Capstone Project Report

Juanyan Li

February 13, 2017

1 DEFINITION

1.1 PROJECT OVERVIEW

Stack Exchange is a network of Q&A communities with different topics. The problem I try to solve here is a simplified version of the original Kaggle competition: Transfer Learning on Stack Exchange Tags. The initial competition aims to explore the possibility of transferring knowledge learned from multi-tagged Stack Exchange questions to a new field. More specifically, it attempts to predict the tags of questions in the field of physics by learning tagged questions from other fields: biology, cooking, cryptography, diy, robotics and travel.

The simplified problem that are being solved in this project will be described in the next subsection. It utilizes the dataset provided by the competition. The dataset originates from the Stack Exchange data dump, containing questions w/ or w/o tags from the aforementioned seven topics.

Next subsection will clarify the problem the project is trying to solve.

1.2 PROBLEM STATEMENT

Different from the original Kaggle competition, the problem here instead focuces on the prediction of a topic each question belongs to. The classifier will be designed to learn the information mentioned in the title as well as the question and correlates it with the topic the question. The input for the classifier would be the context in the question and its output would be one of the seven categories (biology, cooking, cryptography, diy, robotics, travel and physics).

In order to build such a classifier, one must first decide features the classifier uses as the input. Appropriate features should be able to correctly represent the knowledge or information that resides in the each question. A baseline approach for feature engineering is proposed by using term frequency-inverse document frequency (tf-idf) in the field of information retrieval. It provides a straightforward vector representation of text data. In comparison, I implemented another feature engineering model called Distributed Memory Model of Paragraph Vectors (PV-DM, also known as "doc2vec") - a neural network based unsupervised method that learns the vector representation of paragraphs.

Once the mathematical features for the training data is acquired, the next step would be choosing the classification model (classifier). In the project, three types of classification model would be considered: Logistic Regression, Naive Bayes Classifier and Support Vector Machine (SVM), since all of them are proved to be good classifiers in various tasks. By running cross-validation, it is hoped that the best combination of techniques (features and classifier) can be found.

Technical details for each methods shall be revealed in the Analysis section.

1.3 METRICS

To obtain a comprehensive understanding of the performance of each feature-classifier combination, precision, recall and F1-score will be used as the evaluation metrics.

Precision evaluates the classifier's ability to correctly label samples with their true tags. It is calculated as follows:

$$Precision = \frac{tp}{tp + fp} \tag{1.1}$$

where tp is the number of true positives and fp is the number of false positives.

Recall, on the other hand, evaluates the classifier's ability to find positive labels for all samples. It is calculated as follows:

$$Recall = \frac{tp}{tp + fn} \tag{1.2}$$

where fn is the number of false negatives.

F-score attempts to give an overall evaluation of the classifier by aggregating both precision and recall. In this case, F_1-score is used which weights precision and recall equally:

$$F_1 = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{1.3}$$

To extend the calculated scores to a multi-class setting, two approaches are adopted and evaluated separately:

i) The first approach calculates metrics for each label and then find out unweighted means as the overall scores.

ii) The second one also calculates the metrics for individual labels first but then weights them by the number of true instances for each label to come up with the final scores.

The next section will give concrete analysis on the problem itself.

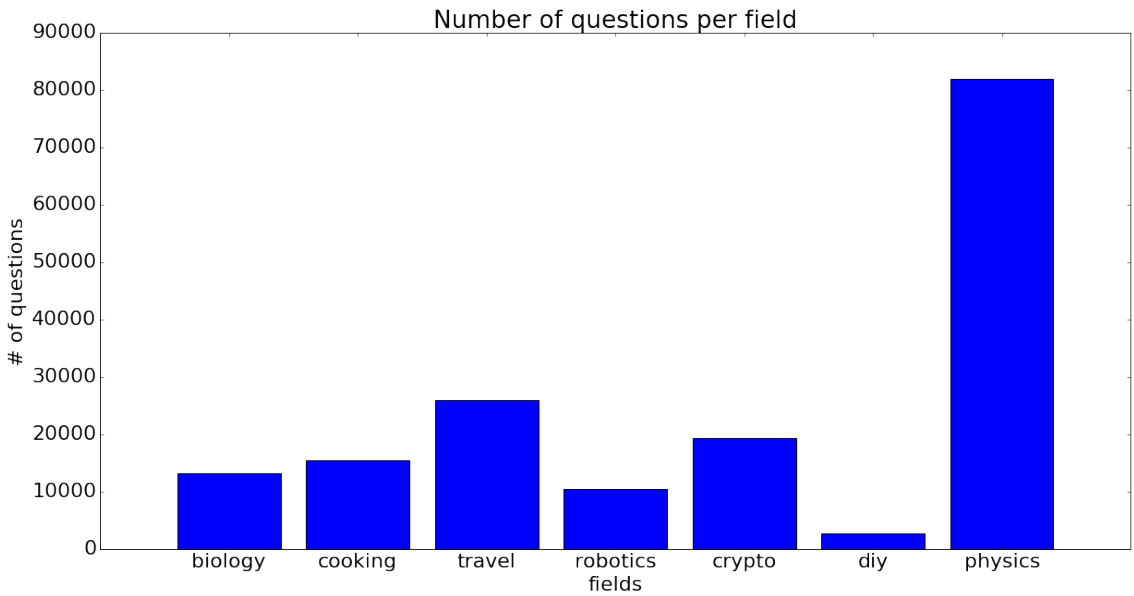
2 Analysis

In the first sub-section, a basic exploration of the dataset will be given, which mainly includes underlying statistics of the dataset. In the follow up section, methods for feature engineering (tf-idf and PV-DM) will be described in details. The last sub-section builds a baseline model using tf-idf as features and Gaussian Naive Bayes model as classifier. Results of the baseline model will be presented as well.

2.1 Data Exploration

As mentioned in the first section, the Stack Exchange dataset used in this project contains seven fields with a number of questions in each fields. See Figure 2.1. for the distribution of number of questions in each fields.

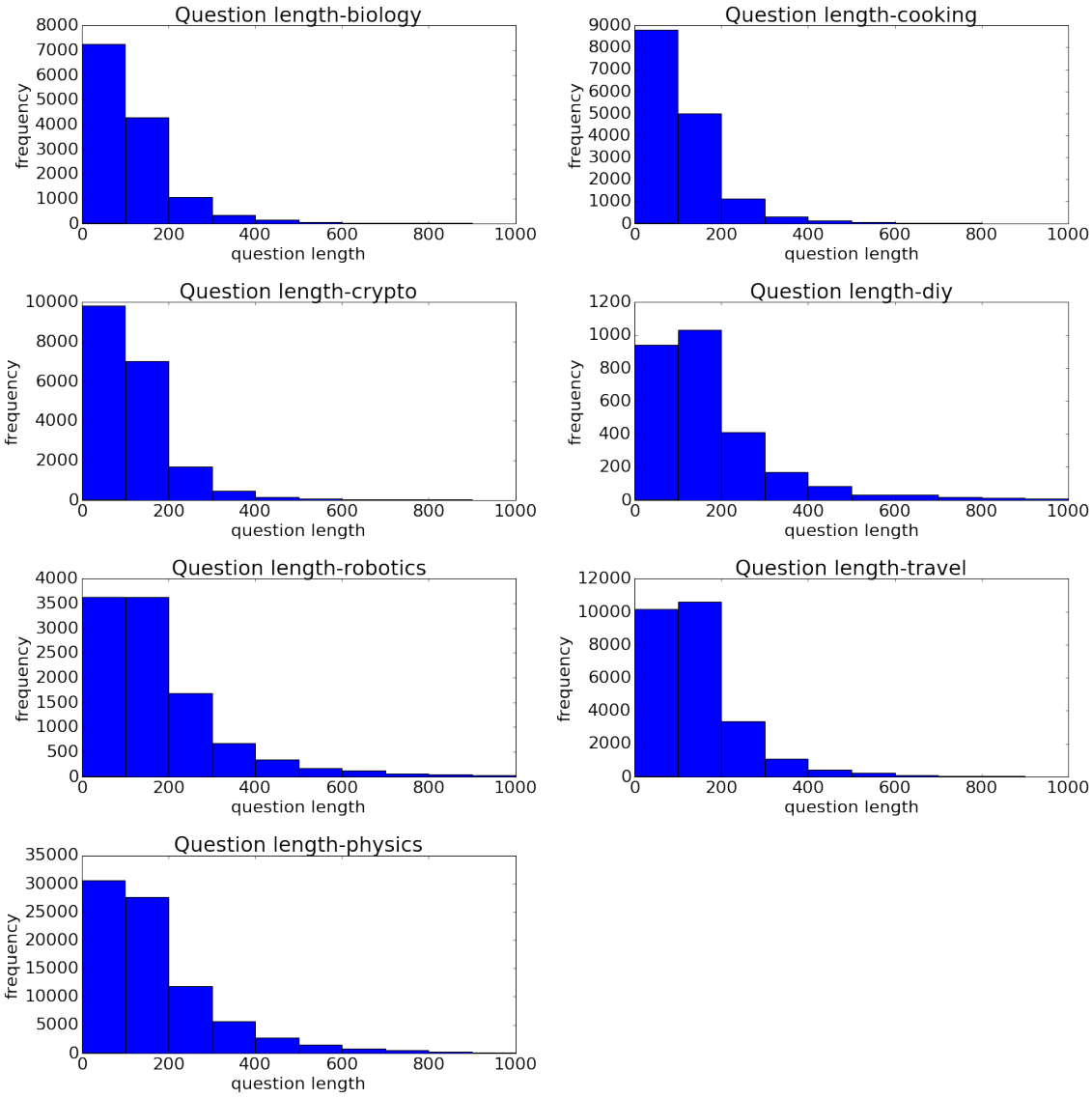
Figure 2.1:



Notice that the data heavily skews towards the physics field. A balanced dataset is preferred in order to avoid getting a biased classifier. This can be achieved by upsampling/downsampling in the training stage.

To take a further look into the dataset, a distribution of question length (number of tokens) is calculated based on each topic (See Figure 2.2). This gives an intuitive understanding of how much information is contained in each individual question.

Figure 2.2:



As can be seen above, most of the questions have less than 500 words/tokens. The figures shown above are capped to 1000 tokens. Actual distributions would have a much longer tail.

Those questions can be seen as outliers. Some of them are so because the author that raised the question pasted a large chunk of materials from other sources, which might not be useful. In the subsequent training stage, those outliers will be removed from the training samples.

2.2 Data Preprocessing

Before feature engineering, the data needs to be preprocessed first. General steps that are used in both TF-IDF and PV-DM are:

i) Data from all seven fields are loaded and combined into one dataframe in pandas;

ii) To simplified things further, all questions' titles and their contents are combined into one chunk of text;

iii) All html tags are removed from the raw combined content;

iv) Remove questions that have too many tokens (>1000).

For TF-IDF, which will be dicussed more in the next subsection, there is existing configurable functionality in scikit-learn. For "doc2vec" model, however, I implemented my own version on TensorFlow both for the purpose of learning and having more control on the training procedures.

2.2.1 Preprocessing for PV-DM

PV-DM essentially uses two large matrices to represent the embeddings of words as well as paragraphs. If the vocabulary for the words is of size V with embedding dimension d_w and the number of paragraphs in the training process is P with embedding dimension d_p, then the matrices' sizes are $V \times d_w$ and $P \times d_p$ respectively.

The first step to prepare inputs for PV-DM model is to build a vocabulary of words based on the training corpus. `UNKNOWN` and `NULL` symbols are also added for infrequent words and padding.

The next step is then to split the documents into sentences since I dont't want the sliding window to move across consecutive sentences. Each sentence belongs to a certain paragraph which is indexed with numbers for embeddings lookup during training stage.

After that, each sentence is then tokenized and transformed into a list of numbers based on the vocabulary built earlier. Sentences with too few or too much tokens are removed from the training set.

The last step happens during training where `NULL`s are padded at the begining of the sentence based on the configurable window size. Each batch data generated contains paragraph ids for each sentence as well as the encoded sentence itself. Batch labels returned are the ids of the next words the model tries to predict.

2.3 Feature Engineering

Let's discuss more about how to extract useful features from the corpus that could serve the purpose of classfying questions into corresponding topics.

2.3.1 TF-IDF

One common way to encode documents as numeric vectors is the vocabulary represnetation. Assume that the dataset has a vocabulary of size V in terms of unique words/tokens. Then for each document, it numeric representation is a vector of length V, which is usually a large number. For each word appearing in the document, the corresponding position in the vector is marked as the counts of that word, otherwise 0. Such representation clearly captures some of the characteristics of the document. However, it also tends to be noisy as many of the words appearing in a sentence are common words that can be easily seen from other sentences.

In order to reduce the noise level, tf-idf is preferred over simple term frequency. Instead, for each cell in the feature vector, term frequency or word counts is normalized by inverse document frequency (idf), hence the name. idf is calculated as follows:

$$idf(w, C) = \log \frac{N}{|\{d \in C; t \in d\}|} \tag{2.1}$$

where w is the word being calculated; C is the corpus data; N is the term frequency and d is each document in the corpus. Intuitively, the more the word appearing across documents, the less important it is in the final vector representation.

2.3.2 PV-DM

Also known as "doc2vec", the model is originally inspired by yet another famous embedding model called "word2vec" which translates words in a corpus into vector representations through neural netword based unsupervised training. The training mechanism is summarized in figure 2.3. Given a window consisting of consecutive words in a document, the input of the model is the concatenation/average of the document embedding and all the word embeddings within that window. The output of the model tries to predict the the next word outside the training window. Gradient descent method is used to optimized the loss function - usually calculated by negative sampling and cross entrophy.

In order to efficiently train the doc2vec model, a GPU implementation is conducted using TensorFlow the training is carried out using a GTX 1080 with 8GB memory. In this implementation, embeddings for paragraph/question is set to be 100 while word embeddings is set to be 50. A sliding window of size 4 is adopted and both paragraph embedding and word embeddings are concatenated to form the final input vector. The major concern here is that I wanted to preserve the sequential nature of the inputs. To do so and to avoid over-consumption of memory, I chose a smaller embedding size and also a smaller window size. The final model is trained using gradient based optimization with a batch size of 128.

Given the output logits x and the target t, the cross entropy for this single input is:

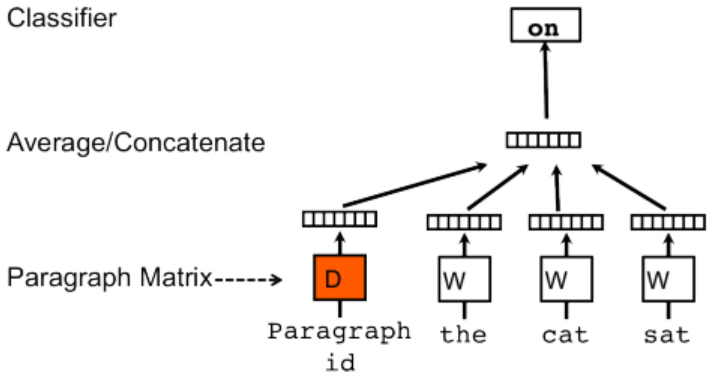
$$E(x,t) = -t \cdot \log\sigma(x) + (1-t)\log(1-\sigma(x)) \tag{2.2}$$

where $\sigma(\cdot)$ is the sigmoid function.

Since the size of the vocabulary is always a huge number, it is computationally expensive to calculate the cross entrophy against the whole vocabulary each time. Also, such computation will certainly obscure the importance of true labels, namely the right next word outside the window the model trying to predict. Instead, a negative sampling is performed to draw the true label with a bunch of other false labels and the cross entrophy is computed based on these samples. In the current implementation, a total of 100 samples are drawn for each input window including the true label.

In the training stage, it turned out that adaptive gradient ("AdaGrad") was able to converge to a smaller loss. AdaGrad is a gradient-based optimization method that takes advantage of historical gradient information and use it as way to adaptively adjust learning rate at each step.

Figure 2.3:



Next sub-section will focus on setting up the baseline model for topic classification.

2.4 Baseline Model

The baseline model used tf-idf vector as input to a Gaussian Naive Bayes classifier. Naive Bayes models assume the independence between features and use Bayes' theorem to infer the class of the each data point:

$$\tilde{y} = arg\max_{y} P(y) \prod_{i=1}^{n} P(x_i|y) \tag{2.3}$$

where $\tilde{y}$ is the estimated class given a data point $x = x_1, x_2, \cdots, x_n$ and y is the true class. In practice, Maximum A Posteriori (MAP) is used to estimate $P(y)$ and $P(x_i|y)$. In the Gaussian setting, the likelihood of a feature given class y is assumed to be Gaussian, namely:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right) \tag{2.4}$$

where σ_y and μ_y is the standard deviation and mean of the underlying Gaussian distribution.

2.4.1 Experiment

The experiment is carried out by splitting the dataset into training and testing subsets. In this case, the testing size is 0.75. To ensure the same train-test split can be reproduced, a fixed random state is set in numpy. In the baseline model, only the 10,000 most frequent words are chosen in the vector representation of each question in order to fit into the memory.

2.4.2 Baseline Results

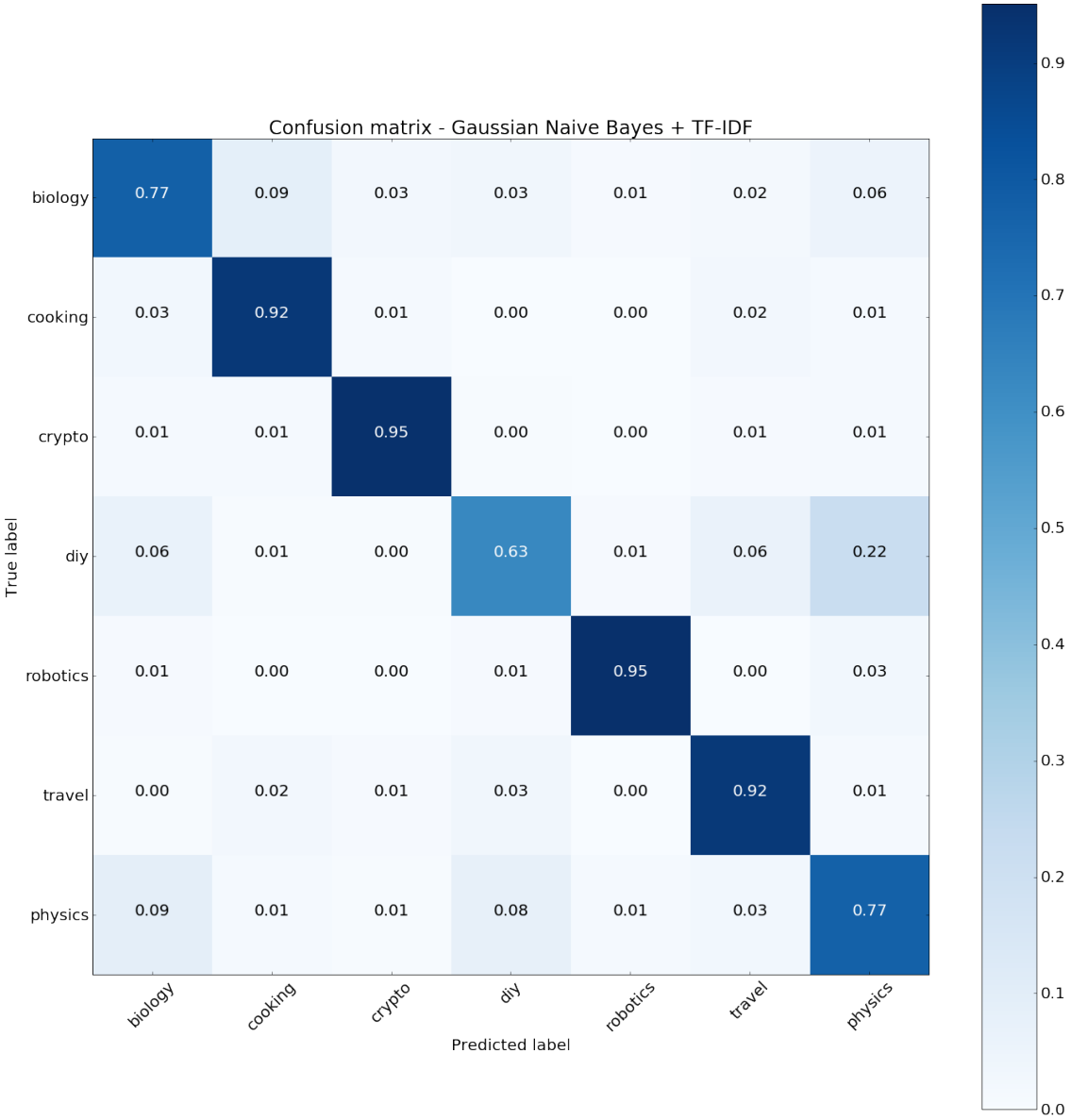
As mentioned in previous section, the performance of the classification model is evaluated from three metrics: precision, recall and F1-score. Since there are in total seven classes, an average of each metric is calculated across each class. In the unweighted case, data imbalance is not considered while in the weighted case metrics are averaged based on the number of instances in each class. The final results are shown in table.

Table 2.1: Precision, recall and F1-score

	Precision	Recall	F1-Score
unweighted	0.75	0.84	0.77
weighted	0.88	0.84	0.85

To obtain more granular details of the results, a confusion matrix is calculated and visualized (see Figure 2.4). As can be seen from the map, topics like cooking, crypography, robotics and travel are easily distinguished by the baseline classification model while other topics such as biology, diy and physics are not well classified. Biology is sometimes confused with cooking (0.09) or physics (0.06) and diy is often confused with physics (0.22). And vise versa physics questions are often mistaken as biology (0.09) or diy (0.08).

Figure 2.4:


Confusion matrix - Gaussian Naive Bayes + TF-IDF
True label
Predicted label
biology
cooking
crypto
diy
robotics
travel
physics
0.77 0.09 0.03 0.03 0.01 0.02 0.06
0.03 0.92 0.01 0.00 0.00 0.02 0.01
0.01 0.01 0.95 0.00 0.00 0.01 0.01
0.06 0.01 0.00 0.63 0.01 0.06 0.22
0.01 0.00 0.00 0.01 0.95 0.00 0.03
0.00 0.02 0.01 0.03 0.00 0.92 0.01
0.09 0.01 0.01 0.08 0.01 0.03 0.77
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0