

ML India-Bangalore Meetup

ML India

- * A place, an effort to catalyze the Machine Learning ecosystem in India involving students, researchers, institutions & corporations(<http://ml-india.org/>)

- * Maintained by a group of Machine Learning and Data Science enthusiasts

Varun Aggarwal, Co-founder, Aspiring Minds

Shashank Srikanth, Researcher, Aspiring Minds

Bhanu Pratap, Researcher, Aspiring Minds

Harsh Nisar, Researcher, Aspiring Minds

Sritulasi Edpuganti(Bangalore chapter), Co-founder, Dolphino

Agenda

- * Intro to ML
- * Intro to ML Algorithms used in paper
- * Paper Highlights and Summary
- * Discussion

What is Machine Learning?

- * Herbert Simon (Turing award 1975, Nobel prize in Economics 1978) – “Learning is any process by which system improves performance from experience”
- * More formally, Tom M. Mitchell – “A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T , as measured by P , improves with experience E ”

What is Machine Learning?

- * Is disrupting AI bigtime...
- * Widely used in everyday life, without us even knowing it
- * Important to understand intuition – handy cross domain skill

Applications of ML

- * Self Driving Cars (Watch: https://www.ted.com/talks/chris_urmson_how_a_driverless_car_sees_the_road?)
- * Face Recognition (Ex: Facebook Autotagging)
- * Speech Recognition (Ex: Siri, Cortana)
- * Personal Recommendations (Ex: Netflix, Amazon,...)
- * Efficient Web Search (Ex: Google, Bing,...)
- * And many more.....

Supervised Learning

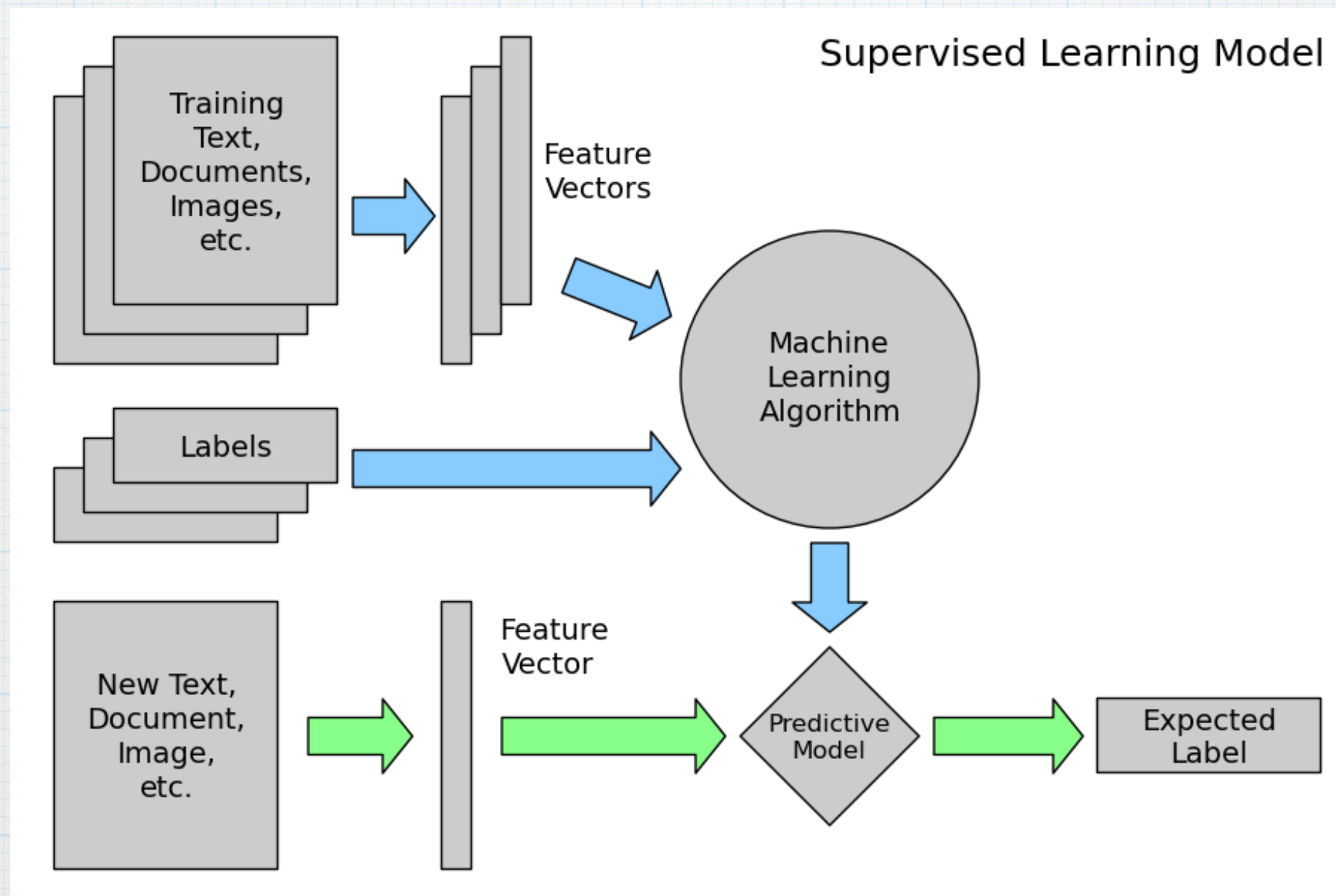


Image credit: Astroml blog

ML Jargon

- * Features/Input data
- * Labels/Observed values
- * Precision
- * Recall
- * F-measure
- * Cross Validation

Supervised Learning

One key research area in machine learning is to find the right features for a given problem.

Supervised Learning

Number of different models and learning techniques for these model

- * Linear Models
- * Ridge/LASSO
- * Polynomial Models
- * Neural Networks
- * Deep/convolutional networks
- * SVMs
- * Decision trees
- * Ensemble models
 - Boosting/bagging
 - ✦ Gradient boosted Decision Trees (more on this later)

Unsupervised Learning

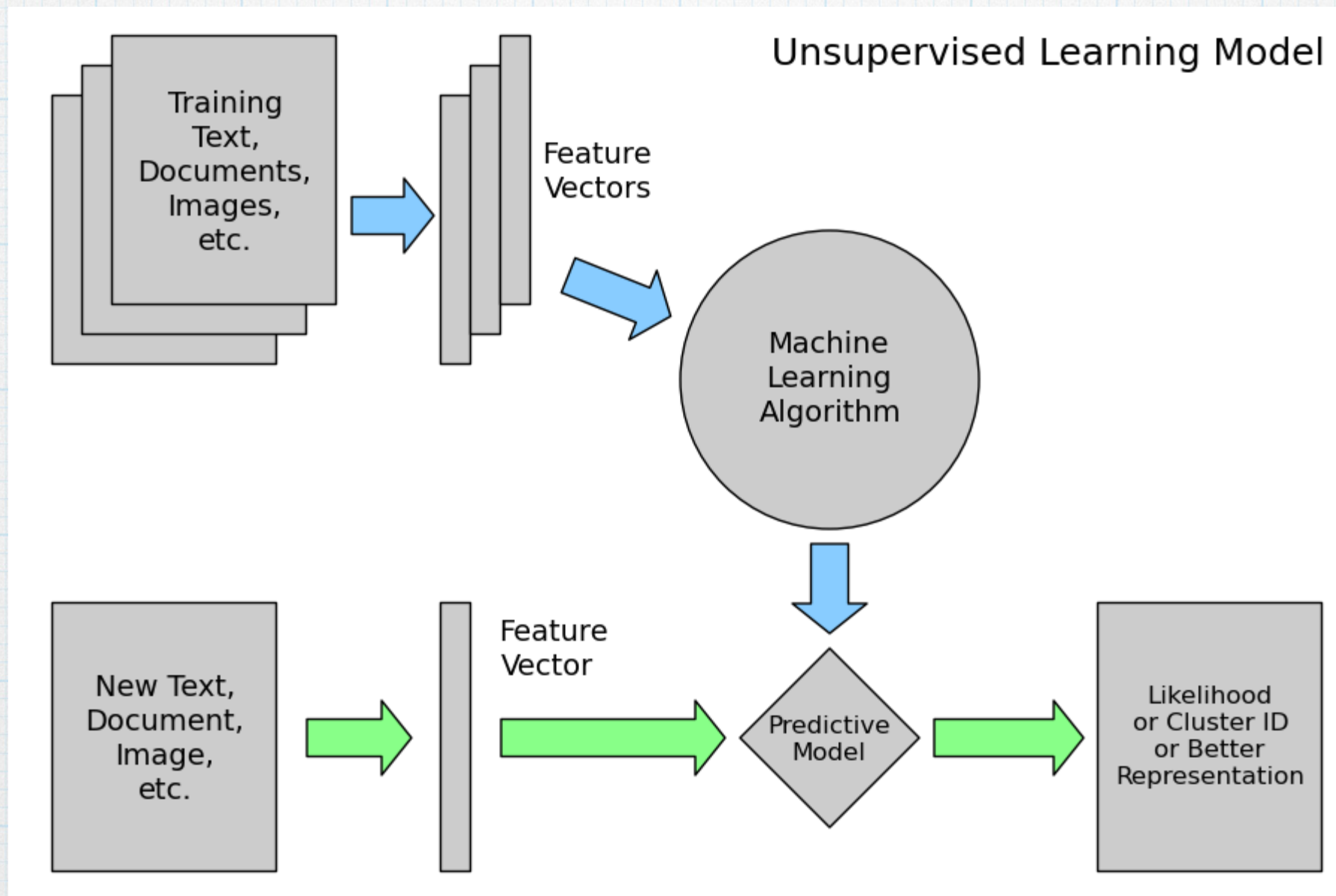


Image credit: Astroml blog

Unsupervised Learning

- * Clustering (K-means, hierarchal,...)
- * Singular value decomposition
- * Dimensionality Reduction
 - Principal Component Analysis(PCA)

Algo/models used in paper

- * Gradient Boosted Decision Tree (ML Algo)
- * Latent Dirichlet Allocation (LDA - a kind of topic modelling technique)

Gradient Boosted Decision Tree

Friedman 2001, Greedy function Approximation: A Gradient Boosting Machine

- * Effective, Off-the-shelf method for predictive models with high accuracy
- * Ensemble of weak learners , usually Decision Trees.
- * In gradient boosting, model assumes an additive expansion

$$F(x, \beta, \alpha) = \sum_{i=1}^n \beta_i h(x, \alpha_i)$$

(x,y)-Input labelled features

F(x) - Function to estimate

h- weak learners

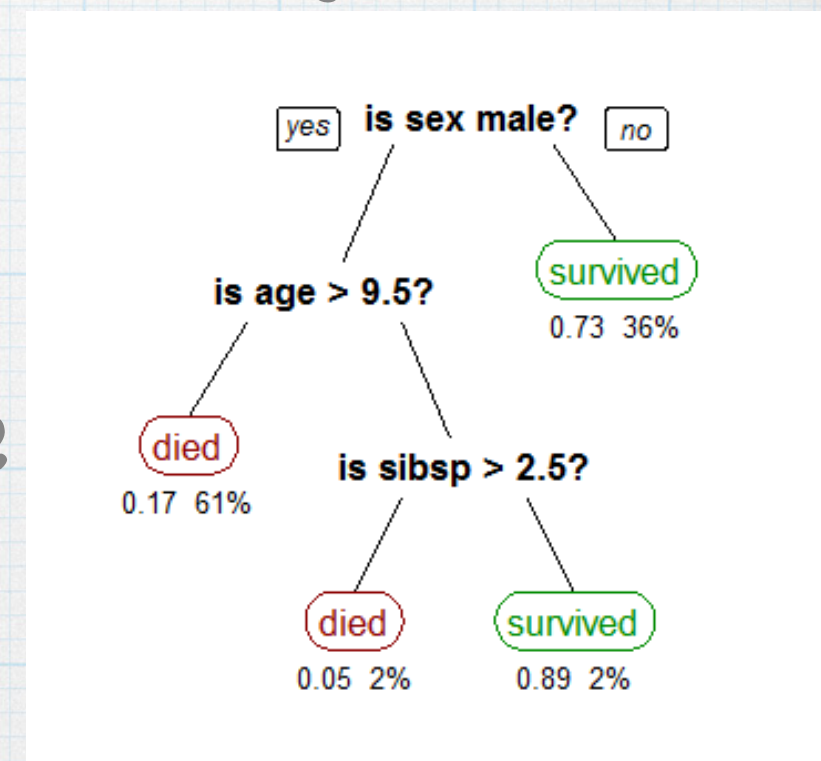
β - (To compute), the weight that a given classifier has in context of ensemble

Gradient Boosted Decision Tree

- * Intuitively, we iteratively build a sequence of predictors, and our final predictor is a weighted average of these predictors
- * At each step, we focus on adding an incremental classifier that improves the performance of the entire ensemble.
- * Good resources
 - <http://tullo.ch/articles/gradient-boosted-decision-trees-primer/>
 - <http://www.analyticsvidhya.com/blog/2015/09/complete-guide-boosting-methods/>

Gradient Boosted Decision Tree

- * Good resource to learn Decision Trees - <http://www.r2d3.us/visual-intro-to-machine-learning-part-1/>
- * A flow chart like tree structure
- * Internal node denotes test on attribute
- * Branch represent outcome of the test
- * Leaf nodes represent class labels or class distributions



Source:Wikipedia

Latent Dirichlet Allocation(LDA)

- * In context of paper - a variant of LDA is used to calculate linguistic features for GBDT.
- * It is a topic modeling technique
- * It is a way of automatically discovering topics from unstructured text.

Latent Dirichlet Allocation(LDA)

Example: Suppose we have following sentences

- I like to eat broccoli and bananas.
- I ate a banana and spinach smoothie for breakfast.
- Chinchillas and kittens are cute.
- My sister adopted a kitten yesterday.
- Look at this cute hamster munching on a piece of broccoli.

LDA will produce result like this

- Sentences 1 and 2: 100% Topic A
- Sentences 3 and 4: 100% Topic B
- Sentence 5: 60% Topic A, 40% Topic B
- Topic A: 30% broccoli, 15% bananas, 10% breakfast, 10% munching, ... (topicA - about food)
- Topic B: 20% chinchillas, 20% kittens, 20% cute, 15% hamster, ... (topicB - cute animals)

Source: Edwin chen's Blog (He is a great ML guy to follow)!


Twitter User Classification

(Marco Pennacchiotti, Ana-Maria Popescu)

Highlights

- * Problem to solve - Automatically infer user attributes like Political Affiliation, Ethnicity, Affinity to a business
- * Features used
 - Profile features (Name, location, bio,...)
 - Tweeting behavior (Avg no of tweets per day, number of replies,...)
 - Linguistic features (Prototypical words, Prototypical Hashtags, topic modeling, sentiment words)
 - Social network (replies, retweets,...)
- * Learning Algo used - Gradient Boosted Decision Trees
- * Results :
 - User Political Affiliation classified into Republicans/Democrats (Precision: 0.894 ± 0.007)
 - User Ethnicity classified into African-American/Not (Precision: 0.646 ± 0.017)
 - User classified for his affinity to starbucks (Precision: 0.763 ± 0.021)
- * Takeaways
 - Linguistic features (esp, topic-based) are proved consistently reliable across these classification tasks
 - Explicit social network features are valuable

Background

How many of you use  ?

Significant Improvements can be made to user experience by knowing demographic attributes, interests,..etc of a user

Recommendations can be given w.r.t which users to follow, what posts to read,...etc.

Related Work

- * Twitter user attribute detection. (Rao et al. 2010)
 - simple features like N-grams
 - Sociolinguistic features (presence of emoticons)
 - User network Statistics (No of followers)
 - Communication behavior(Retweet frequency)
- are used for user attribute detection.

Learning Algo & Features

* Learning Algo: Gradient Boosted Decision Trees (Friedman 2001)

* Profile features (PROF)

- Pilot study to access direct use of profile info (for gender and ethnicity classification tasks)
 - Corpus - 14M users
 - Technique - 30 regular expressions on Bio field
 - Results - could find ethnicity of <0.1% users, gender of 80% users. **Low Accuracy**
- Though profile fields don't contain good quality data, it can be used for bootstrapping training data. Some features derived are
 - length of username
 - use of avatar picture
 - no of followers...etc

Learning Algo & Features

- * Tweeting behavior features - useful for constructing model of user. 20 BEHAV features like
 - No of tweets
 - No & fraction of retweets
 - urls/tweet
 - avg no of #
 - avd time and std. between tweets ,.....

Learning Algo & Features

- * Linguistic features (features extracted from a tweet)

- Prototypical words/ Proto words(LING-WORD)

- classes can be described by these words

- young people - dude, lmao,...

- republicans - healthcare,...

- probabilistic model for automatically extracting proto words

- 2 Features derived:

- ◆ score based on no of proto words used by user

- ◆ score based on number of proto words belonging to a class- for that user

Learning Algo & Features

- Prototypical Hashtags (LING-HASH)

- Hypothesis: users from same class might like similar topics. Topics can be derived from Hashtags they used
- Features calculated similar to LING-Word after finding top hashtags used by users

- Generic LDA (LING-GLDA)

- Adaptation of original LDA; documents are replaced by users
- Users are represented as multinomial distribution of topics
- users = words of tweet
- how used in classification - Democrats higher prob of talking about social reforms, Republicans higher prob of talking about oil drilling
- General topics returned - soccer, music, politics

Learning Algo & Features

- Domain-specific LDA(LING-DLDA)
 - users from a specific domain like republicans, democrats
 - Domain specific topics returned - reforms, conservative approach
- Sentiment Words(LING-SENT)
 - Hypothesis: one class can have a majority opinion on a topic, which is different from other class
 - Ronald Reagan - Republicans(+), Democrats(-)
 - Manually collect set of terms for classes and find sentiment of user w.r.t those terms

Learning Algo & Features

* Social Network Features

- Friend accounts(SOC_FRIE) - Democrats following democrats
 - ◆ Find prototypical friend accounts (similar technique used for proto words)
 - ◆ For each porto-friend account, set value to 1 if user follows, 0 otherwise
- Prototypical replied (SOC-REP) and retweeted (SOC-RET)
 - ◆ Hypothesis : users belonging to same class, reply/retweet messages from specific accounts
 - ◆ Features calculated similarly as Ling-word/Ling-Hash

Experimentation Results

System	PREC	REC	F-MEAS
democrats-B 1	0.989	0.183	0.308
democrats-B2	0.735	0.896	0.808
democrats-FULL	0.894 [‡]	0.936^b	0.915^b
republicans-B 1	0.920	0.114	0.203
republicans-B2	0.702	0.430	0.533
republicans-FULL	0.878 [‡]	0.805^b	0.840^b
ethnicity-B 1	0.878	0.421	0.569
ethnicity-B2	0.579	0.633	0.604
ethnicity-FULL	0.646 [‡]	0.665^b	0.655^b
starbucks-B 1	0.817	0.019	0.038
starbucks-B2	0.747	0.723	0.735
starbucks-FULL	0.762	0.756^b	0.759^b

Overall classification results

Lets Discuss.....