Second Review

CS5154/6054

Yizong Cheng

12/1/2022

Exam 2

- 8:30-9:50am December 6th.
- Close-book exam (calculators allowed)
  - Students with last name initial K-M go to Swift 520
  - Others in Swift 500.
  - Leave an empty seat between you.

10/13: Information Theory

- IIR 13.5: Feature Selection
  - Now “relevant” and “non-relevant” becomes two classes.
  - Terms become “features” or dimensions in vector classification.
- Still, need to fill out counts in the 2x2 table.
  - Or, four intersections of two sets, N11, N10, N01, and N00.
  - Along with row sums and column sums N1x, Nx1, N0x, Nx0.
  - p_t = N11/Nx1 and n_t = N10/Nx0, c_t = log (N11/N01)/(N10/N00)
  - c_t = 0 iff N11 N00 = N01 N10 or feature t is neutral.
  - Quiz 13
- Chi2 and mutual information

10/18: The Bernoulli Model

- IIR 13.3 and 13.1: The text classification problem
  - Training set, learning method, the classifier
  - Probability ranking is classification.
- TrainBernoulliNB and ApplyBernoulliNB
  - What are the parameters? P(c) and P(t|c)
  - How are the parameters learned? N_{ct}, and smoothed P(t|c)
  - How are classification decisions made?
- Assignment 12: IR13A-C.py, Bernoulli with feature selection
- Quiz 14: N, N_c, N_{ct}, prior[c], and condprob[t][c] (P(c) and P(t|c))

10/20: Multinomial NB

- IIR 13.2 and 14.4: Linear classifier
  - For two classes, the log-count ratio
- TrainMultinomialNB and ApplyMultinomialNB
  - What are the parameters? P(c) and P(t|c)
  - How are the parameters learned? T_{ct}, and smoothed P(t|c)
  - How are classification decisions made?
- Assignment 13: IR14
- Quiz 15: T_{ct} and condprob[t][c]

10/25: Competing Classifiers

- IIR 14.2 and 14.3: Rocchio and kNN classifiers
  - Parameters to be learned and used
  - Sklearn: NearestCentrid and KNeighborsClassifier
  - Quiz 16: TrainRocchio and ApplyRocchio, Train-kNN and Apply-kNN
  - ApplyLinearClassifier w and b
  - Macroaveraging and microaveraging
- Other classifiers from sklearn:
  - Logist regression, support vector machine, random forests
  - Multilayer perceptron (MLPClassifier)
- Assignment 14: IR15

10/27: Multiclass Classification

- IIR 14.5 and 15.4:
- Confusion matrix
- Assignment 15: IR16B.py
- Quiz 17: hyperplane, non-separable cases on the hypercube

11/1: From Rocchio to K-Means

- IIR 16.1 and 16.4: flat clustering, k-means
- RSS
- K-means: TrainRocchio and ApplyRocchio
  - Initial centroids determine the final membership.
- IR18
- Quiz 18: RSS

11/3: Evaluation of Clustering

- IIR 16.3
- Purity
- Normalized mutual information
- Rand index, Quiz 19
- F5
- Confusion matrix and purity computation
- Assignment 16: IR19
- Symmetry
- Invariant to doubling?

11/10: Hierarchical Clustering

- IIR 17
- SimpleHAC, based on a similarity matrix
- Combination similarity, dendrogram
- Single-link and complete-link clustering
  - Chaining and outliers
  - Quiz 20
- Sklearn’s AgglomerativeClustering
  - average and ward linkages
  - plot_dendrogram
  - Assignment 17

11/15: Topic Modeling

- IIR 18
- Singular value decomposition of the term-document matrix
  - $C = U\Sigma V^T$
  - Sklearn TruncatedSVD
  - Intermediate dimensions as topics
  - Quiz 21
- Nonnegative matrix factorization (NMF)
  - $C \sim WH$
- Latent Dirichlet allocation (LDA)

11/17: Missing in IIR

- What is new?
- Question answering
  - Knowledge graph
  - Deductive knowledge and inductive knowledge
- Information retrieval is not just document retrieval
  - Term → concept
  - Reasoning and logic
- Quiz 22: still old topics: topics and k-means

11/22: Link Analysis

- IIR 21.3: HITS
- The hyperlink graph
  - The adjacency matrix A
- “Good nodes won’t point to bad nodes”
  - Simple iterative logic for labeling good and bad nodes.
- “A good authority page for a topic is pointed to by many good hubs for that topic”
  - Circular definition and iterative update
  - Scaling/normalization of the vectors a and h
  - Eigenvectors of AA^T and A^TA
  - Quiz 23

11/29: PageRank

- IIR 21.2: PageRank
  - Teleporting probability α
- Algorithm from the adjacency matrix A to the Markov chain P
  - 2. dividing each row by sum of row to turn A into a stochastic matrix
  - 3. multiplying the matrix by $1 - \alpha$
  - 4. adding α/N to the matrix to get stochastic matrix P
- Ergodic Markov chain
- Quiz 24: A to P