Lecture 11

-NLP is all about words, their arrangement and their meaning

-NLP is a branch of AI that tries to emulate/understand the way a human speaks to another human

-Lecture 11 is about classifiers for text classification

List of classifiers

- Distance-based classifiers (Eg. – Manhattan (L1-norm), Euclidean (L2-norm))
- Similarity-based classifiers (Eg. - cosine, jaccard)
- Decision tree
- Random forest of decision trees
- Neural networks
- Logistic regression
- Naïve Bayes (probabilistic)
- Support Vector Machines (SVM)

Text classification task

- Let **v** be the feature vector associated with training document d_1
- Let **w** be the feature vector associated with test document d_2
- Calculation:

  Compute the distance between the two feature vectors

  OR

  Compute the similarity between the two feature vectors
- Given lots of training documents and 1 test document -
- Check for minimum distance OR maximum similarity with the test document (CLOSEST MATCH indicates the CLASS of test document)

Similarity classifier

- Cosine similarity

$$\mathbf{a} \cdot \mathbf{b} = |\mathbf{a}||\mathbf{b}| \cos\theta$$

$$\frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{a}||\mathbf{b}|} = \cos\theta$$

$$\text{cosine}(\mathbf{v}, \mathbf{w}) = \frac{\mathbf{v} \cdot \mathbf{w}}{|\mathbf{v}||\mathbf{w}|} = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2} \sqrt{\sum_{i=1}^{N} w_i^2}}$$