Doctor Bayes

Predicting Diseases from Symptoms

Anthony Perez, Brandon Beckhardt, Leo Keselman

Introduction

According to recent OECD estimates, healthcare expenditures for developed countries account for 10% to 20% of their national economies. In order to alleviate the impact of healthcare spending on the the “last mile” of medicine, we built a system to predict a user’s illness simply based on a description of their systems.

Originally we’d hoped to develop a classification system on top of electronic medical records, clinician notes, and the ICD-10 disease classification system. We were unable to find publically available EMRs so instead we created Doctor Bayes, an illness classification system whose education comes only from reading about medicine on the web.

Data Sources

We collected data from three sources:

- Mayo Clinic
- Freebase
- Wikipedia

The Freebase and Mayo Clinic sources each contain a list of symptoms and a description for each disease. Our training data was made from combinations of the Freebase and Mayo Clinic data (see the results section). The Wikipedia data is more free form and as a result was used as testing data. We tested on either the sentences describing the symptoms of the disease or the main section of the article.

Word & Document Similarity

When working across 3 free text datasets, each has different ways of describing identical symptoms and diseases. We used Freebase’s incomplete alias lists to help tie our data, obtaining 486 diseases with information across all three. However, that’s far short of all the examples (Freebase: 12,595; Wikipedia: 2,772; Mayo: 1,247). We’re investigating word embeddings and document similarity methods to create connections between similar data in the cases where we lack annotations

We’ve implemented baseline latent semantic analysis across our dataset, along with training word2vec vectors on our combined dataset (whose t-SNE visualization is shown to the right)

Algorithms and Learning Models

There are three main sections in our pipeline: training data sources, feature extraction, and model selection. We have six ways to partition our data (symptoms and descriptions from each of our three data sources), six methods of manipulating our features (described below), and several models (four of which we present below).

Our primary algorithm was a Multinomial Naive Bayes with Laplace smoothing due to its success in NLP processing tasks. We couldn’t find a comprehensive list with occurrences of diseases for our priors, so our priors are based on the number of Google results associated with a given disease.

All of our feature manipulation is aimed at reducing the feature space because we identified this problem as having high variance (being prone to overfitting). We also see increasing the data we give our models reduces test error, which is expected with a high variance problem.

Results

Data Set Error Analysis

Using all of our feature manipulations with the given training data, these are the top five accuracy results.

	Free Base Symptoms	X		X	X	X	X	
	Mayo Symptoms		X	X	X	X	X	
	Free Base Descriptions				X		X	X
	Mayo Descriptions					X	X	X
Naïve Bayes	Wikipedia	34%	69%	72%	84%	82%	87%	79%
	Manual	58%	83%	83%	90%	88%	88%	78%
Cosine Similarity	Wikipedia	39%	77%	80%	90%	84%	90%	81%
	Manual	33%	73%	78%	90%	83%	85%	65%
Logistic Regression	Wikipedia	16%	38%	40%	49%	46%	55%	47%
	Manual	13%	30%	25%	33%	28%	33%	40%
Random Trees	Wikipedia	31%	43%	60%	71%	66%	67%	58%
	Manual	25%	0%	28%	33%	33%	38%	8%

Feature Manipulations Error Analysis

Using all of our data sources with the given feature manipulations, these are the top five accuracy results.

	TFIDF		X					X	X	X	X	X	X
	Stop			X				X		X	X	X	X
	Stem				X				X	X	X	X	X
	Document Frequency						X				X	X	X
	Remove NonAlphabetical											X	X
	Priors					X							X
Naïve Bayes	Wikipedia	40%	83%	64%	43%	37%	40%	87%	83%	87%	87%	87%	87%
	Manual	78%	85%	78%	78%	55%	78%	90%	85%	90%	90%	90%	88%
Cosine Similarity	Wikipedia	17%	88%	65%	25%	17%	17%	90%	90%	91%	91%	90%	90%
	Manual	30%	88%	45%	35%	30%	30%	88%	88%	85%	85%	85%	85%
Logistic Regression	Wikipedia	51%	75%	71%	57%	9%	51%	76%	75%	77%	77%	77%	55%
	Manual	23%	35%	23%	23%	13%	23%	38%	43%	43%	43%	43%	33%
Random Trees	Wikipedia	62%	67%	61%	63%	61%	62%	67%	66%	66%	70%	67%	69%
	Manual	23%	35%	15%	23%	25%	23%	30%	33%	30%	35%	35%	33%

Error on Training Set

Training on both Freebase and Mayo Clinic data and testing on the Freebase descriptions, these are our top five accuracy results.

	Basic	X	
	Feature Manipulation		X
Naïve Bayes	Training	99%	100%
Cosine Similarity	Training	99%	100%
Logistic Regression	Training	100%	94%
Random Trees	Training	100%	100%

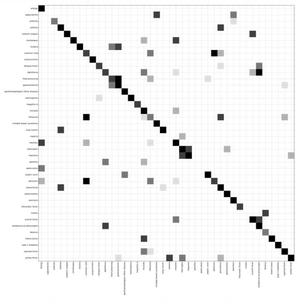
Performance Example

An example of our collected statistics and a single analyzed algorithm condition. This was with Naive Bayes and TF-IDF features.

Metric	Result
Top 1 Accuracy	65.02%
Top 5 Accuracy	87.45%
Mean Rank	8.45
Median Rank	1.0
Trimmed Mean Rank (⅔ of data)	1.38

Confusion Matrix

On our manual, handwritten dataset of illness descriptions (40 examples)



Examples Queries

>> I am having trouble remembering things. I repeat myself a lot.
- **1, Alzheimer's disease**
- 6.8e-06, Transient global amnesia
- 2.1e-06, Dyslexia
- 9.7e-07, Amnesia
- 5.8e-08, Social anxiety disorder

>> I shake a lot and have trouble balancing.
- **1, Parkinson's disease**
- 0.0001, Generalized anxiety disorder
- 4.2e-05, Primary aldosteronism
- 1.4e-05, Social anxiety disorder
- 5.6e-06, Alzheimer's disease

>> Redness, itchiness and difficulty opening eyes in the morning
- **1, Conjunctivitis (Pink Eye)**
- 6.6e-08, Glaucoma
- 7.7e-10, Dermatitis
- 1.6e-10, Blepharitis
- 8.3e-11, Sleepwalking

>> Fever headache and with a rash on my wrists.
- **1, Rocky Mountain spotted fever**
- 3.4e-05, Meningitis
- 2e-06, Lyme disease
- 1.8e-06, Rheumatic fever
- 1.4e-06, Dengue fever

>> Severe throat pain with difficulty swallowing and swollen tonsils.
- **1, Streptococcal pharyngitis** (Strep Throat)
- 2.9e-06, Diphtheria
- 4.5e-08, Infectious mononucleosis
- 1.9e-08, Laryngeal cancer
- 1.7e-09, Tularemia

>> It feels like there is liquid in my lungs. I am coughing a lot and I have a fever.
- 1, Gastroenteritis (Stomach Flu)
- 0.00074, Bronchitis
- **2.5e-05, Pneumonia**
- 1.9e-05, Influenza
- 1.2e-05, Tuberculosis

>>I'm sneezing, have a mild fever, and am very congested.
- 0.97, Pertussis (Whooping Cough)
- **0.027, Common cold**
- 0.00096, Influenza
- 3.6e-05, Pneumonia
- 4.2e-10, Egg allergy

>> I have a headache, fever, cough and fatigue
- 0.66, Common cold
- 0.19, Pneumonia
- **0.14, Influenza**
- 0.0041, Typhoid fever
- 0.0031, HIV/AIDS

>> I have to go to the bathroom a lot and my stomach hurts all the time.
- 1, Social anxiety disorder
- 0.0012, Generalized anxiety disorder
- 0.00021, Bulimia nervosa
- 0.00016, Gastroenteritis
- 4.7e-05, Osteoarthritis

Answer: **Irritable Bowel Syndrome**

>> I have a cut on my arm that seems to be infected. There is a red streak going up my arm.
- 0.84, Complex regional pain syndrome
- 0.079, Psoriasis
- 0.037, Antiphospholipid syndrome
- 0.03, Influenza
- 0.01, Scleroderma

Answer: **Sepsis**

Challenges and Next Steps

Lack of Clinical Data: None of our data sources exactly represent the two types of notes we’d like to learn from: actual clinician notes and people’s descriptions of their own illnesses. Training our system on data which better represents our desired queries would greatly strengthen the quality of the model.

Overconfidence: Our model tends to overfit, guessing either right or wrong with very high probability. We’d like to alleviate this by obtaining additional data and reducing the size of the feature space. Using only word referring to symptoms may be one solution.

Inconsistent Sources: Our sources refer to the same diseases and symptoms by different names and conventions. Part of our work is focused on finding matching diseases across data sources.

Quality Priors: It would be beneficial to access full data records to get the actual incidence of diseases as our current priors are unreliable.