Team (of 3) Project
Points: 20

COSC6340.001
Dr. Minhua Huang

Out: 10/07/2022, Due 11/17/2022

Objective:

Identify an English phrase on Bigram Language Model by Perplexity

You can call functions and facilities on the preprocessing procedure. You are not allowed to call functions for obtaining bigram, corpus cross entropy, and perplexity, and the test accuracy.

Given corpora $\mathbf{D}$, where $\mathbf{D} = < \mathbf{x[i]}, \mathbf{y[i]} > |\mathbf{i} = \mathbf{1}...\mathbf{n}$, s.t. each of $\mathbf{x} =< \mathbf{verb}, \mathbf{noun}, \mathbf{prep}, \mathbf{prepobj} > =< \mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}, \mathbf{x_4} >$ with a class label $\mathbf{y} \in \{\mathbf{V}, \mathbf{N}\} = \{\mathbf{y_1}, \mathbf{y_2}\}$. The corpora D is divided into two sets which are **Dtrain** and **Dtest**, specified by **Dtrain.csv** and **Dtest.csv** files.

- (4.5 points, Implementation) Training procedures.
  - Compute bigram probability for $\mathbf{j^{th}}$ attribute of $\mathbf{i^{th}}$ feature under a class label $\mathbf{y}$ for all $\mathbf{i}$, $\mathbf{j}$ and $\mathbf{y}$ in **Dtrain** by **MLE** algorithm (and a smoothing technique), where $\mathbf{C}$ is a counting function.

$$\mathbf{p}(\mathbf{x_{i,j}}|\mathbf{x_{i-1,j}}, \mathbf{y}) = \frac{\mathbf{p}(\mathbf{x_{i-1,j}}, \mathbf{x_{i,j}}, \mathbf{y})}{\mathbf{p}(\mathbf{x_{i-1,j}}, \mathbf{y})} = \frac{\mathbf{C}(\mathbf{x_{i-1,j}}, \mathbf{C_{x_{i,j}}}, \mathbf{y})}{\mathbf{C}(\mathbf{x_{i-1,j}}, \mathbf{y})}$$

- (4.5 points, Implementation) Testing procedures.
  - Compute the corpus cross entropy for each of the data instances in **Dtest**. A data instance of size $\mathbf{m}$ is associating with a probability distribution $\mathbf{p}$ with $\mathbf{m}$ probabilities.

$$\mathbf{H}(\mathbf{p}|\mathbf{y}) = -\sum_{\mathbf{i=1}}^{\mathbf{m}} \mathbf{p_{i,y}} \mathbf{log_2 p_{i,y}}$$

  - Compute the perplexity of the probability distribution of $\mathbf{p}$.

$$\mathbf{PP}(\mathbf{p}|\mathbf{y}) = \mathbf{2^{H(p|y)}}$$

  - Assign a class label for a data instance in **Dtest**.

$$\mathbf{y} \leftarrow \mathbf{argmin_{y_k}}\{\mathbf{PP}(\mathbf{p}|\mathbf{y} = \mathbf{y_k})\}$$

  - Evaluate your system by the following accuracy measurement.

$$\mathbf{ACC}_{\mathcal{D}_{\mathbf{test}}} = \frac{\mathbf{1}}{|\mathcal{D}_{\mathbf{test}}|} \sum_{\mathbf{i=1}}^{\mathbf{T}} \mathbf{L}(\hat{\mathbf{y}}^{\mathbf{i}}, \mathbf{y}^{\mathbf{i}})$$

    where $\hat{\mathbf{y}}^{\mathbf{i}}$ is the assigned class label by the classifier and $\mathbf{y}^{\mathbf{i}}$ is the true class label of a data instance $\mathbf{x}^{[\mathbf{i}]}$ in $\mathcal{D}_{\mathbf{test}}$ and $\mathbf{T}$ is the number of data instances in $\mathcal{D}_{\mathbf{test}}$.

$$\mathbf{L}(\hat{\mathbf{y}}^{\mathbf{i}}, \mathbf{y}^{\mathbf{i}}) = \begin{cases} 1 & \text{if } \hat{y}^i = y^i \\ 0 & \text{if } \hat{y}^i \neq y^i \end{cases}$$

- (7 points) Report on your design (10 pages).
  Write a 10 page report. The first page should list the names of group members as well the associated tasks.
  - Introduction
  - Describing your algorithms.
    - Preprocessing procedures
    - The algorithm(s) of obtaining bigrams
    - The algorithm(s) of obtaining corpus cross entropy and perplexity
    - The running time complexity of an algorithm (optional)
    - The missing data handling
    - Testing metrics
  - Experiments
    - Experiments and results discussions.
      - Experiment settings.
      - The results discussions and comparisons.
      - Pros and cons of the design.
  - Further improvements.
- (4 points). Presentation.
  - 13 minutes of oral presentation.
  - 2 minutes of question-answering.

Presentation Date

- Nov. 21, 3 -4 groups
- Nov. 23 (reading day), 3 -4 groups
- Nov. 28, 3 -4 groups
- Nov. 30, 3 -4 groups

Submission:

- Submit files
  - ReadMe.txt – describe how to operating your system.
  - Project source codes
  - Project report
  - PPt presentation slides