



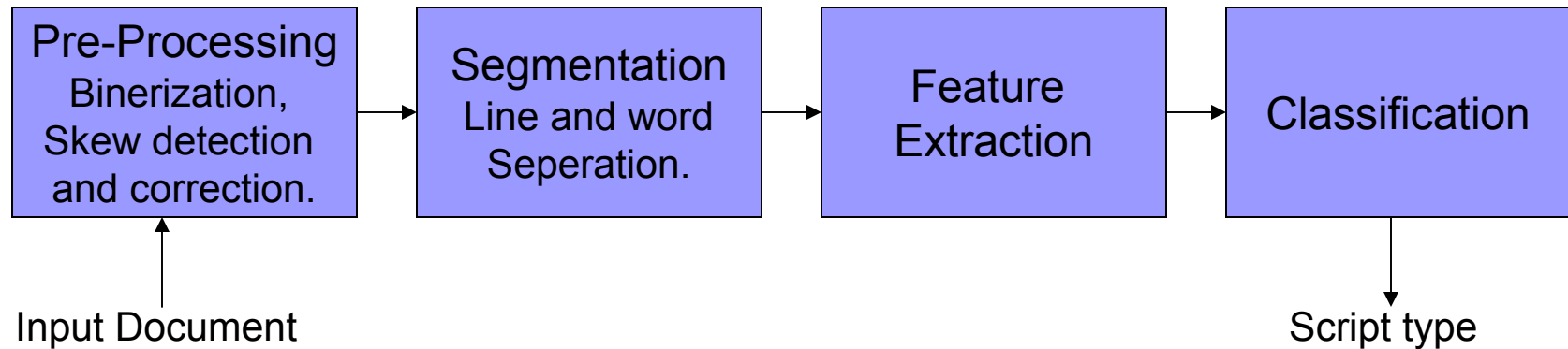
Automatic Script Identification



Why do we need Script Identification

- OCRs are generally language dependent.
- Document layout analysis is sometimes language dependent.
- For Indexing Documents.

Steps in Script Identification





Feature Extraction

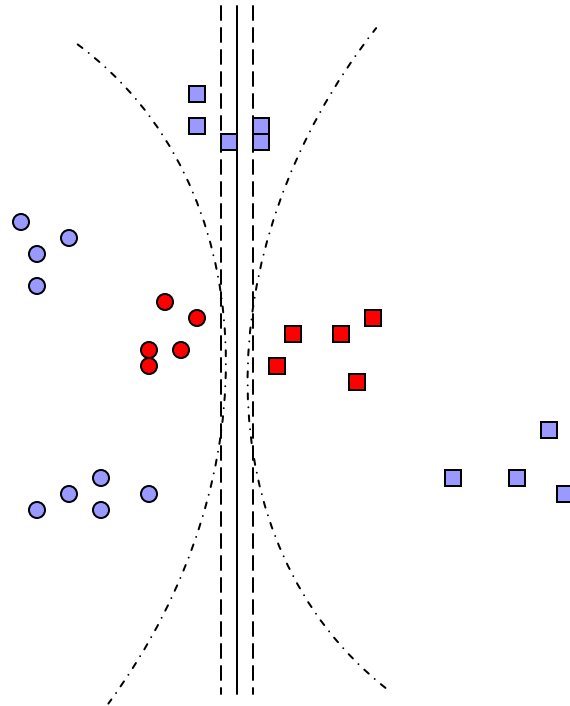
- Projection Profiles.
- Texture features using entropy, energy etc.
- Hough transform based features.
- Gabor filter based directional features.



Nature of Classification Problem

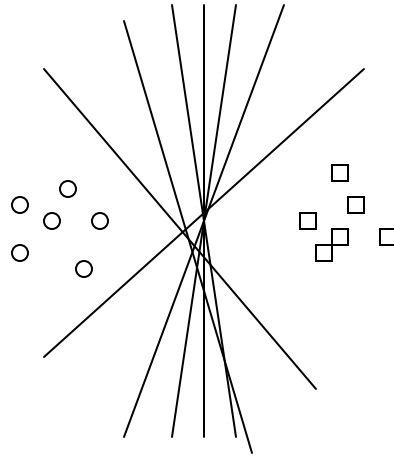
- Classifiers like KNN, MLP, SVM have been used.
- Training samples for a class is limited to a few fonts only.
- However in reality there are many fonts for the same script.
- Hence while the entire data is distributed over much larger space, the training is packed in much smaller region.

Nature of Classification Problem



Red for training data
Blue for remaining data

Nature of Classification Problem



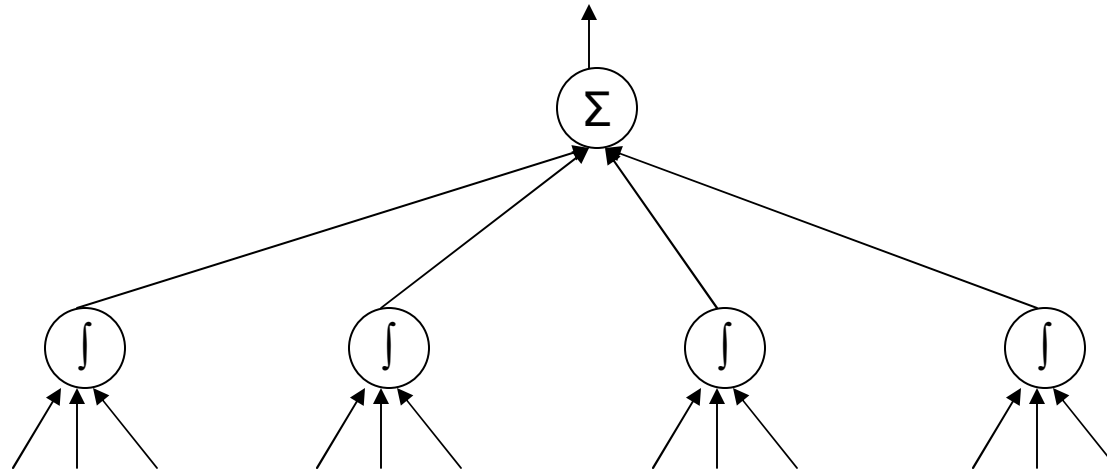
- Bayesian neural networks solve this by sampling from the distribution of weights and integrating over solutions
- But sampling is expensive and the method is time consuming.



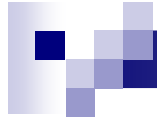
Expectation Maximization Neural Model

- To get non-linear probability fields we need to integrate over all weights that classify well and yet are different from each other.
- Modify cost function or error function in the neural network we need to optimize.
- $M(w_i) = E(w_i) + \cos^2(w_i, w_j)$ where $(j \neq i)$

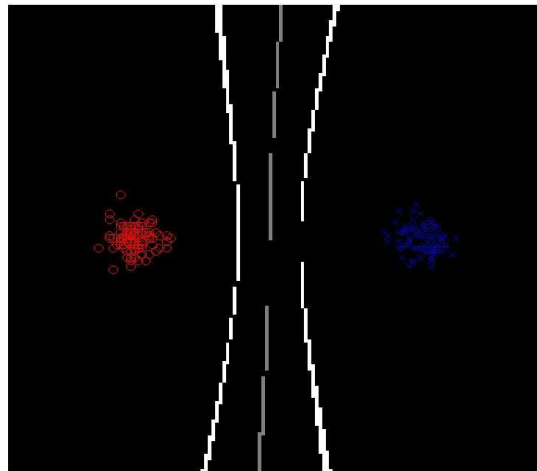
EM Neural Model



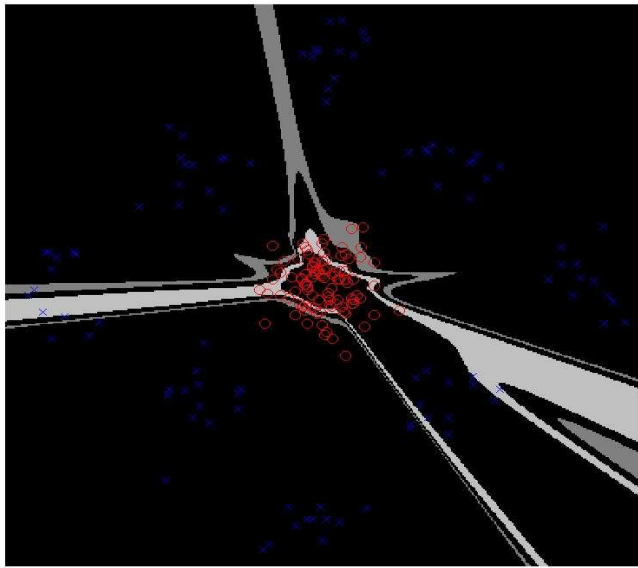
- Training using EM.
- E step is exact as we make prior equal to posterior
- M step through the modified Gradient descent
- Single layer of neurons shows non-linearity!



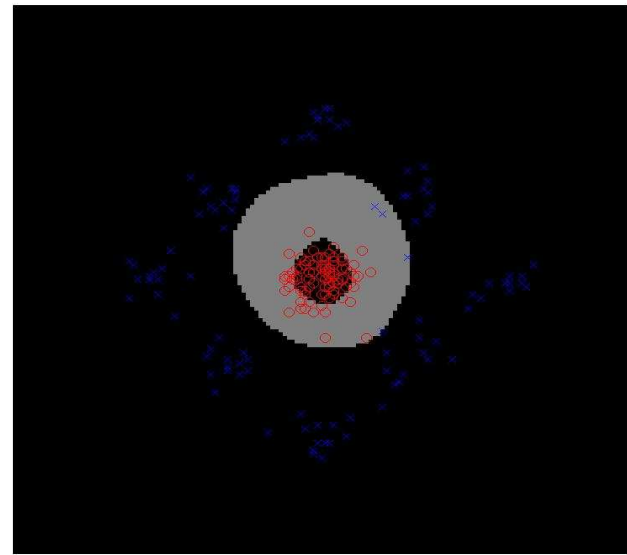
Some Experiments



Some Experiments



MLP



EMNM



References

1. A. Spitz, "Determination of The Script and Language Content of Document Images" IEEE Transactions on PAMI, vol 19, pp 235-245, 1997
2. J Ding, L Lam and C Y Suen, "Classification of Oriental and European Scripts by Using Characteristic Features", Proc. 4th ICDAR, pp 1023-1027, 1997
3. J. Hochberg, P Kelly, T Thomas and L Kerns, "Automatic Script Identification from Document Images using Cluster Based Templates." IEEE Transactions on PAMI, vol 19, pp 176-181, 1997.
4. T.N. Tan, "Rotation Invariant Texture Features and Their Use in Automatic Script Identification" IEEE Transactions on PAMI, vol 20, pp 751-756, 1998
5. J Hochberg, K Bowers, M Cannon and P Kelly, "Script and Language Identification for Handwritten Document Images", IJDAR, vol 2, pp 42-52, 1999
6. U Pal and B B Chaudhuri, " Automatic Identification of English, Chinese, Arabic, Devanagiri and Bangla Script Line", Proc. 6th ICDAR, 2001