Friday, 4 July 2025

Machine learning notes

Cross validation

* Since there are many machine learning models, we might be confused which model to choose, here comes the cross valdation
* There are two steps involved
  + we need to estimate parameters for the model which is also known as training the data
  + Next we need to check how well the model performs
* Suppose we have a large block of data we can use 75% of data for training the model and the rest 25% of data to train the data. But we dont know which 75% of data to take.
* Cross validation takes each 75% data once and estimate the parameters and test the model on the rest 25%. and gives an average proformance of the model.

Confusion matrix

* A confusion matrix is a matrix that tells us what our algorithm did right and what it did wrong
* Example is a model that predicts whether a person has a heart disease or not. So a confusion matrix is such that the rows indicate the prdicted decision and colums represent true decison

|  | Person has heart disease | Person does not have hear disease |
| --- | --- | --- |
| Person has heart disease | 1 | FALSE POSITIVES |
| Person does not have heart disease | FALSE NEGATIVES | 1 |

Sensitivity and specificity

* After making the confusion matrix, we can find the sensitivity and specificity which are defined as follows
  + Sensitivity is defined as the percentage of true positives over true positivies and false negatives.
  + Specificity is defined as the percentage of true negatives over true negatives and fasle positives

Example

|  | Person has heart disease | Person does not have hear disease |
| --- | --- | --- |
| Person has heart disease | 139 | 20 |
| Person does not have heart disease | 32 | 112 |

Sensitivity calculation:

True positives = 139

false negatives = 32

therefore sensitivity = 139/(139 + 32) = 81.3%

Specificity calculation:

True negatives = 112

false positives = 20

therefore specificity = 112/(112 + 20) = 85%

Example for 3 x 3 confusion matrix

|  | Movie A | Movie B | Movie C |
| --- | --- | --- | --- |
| Movie A | 12 | 102 | 93 |
| Movie B | 112 | 23 | 77 |
| Movie C | 83 | 92 | 17 |

Sensitivity for movie A:

True positives = 12

False negatives = 112 + 83 = 195

sensitivity(movie A) = 12/(12 + 195) = 6%

specificity for movie A:  
True negatives = 23 + 77 + 92 + 17 = 209

False positives = 102 + 93 = 195

specificity (movie A) = 209/(209 + 195) = 51.73%