Metrics for comparison of Distributions

Ritik Dutta, IIT Gandhinagar

Progress

- Studied metrics which can be used for comparing distributions
- Implemented one metric (Maximum Mean Discrepancy)
- Generating datasets for populating the Github repository to be used for challenges

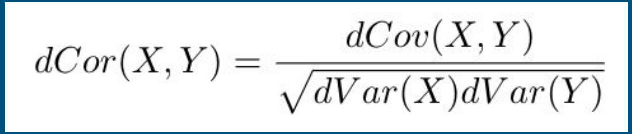
Metrics

- Using metrics (similarity distances/measures) to evaluate similarity between distributions

Kolmogorov-Smirnov Test

- Non-parametric test of the equality of continuous, one-dimensional probability distributions
- Tests such as the Anderson-Darling Test and the Cramer Von-Mises test, are considered to be refinements on this test

Distance Correlation

- Measure of dependence between two vectors of arbitrary, not necessarily equal dimensions
- Measures both linear and non-linear association between two random variables
- 

$$dCor(X,Y) = \frac{dCov(X,Y)}{\sqrt{dVar(X)dVar(Y)}}$$

Relief Divergence

- Used as feature selection method for binary class data
- Computationally efficient and noise-tolerant

Lp - distance

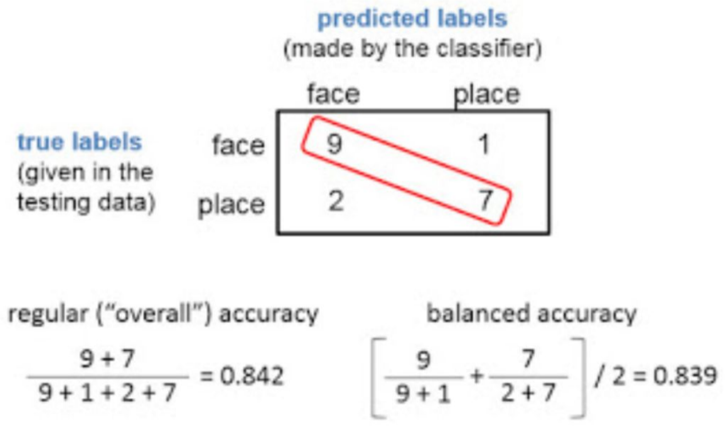
$$\|x\|_p = (|x_1|^p + |x_2|^p + \cdots + |x_n|^p)^{1/p}$$

- In high-dimensionality datasets fractional lp methods will be more effective than Euclidean distance
- An (better?) alternative is Mahalanobis distance

Balanced Accuracy Metric

$$\left(\frac{TP}{P} + \frac{TN}{N}\right) \times 0.5$$

- Datasets might not be balanced, i.e., number of instances in each class and each validation fold is not equal



predicted labels
(made by the classifier)

true labels
(given in the testing data)

	face	place
face	9	1
place	2	7

regular ("overall") accuracy

$$\frac{9+7}{9+1+2+7} = 0.842$$

balanced accuracy

$$\left[\frac{9}{9+1} + \frac{7}{2+7}\right] / 2 = 0.839$$

MMD

- Testing if distributions are different by drawing samples from them
- Find well behaved (smooth) function whose value is high for points from one distribution, low for the other
- Difference between the mean function values on the two samples

Dimension-wise prediction

- Measures how well the model captures the inter-dimensional relationships of the real samples
- Logistic regression used to predict feature values in test set
- Closer the performance of the model trained using synthetic set to the real one, more similar is the synthetic data

Principal Component Analysis

- The first two principal components of the two datasets can be plotted and checked if there are any significant differences w.r.t. the chosen principal components

Other methods

- Dimension-wise probability performance
- Correlation and covariance discrepancy
- Shape and sparsity
- Covariance matrices...