Exploratory, Confirmatory, & Predictive Big Cancer Data Analytics

Ivo D. Dinov

Statistics Online Computational Resource
Health Behavior & Biological Sciences
Computational Medicine & Bioinformatics
Michigan Institute for Data Science

University of Michigan

http://SOCR.umich.edu

SCHOOL OF NURSING
STATISTICS ONLINE COMPUTATIONAL RESOURCE (SOCR)
UNIVERSITY OF MICHIGAN

Model-based (p-value) Statistical Inference

↓

Model-free Machine-Learning Clinical Decision Support (prediction reliability)

- ❑ Predicting univariate clinical outcomes (e.g., cancer staging)
- ❑ Processing unstructured clinical notes and medication data
- ❑ Generating machine-learning models of association

Dinov, Springer (2018)

Head and Neck (HnN) Cancer Dataset

Demographics (n=343)
GENDER
AGE_NOW
MARITAL_STATUS
RACE
DEATH_DATE
SMOKER_STATUS
ALCOHOL_STATUS
ILLICIT_DRUG_USE
CANCER_FAM_HX
DEMENTIA_FAM_HX
HYPERTEN_FAM_HX

Encounters (10,672)
PT_ID
CPI
VISIT_NUM
FINANCIAL_CLASS
SERVICE
VISIT_DATE
DISCHARGE_DATE
LOS_DAYS
LOS_HOURS
CHARGE_SUM
MSDRG_CD
MSDRG_DESC
ADMIT_TYPE
DISCH_DISP
ADMISSION_DX
ADMISSION_DX_DESC
SEER_STAGE

Outpatient Medications (2,815)
PT_ID
CPI
VISIT_NUM
ORDER_DATETIME
RX_ORDER_DESC
RX_ORDER_DOSE_PER_DAY
RX_ORDER_FREQ
RX_ORDER_TOTAL_DOSE_QTY
RX_TOTAL_DOSE_QTY
RX_STRENGTH_UNIT
MEDICATION_SUMMARY

Predicting univariate clinical outcomes (e.g., cancer staging)

- Naïve Bayes Classifier – predict Cancer State (early vs. late)
  - PID: coded patient ID
  - Seer_stage: SEER cancer stage (0=In situ, 1=Localized, 2=Regional by direct extension, 3=Regional to lymph nodes, 4=Regional (both codes 2 and 3), 5=Regional, NOS, 7= Distant metastases/systemic disease, 8=Not applicable, 9=Unstaged, unknown, or unspecified). See: **http://seer.cancer.gov/tools/ssm**
  - Y= 0(early) vs. 1 (late)

Seer_Stage	0	1	2	3	4	5	7	8	9
Proportion	0.03402	0.39886	0.071833	0.147448	0.069943	0.018903	0.124763	0.020794	0.113422

- X=Medication_summary: brief description about medication brand and usage

  hn_med_corpus[[1]]$content = "(Zantac) 150 mg tablet oral two times a day"
  hn_med_corpus[[2]]$content = "5,000 unit subcutaneous three times a day"
  hn_med_corpus[[3]]$content = "(Unasyn) 15 g IV every 6 hours"
  …

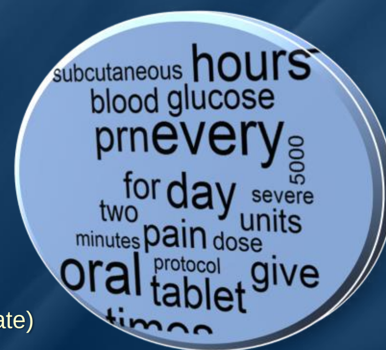
http://Predictive.Space

Predicting univariate clinical outcomes (e.g., cancer staging)

- Visual Analytics

Wordles

Y= 0(early) vs. 1 (late)



- Naïve Bayes Classifier – predict Cancer State (early vs. late)

http://Predictive.Space

Predicting univariate clinical outcomes (e.g., cancer staging)

- Naïve Bayes Classifier – predict Cancer State (early vs. late)

##	hn_med_test$stage		
## hn_test_pred	early_stage	later_stage	Row Total
## early_stage	91	35	126
## later_stage	2	5	7
## Column Total	93	40	133

Independent (out-of-bag) testing/validation, Laplace=15,
Accuracy 72% (*acc*=96/133)

Accuracy can be improved to 80% by model adjustment and by using alternative model-based and model-free classifiers

http://Predictive.Space

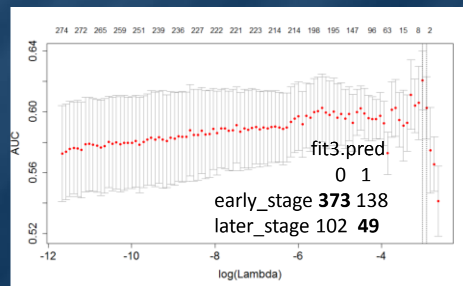
Processing unstructured clinical notes and medication data

`LASSO` estimates minimize a modified cost function

$$\min_{\beta \in \mathbb{R}^k} \left\{ \frac{1}{N} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\},$$

`Ridge` regression minimizes a similar objective function different norm):

$$\min_{\beta \in \mathbb{R}^k} \left\{ \frac{1}{N} \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2 \right\},$$




```
medCorpus<-VCorpus(VectorSource(cancer$MEDICATION_SUMMARY))
dtm.tfidf<-DocumentTermMatrix(medCorpus, control=list(weighting=weightTfIdf))
fit3 <- cv.glmnet(x=dtm.tfidf, y=dtm$stage, family = 'binomial',
        alpha = 1,      # LASSO penalty
        type.measure = "class", # interested in the area under ROC curve
        nfolds = 10,    # 10-fold cross-validation
        thresh = 1e-3, # high value is less accurate, but faster training
        maxit = 1e3     # lower number of iterations for faster training)
```

http://Predictive.Space

TF = ratio (a term's occurrences in a document)/(the number of occurrences of the most frequent word within the same document)
IDF = the inverse of the share of the documents in which the regarded term can be found

Generating Machine-Learning Models of Association

For item-sets X and Y, the `support` of an item-set measures how frequently it appears in the data:

$$support(X) = \frac{count(X)}{N},$$

where *N* is the total number of transactions in the database and *count(X)* is the number of observations (transactions) containing the item-set *X*. Of course, the union of item-sets is an item-set itself, i.e., if $Z = X, Y$, then

$$support(Z) = support(X, Y).$$

For a rule $X \rightarrow Y$, the `rule's confidence` measures the relative accuracy of the rule:

$$confidence(X \rightarrow Y) = \frac{support(X, Y)}{support(X)}$$

This measures the joint occurrence of *X* and *Y* over the *X* domain. If whenever *X* appears *Y* tends to be present too, we will have a high $confidence(X \rightarrow Y)$. The ranges of the support and confidence are $0 \leq support,\ confidence \leq 1$.

http://Predictive.Space

Generating Machine-Learning Models of Association

The `lift` column shows how much more likely one medicine is to be prescribed to a patient given another medicine is prescribed. It is obtained by the following formula:

$$lift(X \to Y) = \frac{confidence(X \to Y)}{support(Y)}$$

Note that $lift(X \to Y)$ is the same as $lift(Y \to X)$. The range of $lift$ is $[0, \infty)$ and higher $lift$ is better. We don't need to worry about the support, since we already set a threshold that the support must exceed.

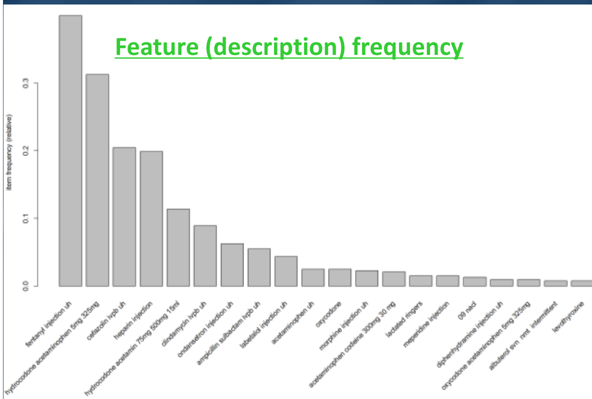
http://Predictive.Space

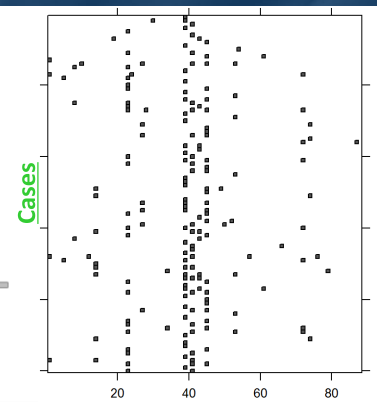
Generating Machine-Learning Models of Association

MEDICATION DESC.1	MEDICATION DESC.2	MEDICATION DESC.3	MEDICATION DESC.4	MEDICATION DESC.5
acetaminophen uh	cefazolin ivpb uh	NA	NA	NA
docusate	fioricet	heparin injection	ondansetron injection uh	simvastatin
hydrocodone acetaminophen 5mg 325mg	NA	NA	NA	NA
fentanyl injection uh	NA	NA	NA	NA
cefazolin ivpb uh	hydrocodone acetaminophen 5mg 325mg	NA	NA	NA

http://Predictive.Space

Generating Machine-Learning Models of Association - Meds

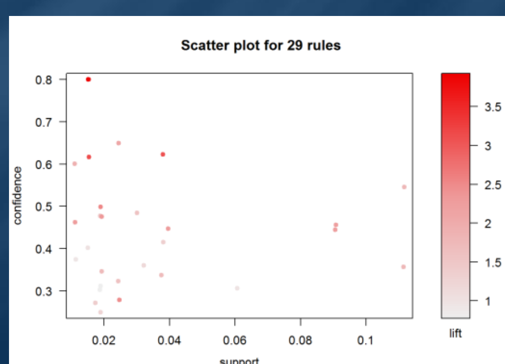




http://Predictive.Space

Generating Machine-Learning Association Mining

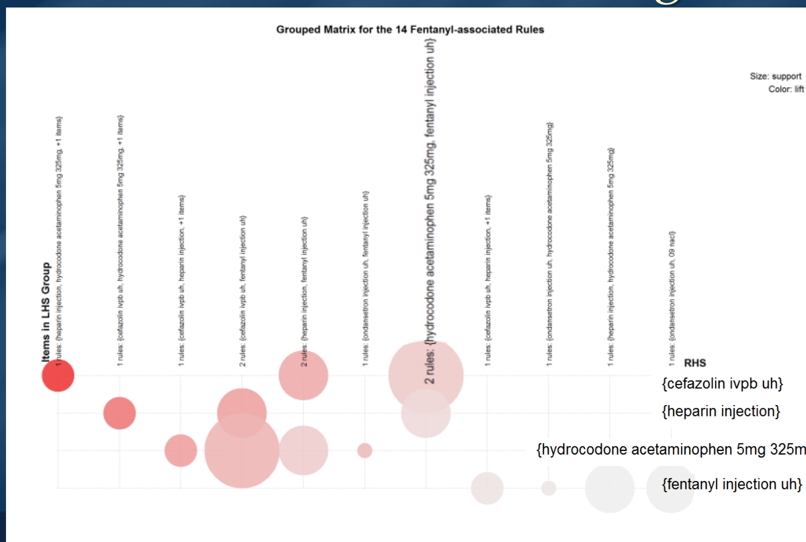
inspect(apriori_med_rule[1:3])




```
##        lhs                              rhs                    support    confidence  lift      count
## [1] {acetaminophen uh}             => {cefazolin ivpb uh}    0.01136364 0.4615385   2.256410  6
## [2] {ampicillin sulbactam ivpb uh} => {heparin injection}    0.01893939 0.3448276   1.733990  10
## [3] {ondansetron injection uh}     => {heparin injection}    0.01704545 0.2727273   1.371429  9
```

http://Predictive.Space

Generating Machine-Learning Association Mining



http://Predictive.Space

Acknowledgments

Funding

NIH: P20 NR015331, U54 EB020406, P50 NS091856, P30 DK089503, P30AG053760, UL1TR002240
NSF: 1734853, 1636840, 1416953, 0716055, 1023115
The Elsie Andresen Fiske Research Fund

http://SOCR.umich.edu

Collaborators

- SOCR: Alexandr Kalinin, Selvam Palanimalai, Syed Husain, Matt Leventhal, Ashwini Khare, Rami Elkest, Abhishek Chowdhury, Patrick Tan, Gary Chan, Andy Foglia, Pratyush Pati, Brian Zhang, Juana Sanchez, Dennis Pearl, Kyle Siegrist, Rob Gould, Jingshu Xu, Nellie Ponarul, Ming Tang, Asiyah Lin, Nicolas Christou, Hanbo Sun, Tuo Wang. Simeone Marino
- LONI/INI: Arthur Toga, Roger Woods, Jack Van Horn, Zhuowen Tu, Yonggang Shi, David Shattuck, Elizabeth Sowell, Katherine Narr, Anand Joshi, Shantanu Joshi, Paul Thompson, Luminita Vese, Stan Osher, Stefano Soatto, Seok Moon, Junning Li, Young Sung, Carl Kesselman, Fabio Macciardi, Federica Torri
- UMich MIDAS/MNORC/AD/PD Centers: Cathie Spino, Chuck Burant, Ben Hampstead, Stephen Goutman, Stephen Strobbe, Hiroko Dodge, Hank Paulson, Bill Dauer, Brian Athey