AI INSTITUTE #AIISC
UNIVERSITY OF SOUTH CAROLINA

UNIVERSITY OF SOUTH CAROLINA

CSCE 771: Computer Processing of Natural Language
Lecture 10: Machine Learning Basics (Review)

PROF. BIPLAV SRIVASTAVA, AI INSTITUTE

20TH SEPTEMBER, 2022

Carolinian Creed: "I will practice personal and academic integrity."

Organization of Lecture 10

- Opening Segment
  - Recent NLP-related talks
  - Review of Lecture 9
  - Recent Talks
  - Announcements
- Main Lecture
- Concluding Segment
  - About Next Lecture – Lecture 11

Main Section
- ML– Supervised
- ML - Unsupervised
- Neural Networks

9	Sep 15 (Th)	Semantics
10	Sep 20 (Tu)	Review: Machine Learning for NLP, Evaluation – Metrics
11	Sep 22 (Th)	Prelim. for Language Model – Vector embeddings, CNN/ RNN
12	Sep 27 (Tu)	Guest Lecture – Dr. Amitava Das: Glove, Word2Vec, Transformer Review: Reasoning for NLP
13	Sep 29 (Th)	Representation: Ontology, Knowledge Graph, QUIZ
14	Oct 4 (Tu)	Representation: Embeddings, Language Models
15	Oct 6 (Th)	Entity extraction
16	Oct 11 (Tu)	Guest Lecture – Dr. Amitava Das: Using lang models to solve NLP tasks

Review of Lecture 9

- Semantics
  - Shallow: similarity, relatedness; frames
  - Propbank
  - Deep: AMR
  - ConceptNet
- Review projects

Discussions: Recent Talks

1. Sep 19, 2022: Distinguished Speaker Series Talk:
   NLP and Society - Sentiment Analysis, Mental Health Monitoring,
   By Prof. Pushpak Bhattacharya, IIT Bombay
   Video: https://youtu.be/d-deCau9ud4

2. Sep 16, 2022: Seminar in Advances in Computing,
   "**Can we ever trust our chatbots? Towards trustable collaborative assistants**",
   By Prof. Biplav Srivastava, UoSC

Main Lecture

Machine Learning – Insights from Data

- Descriptive analysis
  - Describe a past phenomenon
  - **Methods**: classification, clustering, dimensionality reduction, anomaly detection, neural methods
- Predictive analysis
  - Predict about a new situation
  - **Methods**: time-series, neural networks
- Prescriptive analysis
  - What an agent should do
  - **Methods**: simulation, reinforcement learning, reasoning
- New areas
  - Counterfactual analysis
  - Causal Inferencing
  - Scenario planning

Machine Learning – Label Based View

- Label available – Supervised Learning
  - Example: Classification
- Label unavailable – Unsupervised Learning
  - Example: Clustering

Common Textual Data Processing Steps for ML

- Input: strings / documents/ corpus
- Processing steps (task dependent / optional - *)
  - Parsing
  - Word pre-processing
    - Tokenization – getting tokens for processing
    - Normalization* - making into canonical form
    - Case folding* – handling cases
    - Lemmatization* – handling variants (shallow)
    - Stemming* – handling variants (deep)
  - Semantic parsing – representations for reasoning with meaning *
  - Embedding – creating vector representation*

Impt: Contextual Word Embeddings

- Words as discrete
- Words with distributional assumptions:
  - Context: given a word, its nearby words or sequences of words
  - Words used in similar ways are likely to have related meanings; i.e., words used in the same (similar) context have related meanings
    - No claim about meaning except relative similarity v/s dis-similarity of words

TF-IDF based Word Representation -1

- Given N documents

- **Term frequency (TF):** for term (word) t in document d
  = tf(t, d)

Variants to reduce bias due to document length

Variants of term frequency (tf) weight

weighting scheme	tf weight
binary	$0, 1$
raw count	$f_{t,d}$
term frequency	$f_{t,d} \Big/ \sum_{t' \in d} f_{t',d}$
log normalization	$\log(1 + f_{t,d})$
double normalization 0.5	$0.5 + 0.5 \cdot \frac{f_{t,d}}{\max_{\{t' \in d\}} f_{t',d}}$
double normalization K	$K + (1 - K)\frac{f_{t,d}}{\max_{\{t' \in d\}} f_{t',d}}$

Sources:
(a) sci-kit documentation
(b) Wikipedia: https://en.wikipedia.org/wiki/Tf%E2%80%93idf

TF-IDF based Word Representation -2

- Given N documents
- Term frequency (TF): for term (word) t in document d
  = tf(t, d)
- **Inverse document frequency IDF**(t)

  = log [N / **DF**(t)] + 1

DF(t) = **document frequency**, the number of documents in the document set that contain the term t.

- **TF-IDF**(t, d) = TF(t, d) * IDF(t),

Variants of inverse document frequency (idf) weight

weighting scheme	idf weight ($n_t = \|\{d \in D : t \in d\}\|$)
unary	1
inverse document frequency	$\log \frac{N}{n_t} = -\log \frac{n_t}{N}$
inverse document frequency smooth	$\log\left(\frac{N}{1+n_t}\right) + 1$
inverse document frequency max	$\log\left(\frac{\max_{\{t' \in d\}} n_{t'}}{1+n_t}\right)$
probabilistic inverse document frequency	$\log \frac{N-n_t}{n_t}$

Sources:
(a) sci-kit documentation
(b) Wikipedia: https://en.wikipedia.org/wiki/Tf%E2%80%93idf

TF-IDF Example Calculation

Github: https://github.com/biplav-s/course-nl-f22/blob/main/sample-code/l5-wordrepresent/Word%20Representations%20-%20Vectors.ipynb

Classification

Classifier Method Types

- Individual methods
  - Decision Tree
  - Naïve Bayes
- Ensemble
  - Bagging: Aggregate classifiers ("bootstrap aggregation" => bagging)
    - Random Forest
    - Samples are chosen with replacement (bootstrapping), and combined (aggregated) by taking their average
  - Gradient Boosting: aggregate to turn weak learners into strong learners
    - Boosters (aggregators) turn weak learners into strong learners by focusing on where the individual weak models (decision trees, linear regressors) went wrong
    - Gradient Boosting
    - XGBoost: "eXtreme Gradient Boosting."

Source:

- Data Mining: Concepts and Techniques, by Jiawei Han and Micheline Kamber
- https://towardsdatascience.com/getting-started-with-xgboost-in-scikit-learn-f69f5f470a97

ML - Supervised

- By Example:
  - https://github.com/biplav-s/course-nl/blob/master/l9-ml-review/Classification%20-%20Fake%20news.ipynb

- Fake news dataset

Metrics: Accuracy, Precision, Recall

		Predicted class	
		Class = Yes	Class = No
Actual Class	Class = Yes	True Positive	False Negative
	Class = No	False Positive	True Negative

Accuracy =
(TP+TN)/
(TP+FP+FN+TN)

Precision =
(TP)/
(TP+FP)

Recall =
(TP)/
(TP+FN)

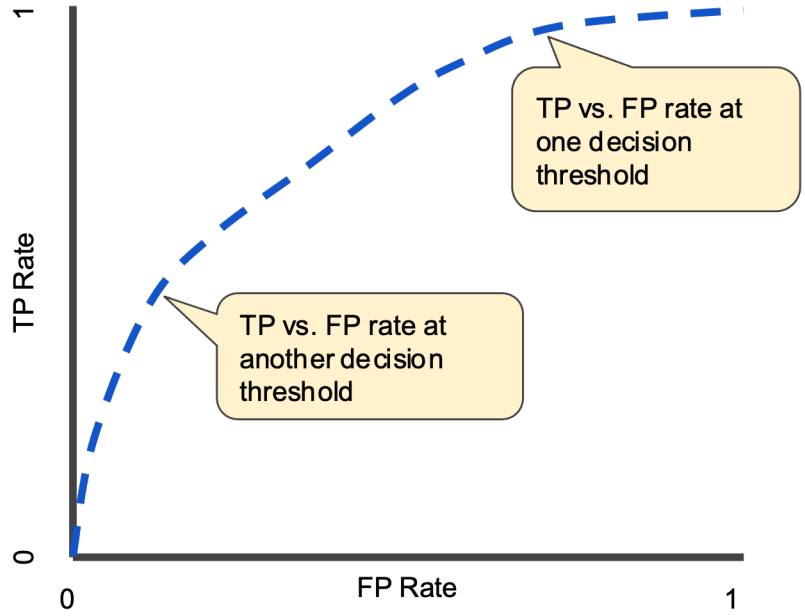
F1 Score: *Harmonic Mean*

1/F1 = 1/Precision + 1/Recall

F1 = 2*(Recall * Precision) /
(Recall + Precision)

ROC – Receiver Operating Characteristic curve

An ROC curve plots TPR vs. FPR at different classification thresholds



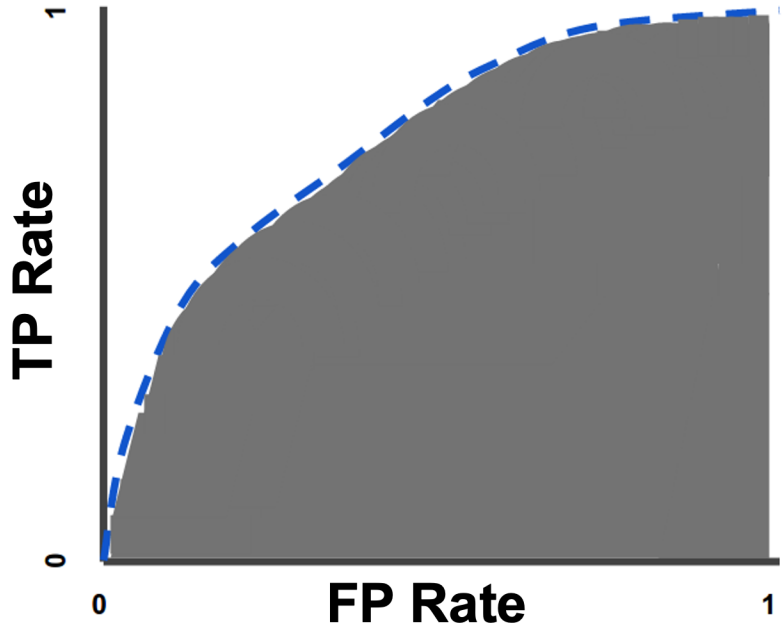
True Positive Rate = Recall = (TP)/ (TP+FN)

False Positive Rate = (FP)/ (FP+TN)

	Predicted class		
Actual Class		Class = Yes	Class = No
	Class = Yes	True Positive	False Negative
	Class = No	False Positive	True Negative

Source: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

AUC – Area Under the ROC Curve



- Aggregate measure of performance across all possible classification thresholds.
- Interpretation: probability that the model ranks a random positive example more highly than a random negative example

Not helpful when the **cost** of false negatives vs. false positives are asymmetric

Source: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

References

- Blogs: https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/
- Google: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

AutoML – Automated Machine Learning

- Automate data preparation
- Automate feature selection
- Hyperparameter tuning

- Demo:
  - IBM Auto AI - https://www.youtube.com/watch?v=ILCsbh9IKT0

Clustering

Unsupervised Machine Learning

- Group data into clusters/ classes without supervision
  - Limited supervision

- What is a good cluster ?
  - Samples within a cluster should be "**near**" to each other (**cohesiveness**)
  - Samples in a cluster should be "**far**" from other samples in other clusters. (**distinctiveness**)

Clustering for Data Understanding and Applications

- Biology: taxonomy of living things: kingdom, phylum, class, order, family, genus and species
- Information retrieval: document clustering
- Land use: Identification of areas of similar land use in an earth observation database
- Marketing: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
- City-planning: Identifying groups of houses according to their house type, value, and geographical location
- Earth-quake studies: Observed earth quake epicenters should be clustered along continent faults
- Climate: understanding earth climate, find patterns of atmospheric and ocean
- Economic Science: market resarch

Content: Jiawei Han, Micheline Kamber and Jian Pei Data Mining: Concepts and Techniques, 3rd ed.

Clustering as a Preprocessing Tool (Utility)

- Summarization:
  - Preprocessing for regression, PCA, classification, and association analysis
- Compression:
  - Image processing: vector quantization
- Finding K-nearest Neighbors
  - Localizing search to one or a small number of clusters
- Outlier detection
  - Outliers are often viewed as those “far away” from any cluster

Content: Jiawei Han, Micheline Kamber and Jian Pei Data Mining: Concepts and Techniques, 3rd ed.

Considerations for a Clustering Algorithm

- Need a distance measure for ***far*** and ***near***
- Be able to explain what a cluster means
- Handle different types of attributes: numeric, categorical (nominal, ordinal), binary
- Detect different shapes of clusters
- Handle noisy data
- Scale
  - Size
  - Dimensions

Major Clustering Approaches (I)

Partitioning approach:

- Construct various partitions and then evaluate them by some criterion, e.g., minimizing the sum of square errors
- Typical methods: **k-means**, k-medoids, CLARANS

Hierarchical approach:

- Create a hierarchical decomposition of the set of data (or objects) using some criterion
- Typical methods: Diana, Agnes, **BIRCH**, CAMELEON

Density-based approach:

- Based on connectivity and density functions
- Typical methods: **DBSACN**, OPTICS, DenClue

Grid-based approach:

- based on a multiple-level granularity structure
- Typical methods: STING, WaveCluster, CLIQUE

Content: Jiawei Han, Micheline Kamber and Jian Pei Data Mining: Concepts and Techniques, 3rd ed.

Major Clustering Approaches (II)

Model-based:

- A model is hypothesized for each of the clusters and tries to find the best fit of that model to each other
- Typical methods: **EM**, SOM, COBWEB

Frequent pattern-based:

- Based on the analysis of frequent patterns
- Typical methods: p-Cluster

User-guided or constraint-based:

- Clustering by considering user-specified or application-specific constraints
- Typical methods: COD (obstacles), constrained clustering

Link-based clustering:

- Objects are often linked together in various ways
- Massive links can be used to cluster objects: **SimRank**, LinkClus

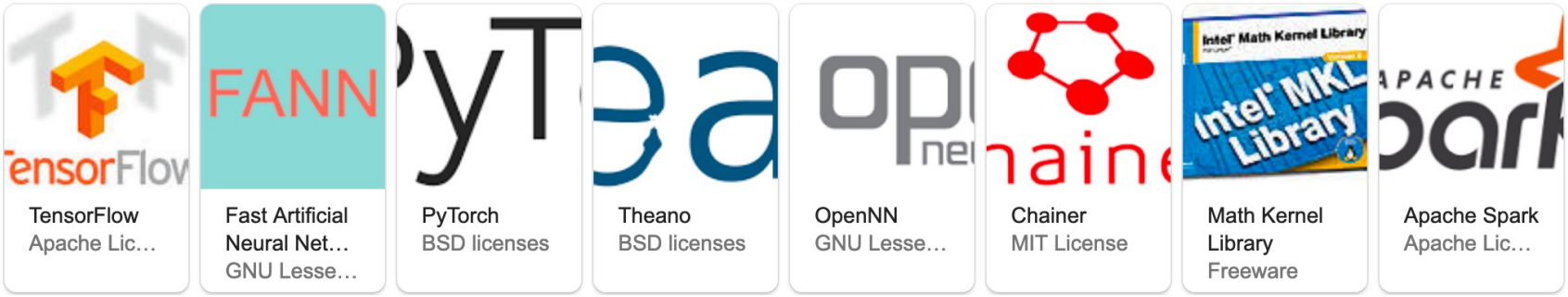
Content: Jiawei Han, Micheline Kamber and Jian Pei Data Mining: Concepts and Techniques, 3rd ed.

ML - UnSupervised

- Example 1
  - https://github.com/biplav-s/course-nl-f22/tree/main/sample-code/l10-ml-review
  - Two mysterious datasets

- Example 2
  - https://github.com/biplav-s/course-nl/blob/master/l9-ml-review/Clustering%20-%20Fake%20news%20Illustration.ipynb
  - Data
    - Bank agreement dataset
    - Fake news dataset

Neural Network Methods



Logistic Regression in a Slide

Function estimate (linear)
W: weight, b: bias

$$f(X_j) = X_j W + b$$

Update Weight

$$W^* = W - \eta \frac{dL}{dW}$$

Error Term (mean squared error)

$$MSE = \frac{1}{n} \sum_{j=1}^{n} \left[f(X_{j\cdot}) - y_j \right]^2$$

Common Code Pattern

```
y = tf.matmul(x, W) + b
loss = tf.reduce_mean(tf.square(y - y_label))
```

Keras and TensorFlow

- By Example:
  - https://github.com/biplav-s/course-nl/blob/master/l9-ml-review/Basic%20TensorFlow%20and%20Keras.ipynb
- TensorFlow's NMIST tutorial
  - https://www.tensorflow.org/tutorials/quickstart/beginner
- More examples
  - Number Addition by sequence learning: https://keras.io/examples/nlp/addition_rnn/
  - AutoEncoder: https://machinelearningmastery.com/lstm-autoencoders/

Discussion

- Platforms/ Tools
  - Periodic changes
  - Interoperation of models
- Data
- Compute resources
  - Development and test
  - Production

Lecture 10: Lecture Summary

- We reviewed Machine Learning methods
- Data preparation is the key
- Watch out for evaluation
- ML is just a step, what happens to the model is also important

Concluding Segment

Reading

- Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2021. Deep Learning--based Text Classification: A Comprehensive Review. ACM Comput. Surv. 54, 3, Article 62 (April 2022), 40 pages. https://doi.org/10.1145/3439726

- Hang Li, Language Models: Past, Present, and Future, Communications of the ACM, July 2022, Vol. 65 No. 7, Pages 56-63 10.1145/3490443

About Next Lecture – Lecture 11

Lecture 11 Outline

- Preliminaries for Language Model and Embeddings
  - Contextual Word Embeddings
- CNN
- RNN
- AutoEncoders