Homework 5

Nicholas Antonov & Patrick Grasso

December 18, 2016

Contents

1 Decision Trees

We tried two different methods of generating decision trees, the one using nltk gets a slightly higher accuracy. In both these versions we removed stop words and stemmed the other words, then split the text into two different sets; 50% training and 50% validation, to train our decision trees on. With sufficient tweaking and depths, they eventually got to a respectable accuracy, one that was better than just fitting the data with a yes or no response.

Some patterns that arose were words like "hospital", "er", and "sample" were consistently in the tree throughout different iterations of the program.

1.1 NLTK version

score	precision	recall
0.832909048831289	0.861	0.239

1.1.1 Output with tree pseudocode

```
python nltk-tree.py
Training on 8642 samples
Vectorizing with CountVectorizer
Vectorization complete. Classifying...
if er == False:
  if emergency == False:
    if hospital == False:
      if seen == False:
        if hep == False:
          if seizure == False:
            if infection == False:
              if cellulitis == False:
                if visit == False: return 'N'
                if visit == True:
                  if did == False: return 'Y'
                  if did == True: return 'N'
              if cellulitis == True:
                if reaction == False:
                  if contactable == False: return 'N'
                  if contactable == True: return 'N'
                if reaction == True:
                  if relevant medical == False: return 'N'
                  if relevant medical == True: return 'Y'
            if infection == True:
              if unknown == False:
                if history concomitant == False:
                  if patient received == False: return 'Y'
                  if patient received == True: return 'N'
                if history concomitant == True: return 'N'
              if unknown == True:
                if normal == False:
                  if route administration reported == False: return 'N'
                  if route administration reported == True: return 'Y'
                if normal == True: return 'Y'
          if seizure == True:
            if unspecified date == False:
              if male patient == False:
                if date outcome == False:
```

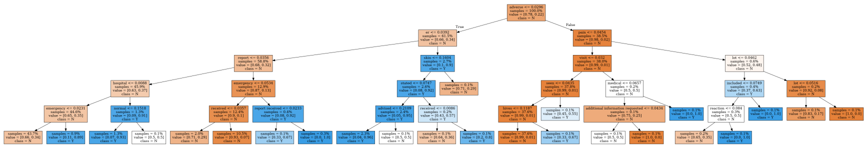
```
            if diagnosed == False: return ’N’
            if diagnosed == True: return ’N’
          if date outcome == True: return ’N’
        if male patient == True: return ’N’
      if unspecified date == True: return ’N’
  if hep == True:
    if stored == False:
      if developed == False:
        if 20 == False:
          if said == False:
            if allergy == False: return ’N’
            if allergy == True: return ’Y’
          if said == True: return ’Y’
        if 20 == True: return ’Y’
      if developed == True:
        if throat == False: return ’Y’
        if throat == True: return ’N’
    if stored == True:
      if events == False: return ’Y’
      if events == True: return ’N’
if seen == True:
  if initial == False:
    if states == False:
      if reporting == False:
        if 01 == False:
          if visit == False:
            if infection == False: return ’Y’
            if infection == True: return ’Y’
          if visit == True: return ’N’
        if 01 == True:
          if dates == False: return ’N’
          if dates == True: return ’Y’
      if reporting == True:
        if office == False: return ’N’
        if office == True: return ’Y’
    if states == True:
      if left == False: return ’N’
      if left == True: return ’Y’
  if initial == True:
    if review == False:
            if resolved == False: return ’N’
            if resolved == True: return ’Y’
          if review == True: return ’Y’
    if hospital == True:
      if stored == False:
        if respectively == False:
          if event reported == False:
            if exact == False:
              if pharmacist refers == False:
                if health professional == False:
                  if unknown reporter == False: return ’Y’
                  if unknown reporter == True: return ’N’
                if health professional == True: return ’N’
              if pharmacist refers == True: return ’N’
            if exact == True: return ’N’
          if event reported == True: return ’N’
        if respectively == True: return ’N’
      if stored == True: return ’N’
  if emergency == True:
    if reported adverse == False:
      if proquad == False:
        if flumist == False:
          if mmr == False:
            if feb 2017 == False:
              if nov 2014 == False: return ’Y’
              if nov 2014 == True: return ’N’
            if feb 2017 == True: return ’N’
          if mmr == True: return ’N’
        if flumist == True: return ’N’
      if proquad == True: return ’N’
    if reported adverse == True: return ’N’
if er == True:
  if warmth == False:
    if initial spontaneous == False:
      if adult == False:
        if dosage == False: return ’Y’
        if dosage == True: return ’N’
      if adult == True: return ’N’
    if initial spontaneous == True: return ’N’
  if warmth == True:
    if administration == False: return 'N'
    if administration == True: return 'Y'

score: 0.832909048831289
```

1.2 Scikit version

score	precision	recall
0.82608192548	0.795	0.361

1.2.1 Visual representation of tree


adverse <= 0.0296 samples = 100.0% value = [0.78, 0.22] class = N
True
False
er <= 0.0392 samples = 61.5% value = [0.66, 0.34] class = N
pain <= 0.0454 samples = 38.5% value = [0.98, 0.02] class = N
report <= 0.0356 samples = 58.8% value = [0.68, 0.32] class = N
skin <= 0.1604 samples = 2.7% value = [0.1, 0.9] class = Y
visit <= 0.032 samples = 38.0% value = [0.99, 0.01] class = N
lot <= 0.0462 samples = 0.6% value = [0.52, 0.48] class = N
hospital <= 0.0088 samples = 45.9% value = [0.63, 0.37] class = N
emergency <= 0.0534 samples = 12.9% value = [0.87, 0.13] class = N
stated <= 0.0747 samples = 2.6% value = [0.08, 0.92] class = Y
samples = 0.1% value = [0.71, 0.29] class = N
seen <= 0.0435 samples = 37.8% value = [0.99, 0.01] class = N
medical <= 0.0657 samples = 0.2% value = [0.5, 0.5] class = N
included <= 0.0749 samples = 0.4% value = [0.37, 0.63] class = Y
lot <= 0.0516 samples = 0.2% value = [0.92, 0.08] class = N
emergency <= 0.0231 samples = 44.6% value = [0.65, 0.35] class = N
normal <= 0.1518 samples = 1.3% value = [0.09, 0.91] class = Y
received <= 0.0357 samples = 12.4% value = [0.9, 0.1] class = N
report received <= 0.0233 samples = 0.4% value = [0.08, 0.92] class = Y
advised <= 0.2109 samples = 2.4% value = [0.05, 0.95] class = Y
received <= 0.0086 samples = 0.2% value = [0.43, 0.57] class = Y
hives <= 0.1187 samples = 37.6% value = [0.99, 0.01] class = N
samples = 0.1% value = [0.45, 0.55] class = Y
additional information requested <= 0.0434 samples = 0.1% value = [0.75, 0.25] class = N
samples = 0.1% value = [0.0, 1.0] class = Y
reaction <= 0.084 samples = 0.3% value = [0.5, 0.5] class = N
samples = 0.1% value = [0.0, 1.0] class = Y
samples = 0.1% value = [0.83, 0.17] class = N
samples = 0.1% value = [1.0, 0.0] class = N
samples = 43.7% value = [0.66, 0.34] class = N
samples = 0.9% value = [0.11, 0.89] class = Y
samples = 1.3% value = [0.07, 0.93] class = Y
samples = 0.1% value = [0.5, 0.5] class = N
samples = 2.0% value = [0.71, 0.29] class = N
samples = 10.5% value = [0.93, 0.07] class = N
samples = 0.1% value = [0.33, 0.67] class = Y
samples = 0.3% value = [0.0, 1.0] class = Y
samples = 2.3% value = [0.04, 0.96] class = Y
samples = 0.1% value = [0.5, 0.5] class = N
samples = 0.1% value = [0.64, 0.36] class = N
samples = 0.1% value = [0.2, 0.8] class = Y
samples = 37.6% value = [0.99, 0.01] class = N
samples = 0.1% value = [0.33, 0.67] class = Y
samples = 0.1% value = [0.5, 0.5] class = N
samples = 0.1% value = [1.0, 0.0] class = N
samples = 0.2% value = [0.65, 0.35] class = N
samples = 0.1% value = [0.0, 1.0] class = Y

1.2.2 Output

```
python decision-tree.py
Train : 8642 (50.00%)
Test  : 8642 (50.00%)
Removing stop words/stemming
Vectorizing with TfidfVectorizer
Vectorization complete. Classifying...
advers                              : 0.4311993467877191
er                                  : 0.16485453975969108
hospit                              : 0.12815002718169066
report                              : 0.09747767691990414
emerg                               : 0.057588328149484 75
test                                : 0.0402387990241915 36
visit                               : 0.013739181566336325
red                                 : 0.012506647351963378
follow inform                       : 0.007722328894573468
mother                              : 0.007646701964854062

score    : 0.82608192548
baseline : 0.7811849109

~ confusion ~
reference:
```

```
[['TN' 'FP']
 ['FN' 'TP']]

confusion [DecisionTreeClassifier]:
[[6667   84]
 [1419  472]]

confusion [DummyClassifier]:
[[6751    0]
 [1891    0]]
```

2 Text topics

score	precision	recall
0.81497338579	0.635	0.309

2.0.1 output with important features

```
python topic-node.py
Train : 8642 (50.00%)
Test  : 8642 (50.00%)
Removing stop words/stemming
Creating tf-idf models
Creating LSI model
Topics:
(0, ['report', 'dose', 'patient', 'medic', 'unknown', 'inform', 'date', 'temperatur',
(1, ['red', 'pain', 'inject', 'site', 'arm', 'swell', 'left', 'day', 'rash', 'pt'])
(2, ['2014.', 'number', 'fluvirin', 'batch', 'oct', 'case', 'initi', 'intramuscularli'
(3, ['excurs', 'temperatur', 'degre', 'hour', 'zostavax', 'minut', 'previou', 'event',
(4, ['allergy/drug', 'red', 'pqc', 'excurs', 'complaint', 'recombivax', 'hb', 'swell',
(5, ['inject', 'site', 'red', 'pain', 'swell', 'arm', 'shoulder', 'fever', 'hb', 'reco
(6, ['zostavax', 'compon', 'menveo', 'pharmacist', 'fluvirin', 'unit', 'arm', 'conjug'
(7, ['none', 'state', 'pain', 'zostavax', 'fluvirin', 'sender', 'held', 'document', 'a
(8, ['none', '67', 'servic', 'person', 'syring', 'pain', 'use', 'rash', 'health', 'hep
(9, ['rash', 'pain', 'syring', 'fever', '67', 'servic', 'person', 'arm', 'compon', 'us
(10, ['rash', 'pain', 'oral', 'rotateq', 'none', 'arm', 'temperatur', 'fluvirin', 'm.a
(11, ['merck', 'proquad', 'gardasil', 'varivax', '9', 'rotateq', '08-jun-2015', 'k0257
(12, ['gardasil', '9', 'pt', 'inject', 'pain', 'site', 'rotateq', 'oral', 'swollen', '
(13, ['rash', 'arm', 'fever', 'zostavax', 'inject', 'site', 'headach', 'gardasil', 'le
(14, ['rash', 'flumist', 'pt', 'pain', 'gardasil', 'inject', 'touch', 'site', 'none',
(15, [’zostavax’, ’rash’, ’pt’, ’flumist’, ’state’, ’2015.’, ’gardasil’, ’2015’, ’none
(16, [’hb’, ’recombivax’, ’qualiti’, ’pqc’, ’complaint’, ’involv’, ’product’, ’expir’,
(17, [’flumist’, ’unspecifi’, ’rash’, ’pt’, ’intranas’, ’quadrival’, ’none’, ’pain’, ’2
(18, [’unspecifi’, ’swollen’, ’pt’, ’touch’, ’flumist’, ’inject’, ’red’, ’fever’, ’site
(19, [’pt’, ’fever’, ’pertin’, ’inject’, ’flumist’, ’drug’, ’touch’, ’red’, ’rotateq’,
(20, [’swell’, ’varivax’, ’swollen’, ’pertin’, ’provid’, ’drug’, ’exact’, ’1024256’, ’2
(21, [’swell’, ’flumist’, ’expir’, ’worker’, ’healthcar’, ’swollen’, ’proquad’, ’fever
(22, [’swell’, ’flumist’, ’swollen’, ’inject’, ’hive’, ’jan’, ’site’, ’2015’, ’sore’,
(23, [’pt’, "’’", ’assist’, ’hive’, ’rash’, ’vaqta’, ’area’, ’locat’, ’anatom’, ’‘‘’])
(24, [’pt’, ’zostavax’, ’expir’, ’swell’, ’day’, ’worker’, ’expiri’, ’reaction’, ’healt
(25, [’hive’, ’itch’, ’zostavax’, ’warm’, ’rash’, ’pneumovax’, ’touch’, ’swollen’, ’23
(26, [’pneumovax’, ’zostavax’, ’23’, ’cm’, ’area’, ’erythema’, ’nurs’, ’assist’, ’swoll
(27, [’arm’, ’pain’, ’sore’, "’’", ’touch’, ’vaqta’, ’‘‘’, ’pt’, ’itch’, ’upper’])
(28, [’erythema’, ’pain’, ’vaqta’, ’arm’, ’expir’, ’upper’, ’cm’, ’reaction’, ’proquad
(29, ["’’", ’‘‘’, ’assist’, ’zostavax’, ’pneumovax’, ’healthcar’, ’worker’, ’state’, ’
(30, [’hive’, "’’", ’touch’, ’pain’, ’warm’, ’sore’, ’‘‘’, ’bexsero’, ’swell’, ’arm’])
(31, [’pneumovax’, ’hive’, ’23’, ’seizur’, ’swell’, ’minut’, ’area’, ’erythema’, ’bodi
(32, [’ii’, ’m-m-r’, ’itch’, ’unspecifi’, ’pneumovax’, ’varivax’, ’certifi’, ’symptom’
(33, [’sore’, ’varivax’, ’ii’, ’warm’, ’itch’, ’touch’, ’m-m-r’, ’rotateq’, ’reaction’
(34, [’itch’, ’area’, ’touch’, ’pneumovax’, ’warm’, ’23’, ’sore’, ’hive’, ’reaction’,
(35, [’degre’, ’bexsero’, "’’", ’‘‘’, ’fever’, ’fahrenheit’, ’23’, ’day’, ’expiri’, ’he
(36, [’bexsero’, ’unspecifi’, ’expect’, ’day’, ’attent’, ’sore’, ’rang’, ’sought’, ’unk
(37, [’sore’, ’swollen’, ’itch’, ’warm’, ’fever’, ’left’, ’day’, ’shot’, ’touch’, ’seiz
(38, [’itch’, ’hive’, "’’", ’varivax’, ’fever’, ’event’, ’ii’, ’administ’, ’m-m-r’, ’‘
(39, [’assist’, ’nurs’, ’bexsero’, ’worker’, ’regist’, ’certifi’, "’’", ’sore’, ’allerg
(40, [’physician’, ’pneumovax’, ’23’, ’varivax’, ’unspecifi’, ’08-jun-2015’, ’1030586’
(41, ["’’", ’minut’, ’sourc’, ’varivax’, ’‘‘’, ’therapi’, ’follow-up’, ’administ’, ’yea
(42, [’right’, ’2015.’, ’left’, ’oct’, ’state’, ’sore’, ’day’, ’pt’, ’swollen’, ’anaphy
(43, [’seizur’, ’itch’, ’febril’, ’touch’, ’itchi’, ’vomit’, ’given’, ’swollen’, ’warm
(44, [’right’, ’ii’, ’m-m-r’, ’shoulder’, ’multipl’, ’regard’, ’seizur’, ’dizzi’, ’2’,
(45, [’right’, ’sore’, ’arm’, ’shoulder’, ’vomit’, ’touch’, ’area’, ’administ’, ’delto
(46, [’left’, ’right’, ’sore’, ’administ’, "’’", ’‘‘’, ’day’, ’celsiu’, ’bexsero’, ’se
(47, [’right’, ’itch’, ’varivax’, ’left’, ’proquad’, ’anaphylaxi’, ’strength’, ’vaqta’
(48, [’given’, ’bexsero’, ’fever’, ’right’, ’ii’, ’shoulder’, ’vomit’, ’m-m-r’, ’dizzi
(49, [’fever’, ’f’, ’degre’, ’healthcar’, ’health’, ’celsiu’, ’expos’, ’public’, ’worke
(50, [’given’, ’right’, ’left’, ’seizur’, ’bodi’, ’itch’, ’ach’, ’vomit’, ’pt’, ’itchi
(51, [’given’, ’right’, ’shoulder’, ’day’, ’state’, ’shot’, ’fever’, ’erythema’, ’left
(52, [’area’, ’swollen’, ’vomit’, ’sore’, ’erythema’, ’bodi’, ’ach’, ’tender’, ’cellul
(53, [’vomit’, ’bodi’, ’ach’, ’seizur’, ’given’, ’diarrhea’, ’sore’, ’nausea’, ’swollen
(54, [’vomit’, ’shoulder’, ’bexsero’, ’pharmacist’, ’hb’, ’diarrhea’, ’schedul’, ’recon
(55, [’shoulder’, ’given’, ’unspecifi’, ’area’, ’erythema’, ’reaction’, ’local’, ’shot
(56, [’shoulder’, ’bexsero’, ’pain’, ’pregnanc’, ’hot’, ’sourc’, ’left’, ’expect’, ’fol
(57, [’left’, ’multipl’, ’vomit’, ’anaphylaxi’, ’event’, ’oct’, ’shot’, ’arm’, ’jan’,
(58, [’hot’, ’shoulder’, ’swollen’, ’pain’, ’right’, ’l’, ’administ’, ’fever’, ’pt’, ’2
(59, [’hot’, ’shoulder’, ’2012’, ’2012.’, ’pain’, ’left’, ’unspecifi’, ’swollen’, ’expe
(60, [’2014.’, ’minut’, ’2012’, ’healthcar’, ’worker’, ’strength’, ’ii’, ’m.a’, ’reacti
(61, [’itchi’, ’fever’, ’seizur’, ’strength’, ’anaphylaxi’, ’vomit’, ’bexsero’, ’state
(62, [’2015.’, ’hot’, ’compon’, ’oct’, ’men’, ’reactions/allergi’, ’licens’, ’conjug’,
(63, [’cellul’, ’hot’, ’anaphylaxi’, ’tender’, ’antibiot’, ’state’, ’l’, ’oct’, ’develo
(64, [’l’, ’red’, ’vomit’, ’swell’, ’bodi’, ’leg’, ’fever’, ’erythema’, ’lump’, ’swolle
(65, [’headach’, ’bodi’, ’itchi’, ’vomit’, ’sore’, ’cellul’, ’hot’, ’tender’, ’minut’,
(66, [’shot’, ’multipl’, ’hot’, ’leg’, ’event’, ’2012’, ’2014.’, ’certifi’, ’state’, ’2
(67, [’l’, ’local’, ’cellul’, ’erythema’, ’licens’, ’given’, ’cm’, ’event’, ’practic’,
(68, [’hot’, ’red’, ’cellul’, ’l’, ’erythema’, ’fever’, ’swollen’, ’15’, ’centigrad’,
(69, [’shot’, ’thigh’, ’local’, ’flu’, ’swollen’, ’state’, ’left’, ’ii’, ’pharmacist’,
(70, [’2012’, ’shot’, ’local’, ’left’, ’area’, ’pregnanc’, ’headach’, ’2012.’, ’around
(71, [’provid’, ’hot’, ’erythema’, ’physician’, ’event’, ’hb’, ’recombivax’, ’headach’
(72, [’itchi’, ’l’, ’erythema’, ’headach’, ’oct’, ’multipl’, ’anaphylaxi’, ’pharmacist
(73, [’headach’, ’cellul’, ’f’, ’l’, ’2014.’, ’weak’, ’itchi’, ’muscl’, ’hot’, ’ach’])
(74, [’day’, ’shot’, ’expir’, ’vaqta’, ’state’, ’l’, ’cm’, ’x’, ’administr’, ’expiri’]
(75, [’day’, ’cellul’, ’week’, ’hot’, ’develop’, ’l’, ’provid’, ’hand’, ’muscl’, ’delto
(76, [’state’, ’none’, ’tender’, ’cellul’, ’leg’, ’public’, ’month’, ’develop’, ’expiri
(77, [’itchi’, ’pregnanc’, ’tender’, ’mom’, ’thigh’, ’headach’, ’child’, ’month’, ’hand
(78, [’tender’, ’event’, ’lump’, ’state’, ’month’, ’pregnanc’, ’anaphylaxi’, ’thigh’,
(79, [’pregnanc’, ’multipl’, ’itchi’, ’expir’, ’regist’, ’sourc’, ’syncop’, ’rang’, ’ce
(80, [’local’, ’upper’, ’tingl’, ’headach’, ’numb’, ’thigh’, ’hand’, ’hot’, ’f’, ’inadv
(81, [’event’, ’minut’, ’headach’, ’expect’, ’sourc’, ’pharmacist’, ’inch’, ’l’, ’hard
(82, [’red’, ’itchi’, ’week’, ’shot’, ’call’, ’vaqta’, ’dizzi’, ’administ’, ’regist’,
(83, [’tender’, ’multipl’, ’cm’, ’x’, ’child’, ’fever’, ’itchi’, ’shingl’, ’anaphylaxi
(84, [’thigh’, ’rang’, ’itchi’, ’inch’, ’anaphylaxi’, ’2’, ’f’, ’relev’, ’pregnanc’, ’c
(85, [’l’, ’muscl’, ’leg’, ’tender’, ’administ’, ’multipl’, ’4’, ’anaphylaxi’, ’frequen
(86, [’upper’, ’3’, ’rang’, ’thigh’, ’around’, ’symptom’, ’physician’, ’extrem’, ’vaqta
(87, [’f’, ’deltoid’, ’tender’, ’2’, ’care’, ’2015’, ’2012’, ’2012.’, ’back’, ’itchi’]
(88, [’headach’, ’muscl’, ’local’, ’tender’, ’syncop’, ’event’, ’3’, ’dizzi’, ’bodi’,
(89, [’dizzi’, ’week’, ’fatigu’, ’upper’, ’1’, ’3’, ’extrem’, ’tender’, ’2’, ’chill’])
(90, [’leg’, ’administ’, ’symptom’, ’event’, ’muscl’, ’develop’, ’dizzi’, ’centigrad’,
(91, [’leg’, ’skin’, ’client’, ’inch’, ’shot’, ’tender’, ’9’, ’syncop’, ’cellul’, ’shin
(92, [’lump’, ’chest’, ’state’, ’leg’, ’hard’, ’upper’, ’inch’, ’itchi’, ’regist’, ’hot
(93, [’upper’, ’lump’, ’cellul’, ’around’, ’schedul’, ’symptom’, ’hard’, ’shingl’, ’ina
(94, [’lump’, ’hard’, ’muscl’, ’skin’, ’red’, ’pregnanc’, ’month’, ’swollen’, ’develop
(95, ['leg', 'deltoid', 'local', '1', 'skin', 'mg', 'muscl', 'expiri', 'expect', 'inco
(96, ['erythema', 'tender', 'skin', 'thigh', 'itchi', 'rang', 'elbow', 'deltoid', 'moti
(97, ['around', 'muscl', '3', 'develop', 'month', 'upper', 'could', 'inch', 'back', 'mo
(98, ['skin', 'around', 'dizzi', 'inch', '1', 'week', 'administr', 'lump', 'leg', 'extr
(99, ['local', 'face', 'back', 'rang', 'inch', 'tender', 'leg', 'pregnanc', 'diamet',
Training a RandomForestClassifier on the topic probability matrix

score    : 0.81497338579
baseline : 0.782226336496

~ confusion ~
reference:
[['TN' 'FP']
 ['FN' 'TP']]

confusion [RandomForestClassifier]:
[[6464  296]
 [1303  579]]

confusion [DummyClassifier]:
[[6760    0]
 [1882    0]]
```

3 Conclusion

In general, the decision tree is better, but the text topic method would be what you would most likely use in a real world case, as it is better at catching more of the truly serious cases.