CIFAR-10 2-class classification

Data

- Data consists of a training dataset consisting of 2000 images, intersparsed between the *airplane* and *cat* class and a test dataset of the same size.
- The dimensions of the dataset are (2000, 10), 10 stands for the word to vec encoding of the descriptors for each image.
- 10 clusters of the SIFT features were taken and clustering was performed.
  - The 10-repr of the input represents the scaled count of the number of SIFT features per cluster.
  - This gives a homogeneous representation of the input irrespective of the number of SIFT features per image.

Models

- The data is trained using both `kNN` and `SVM` (*linear and gaussian kernel*).
- Standard python machine learning libraries have been used.
  - sklearn (For SVM, SVR, LinearSVC, LinearSVR)
  - numpy (data manipulation)
  - pandas (intermediate and long term storage)
  - h5py (dense effficient storage)
- None of the hyperparameters have been changed. Also, given that the `test_data` is just as big as the training, **and** that the input vector size is much less than the training examples, data is dense enough to prevent overfitting.

Results

Method	Score (%)
kNN	65.95
SVM	70.35
Linear SVM*	**71.20**
radius-NN	64.4