1 Logistic Regression (30 points)

In this problem we look at the English dative alternation. You will work with a dataset of annotated examples of the dative alternation from spontaneous speech in conversation, reported in Bresnan et al. (2007) and made available through the `languageR` package (Baayen, 2007). The Colab notebook for this problem automatically downloads this dataset as `dative.csv`.

The dative alternation involves DITRANSITIVE verbs,

(1) a. Kim mailed $\overbrace{\text{me}}^{\text{Recipient}}$ $\overbrace{\text{a gift}}^{\text{Theme}}$. [D(ouble) O(bject)]
 b. Kim mailed $\underbrace{\text{a gift}}_{\text{Theme}}$ to $\underbrace{\text{me}}_{\text{Recipient}}$. [P(repositional) D(ative)]

and the THEME and RECIPIENT arguments are, respectively, what gets acted upon (usually transferred in physical location or possession), and the recipient or destination of the action. One qualitative intuition often reported about the dative alternation is that cases where the recipient argument is a large phrase (as measured, e.g., in number of words) are awkward in the Double Object construction, such as the below example:

(2) ?Kim mailed everyone who had attended the party yesterday a gift.

Alternatively, some researchers have proposed that what is more crucial is whether the recipient and theme arguments are pronouns. In this problem, we test these ideas by building simple logistic regression models of the dative alternation to look at the predictive effects of the length and pronominality of the recipient and theme arguments. We will use the `pandas` and `statmodels` Python packages for this.

In general, in studying the factors influencing speaker preference in the dative alternation, we will be interested in estimating the following probabilistic model:

$$P(\text{Construction} = \text{Double Object}|\text{Subject}, \text{Verb}, \text{Recipient}, \text{Theme})$$

where any of a number of features of the subject, verb, recipient, and theme might influence the speaker or writer's choice of linguistic construction. For purposes of this problem, however, we will dramatically simplify. First, we will ignore the subject and verb altogether, simplifying our problem to:[1]

$$P(\text{Construction} = \text{Double Object}|\text{Recipient}, \text{Theme})$$

Also, we'll only use a few features of the recipient and theme in our predictive model. Recall that logistic regression is characterized by the following equations:

$$\begin{aligned} \eta &= \sum_i \beta_i X_i && \text{(linear predictor)} \\ P(\text{success}) &= \frac{e^\eta}{1+e^\eta} && \text{(logistic transform of linear predictor)} \end{aligned}$$

For the dative alternation, the two possible outcomes are the double object construction (**DO**) and prepositional dative construction (**PD**). We arbitrarily choose DO as the outcome corresponding to "success".

Unlike the case of binomial ordering choice, there is a systematic difference between the possible outcomes, that holds across all instances of the dative alternation: it is always the same two constructions that are being chosen between. To capture the possibility of an overall preference for one construction or the other, we add what is called an "intercept" or "bias" term to the equation determining the linear predictor. This is often expressed in the statistics literature as (assuming M predictors):

$$\eta = \alpha + \sum_{i=1}^{M} \beta_i X_i \qquad \text{(linear predictor)},$$

but it can equivalently be expressed by defining a "dummy" predictor X_0 whose value is always 1, and writing the linear predictor as:

$$\eta = \sum_{i=0}^{M} \beta_i X_i \qquad \text{(linear predictor)},$$

[1] This is an oversimplification, actually: in particular, the identity of the verb has a *lot* of predictive information. This information is most effectively brought into our model by introducing a hierarchical component, endowing verbs with idiosyncratic preferences for which construction they prefer and by how much. Bresnan et al. (2007) and Morgan and Levy (2015) are good references for seeing how to include such a hierarchical (sometimes called "mixed-effects") component for the dative alternation and for binomials, respectively.

a formulation more common in the machine learning literature. (Note that it would be inappropriate to include an intercept in the model for binomial ordering preferences, because there is no intrinsic difference between "success" and "failure" outcomes that is consistently defined across different specific cases of binomial ordering choice.)

We can evaluate the quality of a logistic regression model in a couple of ways. One is predicting the CLASS of the outcome: here, DO or PD? We say that a logistic regression model predicts "success" for a datum if it assigns $P(\text{success}) > 0.5$ for that datum, otherwise "failure". A second is the LOG-LIKELIHOOD of the dataset—the summed log-probabilities of all the observations in the dataset under the fitted model.

Tasks:

1. Define and implement an 80/20 train/test random split of the `dative` dataset.

2. Fit a logistic regression model to the training set that uses *only* recipient pronominality and an intercept term. What is its classification accuracy on the held-out test dataset? How about its log-likelihood?

3. Add theme pronominality as a predictor to the model and see whether that improves the model's predictive power as assessed by held-out classification accuracy and log-likelihood.

4. Determine whether additionally adding theme and recipient length (in number of words) to the model further improves fit. Try both raw length or log-transformed length. Which gives better performance?

5. Look at the β coefficients of a fitted model with all four predictors and interpret them theoretically (don't worry about interpreting the intercept α, whose value will depend on the numeric coding scheme used for the predictors). Are there any general linguistic principles manifested in the values of all four predictors? Do you see any ways to simplify the model (reduce the number of predictor weights that have to be learned) based on general linguistic principles, without sacrificing much predictive accuracy? This may involve creating a new set of predictors that are a function of the four predictors you've been working with up until now.