	True Positive 2

```
# Example of a confusion matrix in Python
from sklearn.metrics import confusion_matrix

expected  = [1, 1, 0, 1, 0, 0, 1, 0, 0, 0]
predicted = [1, 0, 0, 1, 0, 0, 1, 1, 1, 0]
results = confusion_matrix(expected, predicted)
print(results)
```

Output Confusion Matrix: [[4 2]
[1 3]]

Confusion Matrix

Several standard terms have been defined for the 2 class matrix

- The *accuracy* (*AC*) is the proportion of the total number of predictions that were correct

$$Accuracy = \frac{TN + TP}{TN + FN + TP + FP}$$

- Accuracy = 3 / 4 = 75%

Confusion Matrix

- *The True positive rat*e (*TPR*) is the proportion of positive cases that were correctly identified

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

- The *false positive rate* (*FPR*) is the proportion of negatives cases that were incorrectly classified as positive

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

- TPR or **recall** = 2 / 3 = 66.7%
- FPR = 0 / 1 = 0 %

	Predicted Negative	Predicted Positive
Actual Negative	True Negative (TN)	False Positive (FP)
Actual Positive	False Negative (FN)	True Positive (TP)

Confusion Matrix

- The *true negative rate* (*TNR*) is defined as the proportion of negatives cases that were classified correctly,

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

- The *false negative rate* (*FNR*) is the proportion of positives cases that were incorrectly classified as negative

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

- TNR = 1 / 1 = 100%
- FNR = 1 / 3 = 33.3%

	Predicted Negative	Predicted Positive
Actual Negative	True Negative (TN)	False Positive (FP)
Actual Positive	False Negative (FN)	True Positive (TP)

Confusion Matrix

- *Precision* (*P*) is the proportion of the predicted positive cases that were correct,

$$precision = \frac{tp}{tp + fp}$$

- Precision = 2/2 = 100%
- F measure is harmonic mean of precision and recall

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

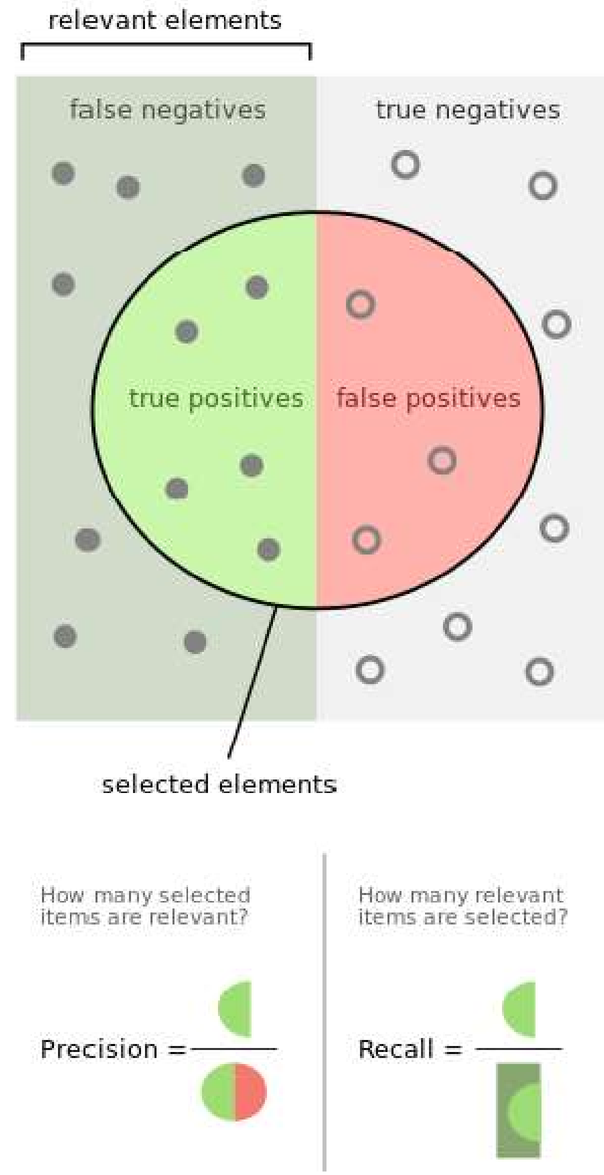
- F1 = (2 * 1 * 0.667)/(1+0.667) = 0.8

Exercise

		Actual	
		Negative	Positive
Predicted	Negative	9760	40
	Positive	140	60

- **Precision** (also called positive predictive value) is the fraction of relevant instances among the retrieved instances
- **Recall** (also known as sensitivity) is the fraction of the total amount of relevant instances that were actually retrieved instances

- Suppose a computer program for recognizing dogs in photographs identifies 8 dogs in a picture containing 12 dogs and some cats. Of the 8 identified as dogs, 5 actually are dogs (true positives), while the rest are cats (false positives)
  - The program's precision is 5/8 while its recall is 5/12(=5/(8+4 for failing to recognize the other 4 cats))
- When a search engine returns 30 pages only 20 of which were relevant while failing to return 40 additional relevant pages, its precision is 20/30 = 2/3 while its recall is 20/60 = 1/3. So, in this case, precision is "how useful the search results are", and recall is "how complete the results are"



Bias and variance of a statistical estimate

- **Consider the problem of estimating a parameter α of an unknown distribution G**
  - To emphasize the fact that α concerns G we will refer to it as $\alpha(G)$
- **We collect N examples $X=\{x_1, x_2, \ldots, x_N\}$ from the distribution G**
  - These examples define a discrete distribution G' with mass 1/N at each of the examples
  - We compute the statistic $\alpha'=\alpha(G')$ as an estimator of $\alpha(G)$
    - In the context of this lecture, $\alpha(G')$ is the estimate of the true error rate for our classifier
- **How good is this estimator?**
- **The "goodness" of a statistical estimator is measured by**
  - BIAS: How much it deviates from the true value

$$\boxed{\text{Bias} = E_G[\alpha'(G)] - \alpha(G)} \qquad \text{where } E_G[X] = \int_{-\infty}^{+\infty} x\,g(x)\,dx$$

  - VARIANCE: How much variability it shows for different samples $X=\{x_1, x_2, \ldots, x_N\}$ of the population G

$$\boxed{\text{Var} = E_G\left[(\alpha' - E_G[\alpha'])^2\right]}$$

Bias and Variance

- An **estimator** is a rule for calculating an **estimate** of a given quantity based on observed data: thus the rule (the **estimator**), the quantity of interest (the estimand) and its result (the **estimate**) are distinguished.

- The **bias** error is an error from erroneous assumptions in the learning algorithm.
  - *Bias* is the simplifying assumptions made by the model to make the target function easier to approximate.
  - High **bias** can cause an algorithm to miss the relevant relations between features and target outputs (underfitting).

- The **variance** is an error from sensitivity to small fluctuations in the training set.