Over-fitting vs. Under-fitting

COMP337/COMP527 - Data Mining and Visualisation

Procheta Sen

UNIVERSITY OF LIVERPOOL

Over-fitting vs. Under-fitting

- If a model M, trained on train data D_{train} performs well on D_{train}, but poorly on a separate test dataset D_{test}, then it is likely that M is **over-fitting** to D_{train}
- Typically you will see 90-99% accuracy on D_{train} and 40-60% accuracy on D_{test} in the case of binary classification on balanced (equal no. of positive and negative) datasets
- This is because M has more than required parameters that it can “fit” to D_{train} (too much flexibility), and it fits all of those on D_{train}, generalizing poorly to D_{test}
- **Under-fitting** is on the other hand the situation where you get poor performance on D_{train} because your model is not sufficiently “fitted” to the train data.

Solutions to Under-fitting

- Learning has not converged
  - Let the training proceed for more iterations
- Your feature space is too small/inadequate
  - Implement more/better features
- Your train data is bad/noisy/missing values
  - Cleanse/re-annotate train data
- Your algorithm is not training well
  - Select a different training algorithm

Solutions to Over-fitting

- Reduce the flexibility of your model
  - Regularisation
  - Remove features
- Early stopping
  - Premature termination of training to prevent parameter overfitting
- Training with more data
- Cross-validation