COMP550 - Assignment 1 Solutions

Fall 2019

1 Find Cases of Ambiguity (40 points)

1.1 Every student took a course. (8 points)

1.1.1 Every... a...

1. Ambiguity/Interpretations (3 points): One course for all students, or a potentially different course per student.
2. Cause/Domain (2 points): Scopal/semantic ambiguity.
3. Knowledge needed (3 points): Knowledge about typical student behavior; access to school records.

1.1.2 Alternative answer: "took a course"

1. Ambiguity/Interpretations (3 points): Enroll in a learning experience; physically take one portion of a meal.
2. Cause/Domain (2 points): Lexical ambiguity.
3. Knowledge needed (3 points): Statistical knowledge about which reading is more likely/natural; contextual knowledge about topic of discussion; database of idioms.

1.2 John was upset at Kevin but he didn't care. (8 points)

He = John? OR he = Kevin?

1. Ambiguity/Interpretations (3 points): Reference ambiguity.
2. Cause/Domain (2 points): Pragmatics/discourse.
3. Knowledge needed (3 points): Knowledge about the event and the attitudes of the people involved; knowledge about people who tend to care vs. people who are upset; knowledge of the meaning of "but" being counter to usual expectation.

1.3 Sara owns the newspaper. (8 points)

... "the newspaper"

1. Ambiguity/Interpretations (3 points): The paper material used to make a newspaper; copy of a newspaper; company/organization.
2. Cause/Domain (2 points): Lexical ambiguity.
3. Knowledge needed (3 points): Access to knowledge about newspaper organization's owner; knowledge about Sara's financial background.

1.4 He is my ex-father-in-law-to-be. (8 points)

... "ex-father-in-law-to-be"

1. Ambiguity/Interpretations (3 points):
   Ex-(father-in-law-to-be): Broke up with fiancé(e); got married.
   (Ex-father-in-law)-to-be: About to break up with fiancé(e); about to get married.
2. Cause/Domain (2 points): Caused by ambiguity in morphological structure.
3. Knowledge needed (3 points): Need knowledge of situation to be able to disambiguate.

1.5 ttyl ;) [text message] (8 points)

;)

1. Ambiguity/Interpretations (3 points): ;) has multiple meanings e.g. flirtations/ joking, secret joke.
2. Cause/Domain (2 points): Lexical ambiguity.
3. Knowledge needed (3 points): Need knowledge of sender and recipient and their relation.

Note: If ttyl has multiple meanings, that could be an alternate solution. One meaning is "talk to you later".

2 Naive Bayes and Logistic Regression (15 points)

Naive Bayes model:

$$p^{NB}(y|\vec{x}) = \frac{p^{NB}(y)\Pi_i p(x_i|y)}{p(\vec{x})} \tag{1}$$

Rewrite in LR form:

$$p^{LR}(y|\vec{x}) = \frac{1}{Z} \exp \Sigma_j a_j x_j \tag{2}$$

Prior:
Define LR feature $x_{class=k}$ for each class label k:

$$x_{class=k} = \mathbb{1}(y = k) \tag{3}$$

$$a_{class=k} = \log \left(p^{NB}(y = k) \right) \tag{4}$$

For each NB feature i with value v, define LR feature x_{kiv}:

$$x_{kiv} = \mathbb{1}(y = k)\mathbb{1}(\text{feature } i \text{ takes value } v) \tag{5}$$

$$a_{kiv} = \log \left(p^{NB}(x_i = v|y) \right) \tag{6}$$

So:

$$\begin{aligned} p^{NB}(y = k|\vec{x}) &= \frac{p^{NB}(y = k)\Pi_i p^{NB}(x_i|y)}{p(\vec{x})} \\ &= \frac{\exp \log p^{NB}(y = k)\ \Pi_i \exp \log p^{NB}(x_i|y = k)}{p(\vec{x})} \\ &= \frac{\exp \left(a_{class=k} x_{class=k} + \sum_i a_{kiv} x_{kiv} \right)}{p(\vec{x})} \end{aligned} \tag{7}$$

with:

$$\begin{aligned} p(\vec{x}) = \sum_k p(\vec{x}, y = k) &= \sum_k p(y = k) p(\vec{x}|y = k) \\ &= Z \\ &\text{by similar argument.} \end{aligned} \tag{8}$$

So if LR has features $\bigcup_k \{x_{class=k}\} \cup \{x_{kiv}\}$, the above is exactly in the form of a LR model.

3 Sentiment Analysis (45 points)

Grading Scheme:
Code [20 points]

- 6 points: Does the code have no major errors?
- 3 points: Calls the right functions?

- 2 points: Does preprocessing?
- 3 points: Seems to do feature extraction correctly.
- 2 points: Correctly sets aside a dev set or does k-fold cross validation?
- 4 points: Tests a number of variants to the model on the dev set?

Report [25 points]

- 4 points: Linguistic quality of report
- 5 points: Does it clearly describe the experiments?
- 3 points: Does it show the range of parameter settings tested?
- 3 points: Reports accuracy?
- 4 points: Results/Conclusions/Successes and failures.
- 3 points: Confusion matrix as a form of error analysis?
- 3 points: Explains role of dev set?

General note on Q3:
Many students did not have any kind of validation/development. Instead they did their model selection based on the performance on the training set. This is actually not how supervised learning is done. You need to have a separate validation set (or do cross-validation) on which you evaluate your model performance based on the different settings/hyperparameters and this is how the best model/configuration is chosen before ultimately testing the model on the test set.

Notes on the report part:

- For the second point (experiments), points were deducted if the dataset split was not specified.
- For the hyperparameters, many students did not attempt to do any kind of experimentation with the hyperparameters of the learning algorithms. Full grade for this item was given only if the student tested at least one algorithm hyperparameter *and* specified the range/values of these hyperparameters. If no range/values were presented, one point was deducted (since the range was specifically asked).
- Some students did not show any confusion matrix in their report. That led to points being deducted even if the code would show it (the confusion matrix was asked for the report part).
- If you did not use any kind of validation, you lose the points for the last item (explaining the role of the validation set). If you do have validation but do not talk (in any way whatsoever) about the role of validation, points were deducted.