CSC2541 - Problem Set 1

Sumeet Ranka

February 6, 2019

Part 1

1. With respect to the columns in the table `admissions`, the patient with id = 40080 is a African-American widowed woman. She lived for 79 years. Her religious beliefs are unknown. She can speak Hait.

```
select * from admissions,patients where admissions.
    subject_id='40080' and patients.subject_id='40080';
```

2. The patient was primarily diagnosed for the condition: *Acute on chronic combined systolic and diastolic heart failure.* The corresponding ICD9 Code is 42843.

```
select * from d_icd_diagnoses where icd9_code=(select
    icd9_code from diagnoses_icd where subject_id
    ='40080' and seq_num=1);
```

3. She was admitted in the ICU for approximately 5 days (to be exact 4.8577 days). She was discharged with a minimally clear and coherent mental status. She had a minimally alert and somewhat interactive level of consciousness. She was reported to be bed bound and a dependent hemiplegia.

```
select * from noteevents where subject_id='40080'
    order by chartdate desc,charttime desc;
```

4. The highest heart rate recorded is 141 and the lowest heart rate is 80.

```
select min(valuenum), max(valuenum) from chartevents
    where itemid in (select itemid from d_items where
    label='Heart Rate') and subject_id='40080';
```

Part 2

2a. The following columns are dropped after splitting the data into training and validation set:

- train

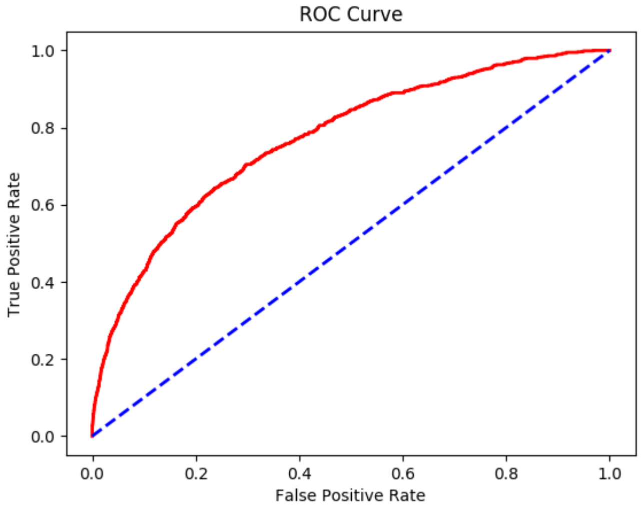


Figure 1: ROC Graph for model trained on `adult_icu.gz` for predicting mortality

Top 5	**Lowest 5**
bilirubin (+)	admType_NEWBORN
lactate (+)	glucose_mean
tempc_min (-)	eth_asian
tempc_max (+)	ptt
creatinine (-)	eth_white

Table 1: Top 5 and Lowest 5 risk factors associated with mortality. The sign indicates whether the factor is positively or negatively correlated.

- subject_id
- hadm_id
- icustay_id
- mort_icu

The numerical-valued columns are min-max normalized before feeding into the logistic regression model.

After the above stated preprocessing, the AUC Score obtained is 0.7709. Figure 1 shows the ROC Graph obtained.

To compare the coefficients, the absolute values of the corresponsing weights are considered. The top 5 and lowest 5 risk factors obtained are mentioned in the table 1.

The model takes 11 iterations to converge when the data is standardized. However, it takes 203 iterations if not. The AUC score in the non-standardized case is slightly higher (0.7768). However, in this case we cannot comment on the risk factors as the values for every column is not in the same range. From the ROC Curve, we can conclude that the model performes moderately good. It is more accurate than the baseline but not sufficiently steep in the left to make it a deployable model.

The following are the explanation of the top 5 and the lowest 5 mortality factors.

- High levels of `bilirubin` indicates liver inefficiency or indication of diseases such as jaundice.
- A general increase in the `lactate` means a greater severity of the condition. It indicates the concentration of lactate in the blood.
- `tempc_min` indicates the minimum temperature (in celsius) that the patient had. Lower body temperature means the person may be suffering from hypothermia.
- `tempc_max` indicates the maximum temperature (in celsius) that the patient had. Higher temperature means the harder your body is working to fight the infection.
- `creatinine` is a waste product from the normal breakdown of muscle tissue. High levels indicate the inefficiency of kidney functioning.
- `admType_NEWBORN`: Boolean variable indicating whether the person admitted is a new born.
- `glucose_mean`: Mean value of the glucose across all the readings. The glucose level could be controlled for the patients.
- `eth_asian`: Boolean variable stating whether the patient belongs to asian ethnicity.
- `eth_white`: Boolean variable stating whether the patient belongs to white ethnicity.
- `ptt`: Partial Thromboplastin Time. It indicates the ability of the body to form blood clots. Some literature points to the fact that high values of ptt increases the mortality rate. However, the patients may be under the medication that controls the clotting ability.

2b. `RegexpTokenizer` is used to tokenize the clinical notes. On training a L1-Regularized Logistic regression model using TF-IDF feature vectors, AUC score of 0.8250 is observed. The model converged in 16 iterations. Figure 2 shows the ROC Graph obtained.

The top five (both positively and negatively correlated) and lowest five words associated with mortality obtained from the data are mentioned in the table 2.

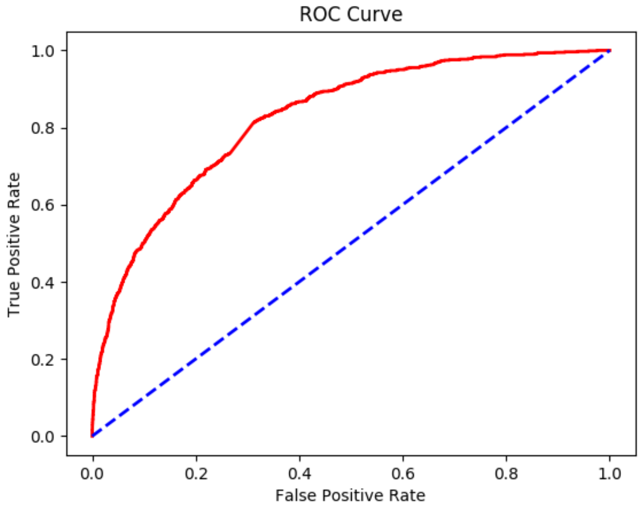


Figure 2: ROC Graph for model trained on `adult_notes.gz` for predicting mortality

Top Five Positive	**Top Five Negative**	**Lowest Five**
prognosis	diet	00
cmo	extubation	000
dnr	clear	0000
corneal	extubated	0000am
worsening	good	0000d

Table 2: Top Five positively and negatively correlated words and Lowest Five words associated with mortality

The word obtained somewhat reflects the condition of the patient. For example, the words like *worsening* and *dnr* reflects a bad condition of the patient whereas words like *extubation* and *good* reflects normal conditions. Words in the last column seems to not hold importance when trying to interpret the condition of the patient.

2c. Combining the model increases the AUC Score to 0.8471. The graph is more closer to the left border and the top border. Figure 3 displays the ROC Graph.

Part 3

This part deals with the hypertension prediction based on various vital measurements.

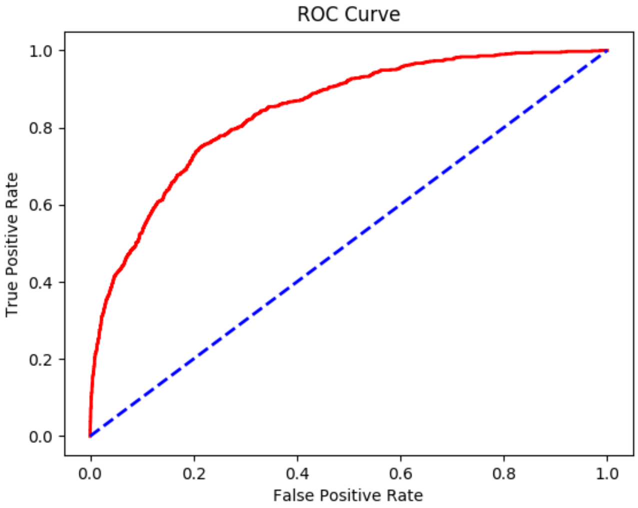


Figure 3: ROC Graph for model trained on both `adult_icu.gz` and `adult_notes.gz` for predicting mortality

3a. AUC Scores and F1 Scores obtained by training a logistic regression model on different feature sets are mentioned in the table 3. ROC curves are displayed in the figure 4.

Feature Set	**AUC Score**	**F1 Score**	**# patients removed**
Heart Rate	0.5262	0.0	13
Respiratory Rate	0.5275	0.0013	16
O2 Saturation	0.5129	0.0	21
Blood Pressure	0.5434	0.012	157

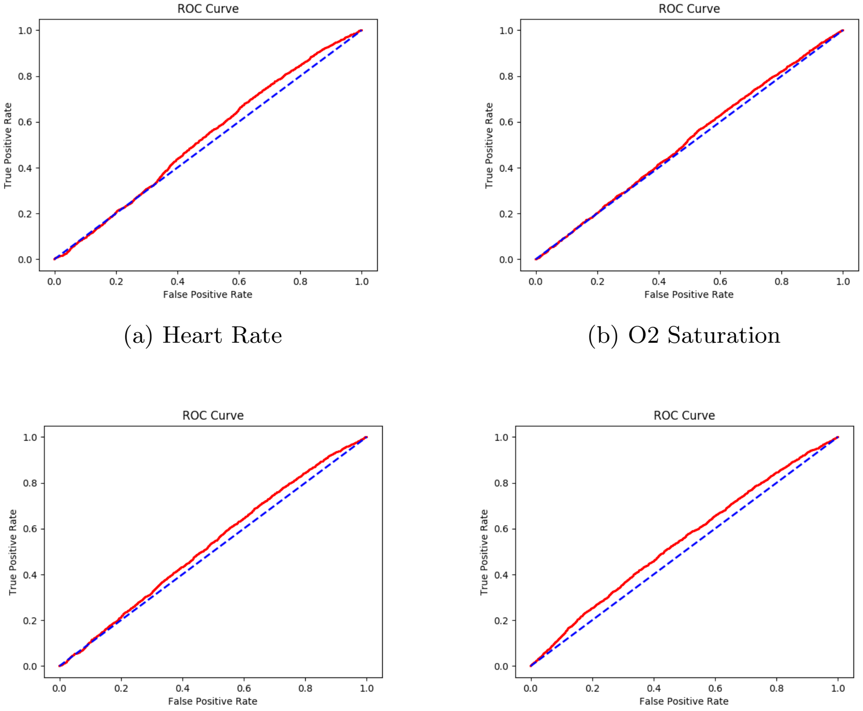
Table 3: AUC Scores and F1 scores obtained by using Logistic Regression to predict hypertension using various feature sets

3b. AUC Scores and F1 Scores obtained by training a LSTM model on different feature sets are mentioned in the table 4. ROC Curves are displayed in the figure 5.

3c. Based on the performance, none of the models should be used.

The aggregated values such as min, max and mean do not provide complete information to predict hypertension.

The sequences provided to LSTM may be overwhelmed by the normal values.



(a) Heart Rate

(b) O2 Saturation

(c) Respiratory Rate

(d) Blood Pressure

Figure 4: ROC curves for different feature sets for a trained logistic regression

Feature Set	AUC Score	F1 Score	# patients removed
Heart Rate	0.5080	0.0	13
Respiratory Rate	0.5373	0.0006	16
O2 Saturation	0.5	0.0006	21
Blood Pressure	0.51	0.0	157

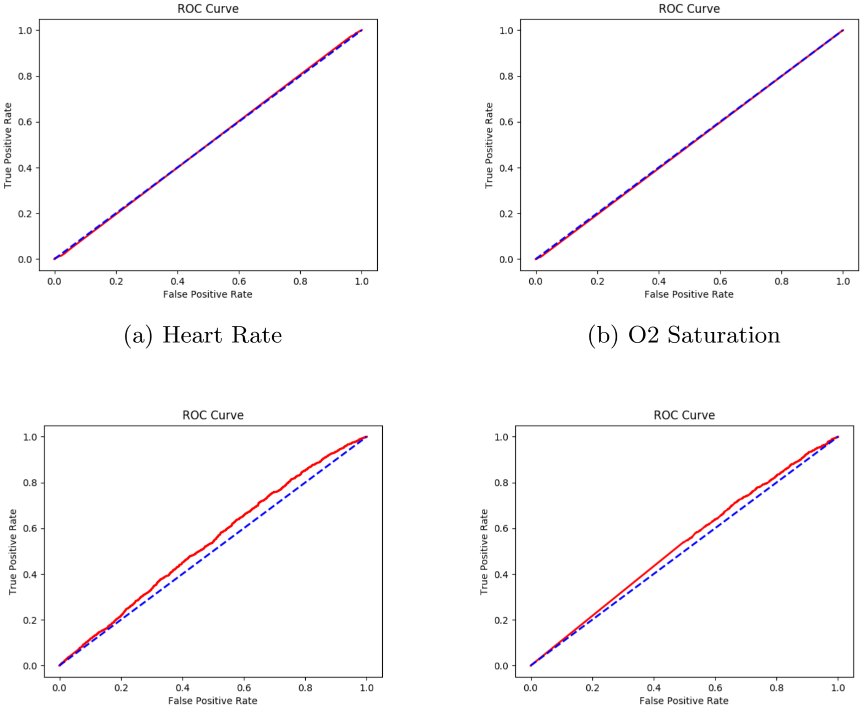
Table 4: AUC Scores and F1 scores obtained by using LSTM to predict hypertension using sequences obtained from various feature sets

Part 4

The text is tokenized using the `RegexpTokenizer` with the regex `[a-zA-Z0-9]+`. All the stopwords and the punctuations are removed and the tokens are converted to lowercase. UMass is calculated as a coherence score.

4a. Table 5 contains the coherence score for LDA model trained for different number of topics. 20 number of topics gives the highest coherence score (i.e. least negative).

4b. The relevant words are filtered from the top 100 words that belong to



(a) Heart Rate

(b) O2 Saturation

(c) Respiratory Rate

(d) Blood Pressure

Figure 5: ROC curves for different feature sets for a trained LSTM

# Topics	Coherence Score
20	-0.2357
50	-0.2417
100	-0.3203

Table 5: Coherence Score of LDA model for different number of topics

the same topic.

- **respiratory**: response, lung, tube, abg, intubated, airway, intubation, sputum, chest
- **vomiting**: pain, dilaudid, nausea, drain, renal, fluid, blood
- **urine**: pain, pneumonia, blood, fluid, acute
- **pulse**: patient, icu, respiratory, rhythm, hr, bp, review

The topics can be identified as follows based on the top words:

- **Initial Procedure to admit or look up a patient**: assessment, plan, patient, name, status, response, sounds, last, hr
- **Brain related**: head, seizure, ct, mental, neuro, mri, left, right
- **Torso related**: contrast, pelvis, abdomen, fluid, bowel, abdominal, chest, radiology
- **Circulatory system**: artery, right, aneurysm, left, carotid, catheter, embolization, hemorrhage, angio, sheath
- **Lung/Chest related**: chest, tube, right, left, pneumothorax, pleural, effusion, radiology
- **Kidney/Liver related**: liver, cirrhosis, hepatic, renal, ascites, transplant, ercp
- **Heart related**: cabg, artery, coronary, gtt, aortic, iabp, wires, left, bypass, valve

These are the most interpretable topics. The others are mostly related to the first topics, *Initial Procedure to admit or look up a patient*, or contain the words that are frequent such as *pt*, *name*, *hr*, *c* and *w*.