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 "last mile" of medicine, we built a system to predict a user’s illness simply based on a description of their symptoms. 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 publicly 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: 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 with 3 free text datasets, each has different ways of describing identical symptoms and diseases. We used Freebase’s incomplete alias lists to help us in our data, containing 48 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 Naïve 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 Sources

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mayo Clinic</td>
<td>Contains a list of symptoms and a description for each disease.</td>
</tr>
<tr>
<td>Freebase</td>
<td>Contains a list of symptoms and a description for each disease.</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>More free form and used as testing data.</td>
</tr>
</tbody>
</table>

We used 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 with 3 free text datasets, each has different ways of describing identical symptoms and diseases. We used Freebase’s incomplete alias lists to help us in our data, containing 48 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 Naïve 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.

Examples Queries

- I am having trouble remembering things. I repeat myself a lot.
  - 1. Alzheimer’s disease
  - 2. Parkinson’s disease
  - 3. Psychosis
  - 4. Bulimia nervosa
  - 5. Alcoholism

- I feel it’s like there is liquid in my lungs.
  - 1. Pneumonia
  - 2. Bronchitis
  - 3. Common cold
  - 4. Asthma
  - 5. Allergies

- I have a headache, fever, cough and fatigue.
  - 1. Influenza
  - 2. Common cold
  - 3. Bronchitis
  - 4. Asthma
  - 5. Allergies

- I have a cut on my arm that seems to be infected.
  - 1. Staphylococcus aureus (Staphylococcal infection)
  - 2. Streptococcus pyogenes (Streptococcal infection)
  - 3. Neisseria gonorrhoeae (Gonorrhea)
  - 4. Chlamydia trachomatis (Chlamydia infection)
  - 5. Human papillomavirus (Human papillomavirus infection)

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 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.

Doctor Bayes: an illness classification system whose education comes only from reading about medicine on the web.