Assignment 3

CMSC 678 — Introduction to Machine Learning

Due Friday March 29th, 2019, 11:59 PM

In this assignment you will experiment with maximum entropy models and explore intricacies of evaluating multiclass classifiers.

You are to complete this assignment on your own: that is, the code and writeup you submit must be entirely your own. However, you may discuss the assignment at a high level with other students or on the discussion board. Note at the top of your assignment who you discussed this with or what resources you used (beyond course staff, any course materials, or public Piazza discussions).

The following table gives the overall point breakdown for this assignment.

<table>
<thead>
<tr>
<th>Item</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assigned</td>
<td>Saturday March 9th, 2019</td>
</tr>
<tr>
<td>Due</td>
<td>Friday March 29th, 2019</td>
</tr>
<tr>
<td>Topic</td>
<td>Multiclass Classification (II): Maximum Entropy Models and Evaluation</td>
</tr>
<tr>
<td>Points</td>
<td>120</td>
</tr>
</tbody>
</table>

This handout may be lengthy, but think of it as both a tutorial and assignment. I provide a lot of explanation and re-summarizing of course concepts.

However, because this assignment handout is lengthy, I am first providing a task list. This task list captures the essence of the questions; it details, without other explanatory text, the tasks you are to do and what your completed assignment should answer. The task list enumerates what you must do, but it does not necessarily specify how—that’s where the full questions come in.

Following the task list are the full questions. The full questions do not require you to answer additional questions, but they do provide specific details, hints, and explanations. You should still read and reference the full questions.

What To Turn In Turn in a PDF writeup that answers the questions; turn in all requested code necessary to replicate your results. Be sure to include specific instructions on how to build (compile) and run your code. Answers to the following questions should be long-form. Provide any necessary analyses and discussion of your results.

How To Submit Submit the assignment on the submission site:

https://www.csee.umbc.edu/courses/graduate/678/spring19/submit

Be sure to select “Assignment 3.”
1. (a) Let $\mathcal{X}$ represent the flips of two different (two-sided) coins. Define two lexical indicator (binary) features $f_1(x)$ and $f_2(x)$, i.e., $f_1(x)$ returns 1 if the first coin is heads.
   (i) Write out (and simplify) the unnormalized probability $p(x)$, for each of the $x \in \mathcal{X}$.
   (ii) What is $\sum_{x \in \mathcal{X}} p(x)$?
   (iii) Write out (and simplify) the normalization constant $Z$.

(b) Let $\mathcal{X}$ be all 10x10 black and white images.
   (i) How many elements are in $\mathcal{X}$?
   (ii) Write out the normalization $Z$.
   (iii) Describe how, and why, this $Z$ is different from the normalization you would compute in a maxent classifier $p(y|x)$.

2. Turn in a written summary of maxent modeling. Your summary should consider feature design, the issue of matching expected to observed feature values, regularization, and joint vs. conditional modeling. There is no minimum or maximum page limit, but a 1/2 page summary is reasonable.

3. (a) Describe the different requirements between the $y_{\text{pred}}$ argument for metrics.classification_report and metrics.log_loss. (Don’t focus on the shape; assume you’re dealing with the same shape.)

(b) For a specific value of predictions $\hat{y}$ and correct values $y^*$ (parametrized by “A” and “B”):
   (i) What is the accuracy of the predictions $\hat{y}$ for the correct values $y^*$?
   (ii) If $A = 1, B = 0$, what is f1_score($y_{\text{true}}=y^*, y_{\text{pred}}=\hat{y}$)?
   (iii) If $A = 1, B = 0$, discuss the differences in $F_1$ for different values of the parameter average.
   (iv) If $A = 1, B = 0$, discuss the differences in $F_1$ for different values of both the parameter average and the parameter pos_label. Discuss the meaning and importance of this pos_label parameter.
   (v) If $A = 1, B = 0$, discuss the differences in $F_1$ for different values of both the parameter average and the parameter labels. Discuss the meaning and importance of this labels parameter.
   (vi) Now let $A = \text{foo}, B = \text{bar}$: answer the above question again? Did your answer change?

(c) Discuss the overall importance of the pos_label and labels parameters, in both a binary and multiclass setting. In your answer, be sure to address the implications if you do (or do not) include them in your evaluation.

4. Discuss the differences between training a $C$-class maxent classifier (trained to minimize cross-entropy loss) and $C$ one-vs-all binary logistic regression classifiers (each trained to minimize cross-entropy loss). In particular, discuss the different sum-to-one requirements—and any implications those may have—and formalize the different objectives.

5. Train a maxent classifier on int-train and evaluate it on int-dev. Try to find the best model configuration possible, from at least 4 different configurations. In your report, document the internal development progress, i.e., how different model configurations perform on int-dev. You may use an existing maxent implementation.
6. Using the maxent configuration found in the previous question, train a new model (including a baseline) on the entire, original train set. At the end of this training, you should have two models trained on the entire 60,000 training set. Evaluate these models on the original 10,000 image test set. Include these evaluation results, and a discussion of them, in your report.

This concludes the “task list.” If this document only has 3 pages (and not 6 pages), please see the fully detailed PDF at https://www.csee.umbc.edu/courses/graduate/678/spring19/materials/a3.pdf. While the detailed PDF does not require any additional items to be completed, it specifies and clarifies many of the details.
Full Questions

Theory

1. (15 points) In class we discussed maxent classifiers of the form \( p(y|x) \propto \exp(\theta^T f(x, y)) \), for example, given an image \( x \), predict the label \( y \). However, maxent models can also be used to simply model observations \( x \), e.g., to develop a probabilistic model over images \( x \). In this usage, our maxent model \( p(x) \propto \exp(\phi^T f(x)) \) provides a likelihood over the possible inputs. Let \( x \in \mathcal{X} \), e.g., if \( x \) is an image, then \( x \) is an element of all “possible” images \( \mathcal{X} \), but if \( x \) is a document, then it is an element of all “possible” documents \( \mathcal{X} \). We’ll call this \( \mathcal{X} \) the domain; in each of the following, the domain may change and will be specified.

(a) Let \( \mathcal{X} \) represent the flips of two different (two-sided) coins: \( \mathcal{X} = \{(H, H), (H, T), (T, T), (T, H)\} \).

Each \( x \in \mathcal{X} \) is one of these pairs. Define two lexical indicator (binary) features \( f_1(x) \) and \( f_2(x) \): \( f_1(x) \) returns 1 if the first coin is heads, and \( f_2(x) \) returns 1 if the second coin is heads. For example, \( f(x = (H, H)) = (1, 1) \) while \( f(x = (T, H)) = (0, 1) \).

(i) Write out (and simplify) the unnormalized probability \( p(x) \), for each of the \( x \in \mathcal{X} \). You may simply write \( p(x) \propto ... \), or \( p(x) = \frac{1}{Z} ... \) (where ... is the simplified expression).

(ii) What is \( \sum_{x \in \mathcal{X}} p(x) \)?

(iii) Write out (and simplify) the normalization constant \( Z \).

(b) Let \( \mathcal{X} \) be all 10x10 black and white images. That is, each \( x \in \mathcal{X} \) is a 10x10 matrix, and each cell \( x_{i,j} \) is either 0 or 1.

(i) How many elements are in \( \mathcal{X} \)?

(ii) Write out the normalization \( Z \). You may leave your answer in a sigma notation (e.g., \( \sum_j \)), but be sure to clearly specify what the summation indexes over.

(iii) Describe how, and why, this \( Z \) is different from the normalization you would compute in a maxent classifier \( p(y|x) \).

2. (25 points) In long-form prose, discuss the following aspects of maxent modeling:

- The effects of symmetric and conjoined features (conjunctions of features) [roughly lessons 1-4].
- The impact of matching observed vs. expected feature values [roughly lessons 3-6].
- The intended effects of \( \ell_2 \) vs. \( \ell_1 \) vs. no regularization [roughly lessons 8-10].
- The difference in maximizing joint (sometimes also called global, or non-conditional) log-likelihood vs. maximizing conditional log-likelihood [roughly lessons 11-15].
- The impact of context-only (vs. outcome-only or joint outcome-and-context) features in conditional modeling [roughly lessons 12-16].

You may use math where appropriate, but do try to come up with succinct written (English) explanations. There is no minimum or maximum page limit, but a 1/2 page summary is reasonable.

The lessons refer to the online tutorial at [https://www.csee.umbc.edu/courses/graduate/678/spring19/loglin-tutorial](https://www.csee.umbc.edu/courses/graduate/678/spring19/loglin-tutorial) You are highly encouraged to work through the tutorial, but you do not have to. Each lesson has some thought questions posed in the instruction box; you should think about them if you work through the tutorial, but you do not have to turn in your responses to them.
3. (25 points) In this question, you will be using the [classification metrics] from the sklearn.metrics module.

(a) Describe the different requirements between the y_pred argument for metrics.classification_report and metrics.log_loss. (Don’t focus on the shape; assume you’re dealing with the same shape.)

(b) Define your array of correct values as $y^\ast = (A, A, A, B, A, B)$, and your predicted values as $\hat{y} = (A, A, B, A, B, A)$. This question will explore what happens as you use different values for $A$ and $B$, and different settings of the parameters for a function like f1_score. I would recommend defining a function such as the following:

```python
def get_correct(A, B):
    return numpy.array((A, A, A, B, B, A))
```

(i) What is the accuracy of the predictions $\hat{y}$ for the correct values $y^\ast$?

(ii) If $A = 1$, $B = 0$, what is f1_score(y_true=y*, y_pred=\hat{y})?

(iii) If $A = 1$, $B = 0$, discuss the differences in $F_1$ for different values of the parameter average.

(iv) If $A = 1$, $B = 0$, discuss the differences in $F_1$ for different values of both the parameter average and the parameter pos_label. Discuss the meaning and importance of this pos_label parameter.

(v) If $A = 1$, $B = 0$, discuss the differences in $F_1$ for different values of both the parameter average and the parameter labels. Discuss the meaning and importance of this labels parameter.

(vi) Now let $A = \text{foo}$, $B = \text{bar}$: answer the above question (v) again? Did your answer change?

(c) Discuss the overall importance of the pos_label and labels parameters, in both a binary and multiclass setting. In your answer, be sure to address the implications if you do (or do not) include them in your evaluation.

4. (15 points) We discussed multiple options for turning a binary classification problem into a $C$-class one. One is to simply generalize the binary classifier to handle the $C$ different classes; this constructs a single classifier. Another is to form $C$ different one-vs-all (also sometimes called one-vs-rest) binary classifiers. Discuss the differences between training a $C$-class maxent classifier (trained to minimize cross-entropy loss) and $C$ one-vs-all binary logistic regression classifiers (each trained to minimize cross-entropy loss). In particular, discuss the different sum-to-one requirements—and any implications those may have—and formalize the different objectives.

**Classification Project: Implementation, Experimentation, and Discussion**

The next two questions continue the previous assignment’s classification project. As before, the core deliverables for these questions are:
The data we’ll be using is the MNIST digit dataset from A2. You may reuse your data splits from A2 (but if not, tell us how you re-split the data, and provide a similar analysis to A2, Q4).

5. (25 points) Design a maxent classifier to do the prediction task. You do not have to implement your own maxent classifier. You can use any existing maxent implementation; just tell us which one it is.

Train a maxent classifier on int-train and evaluate it on int-dev. Try to find the best model configuration possible. In your report, document the internal development progress, i.e., how different model configurations perform on int-dev.

Experiment with at least 4 different configurations, two of which must involve the design of different features. For example, your configurations could be (feature set 1, regularization 1), (feature set 1, regularization 2), (feature set 2, regularization 1), (feature set 2, regularization 2).

A popular implementation is liblinear. It is on GL at /afs/umbc.edu/users/f/e/ferraro/pub/678-s19/liblinear-2.20.

Within this directory there are compiled binaries in bin/train and bin/predict; they have been compiled for GL. There are also Python and MATLAB modules of liblinear. You may also use the sklearn implementation, or build one using another toolkit (like Pytorch or Tensorflow). However, if you use Pytorch or Tensorflow, you must include a virtual environment in your submission.

6. (15 points) Using the best maxent configuration found in the previous question, and a baseline of your choice, train new models on the entire, original train set. At the end of this training, you should have two models trained on the entire 60,000 training set. Evaluate these models on the original 10,000 image test set. Include these evaluation results in your report and discuss the results.

1 https://www.csie.ntu.edu.tw/~cjlin/liblinear