PMDS505P Data Mining and Machine Learning
Experiment 5

February 2025

1 Work to do today

Note: Make a single pdf file of the work you are doing in jupyter notebook. Upload with proper format. Please mention your name and roll no properly with Experiment number in the first page of your submission.

Softmax Regression

Q1. Today we will try to implement softmax regression for solving a multiclass classification problem. For that lets create a synthetic dataset using the make classification function that we used in the previous class. Lets assume we have two features X1 and X2 to predict three classes to which they belong to. And try to plot the decision boundaries in that cases.

- First lets try to create a dataset for our purpose. You can use the code
  ```
  X, y = make_classification(n_classes=3, n_features=2, n_redundant=0, n_clusters_per_class=1, random_state=42)
  ```
- Do the train test split of the data with test size 20%.
- Next we will create an object of LogisticRegression class as clf. The same class can be used for softmax regression
  You can use the code.
  ```
  clf = LogisticRegression(multi_class=’multinomial’)
  ```
- Now we can fit the model.
  ```
  clf.fit(X_train, y_train)
  ```
- Next print the accuracy of your model.
- Further, we can try to plot the decision boundaries in this case. For that we prepare the mesh grid. you can use this code
  ```
  x_min, x_max = X[:, 0].min() - 1, X[:, 0].max() + 1
  y_min, y_max = X[:, 1].min() - 1, X[:, 1].max() + 1
  xx, yy = np.meshgrid(np.linspace(x_min, x_max, 200), np.linspace(y_min, y_max, 200))
  ```

- the above code helps you to create a mesh grid with 200*200 = 40000 points (x1,x2), for which you can find the respective classes to which they belong to according to your models prediction which is saved as Z in the next line of codes. Here np.c_[xx.ravel(), yy.ravel()] stacks these points together as coordinate pairs (x1,x2).

  ```
  Z = clf.predict(np.c_[xx.ravel(), yy.ravel()])
  ```

- Now we reshape back our Z.

  ```
  Z = Z.reshape(xx.shape)
  ```

- Further we use a function known as contour plot to plot the different class feature coordinates(xi,yi) in different colors. plt.contourf() creates filled contour plots to visually represent different decision regions. The x-coordinates come from xx, and the y-coordinates come from yy. The color of each region is determined by Z, which contains the predicted class for each grid point. So, that our decision boundary gets visualized in terms of that process. Here we do the scatter plot of training data as dots and testing data is plotted as ×.

  ```
  plt.contourf(xx, yy, Z, alpha=0.3)
  plt.scatter(X_train[:, 0], X_train[:, 1], c=y_train, edgecolors=’k’, label=’Train’)
  plt.scatter(X_test[:, 0], X_test[:, 1], c=y_test, marker=’x’, label=’Test’)
  plt.xlabel(”Feature 1”)
  plt.ylabel(”Feature 2”)
  plt.title(”Decision Boundaries of Softmax Regression”)
  plt.legend()
  plt.show()
  ```

Q2. Next try to implement softmax regression to fit a model in connection with the dataset ”Croprecommendation.csv” available for you to download in moodle. Since we have more features here it would be difficult to do the visualization in this case.

- Download the dataset ’Croprecommendation.csv’ from moodle. This Crop Recommendation Dataset contains soil and climatic parameters (N, P, K, temperature, humidity, pH, rainfall) along with the best-suited crop for those conditions. Here’s what each column represents:

  Columns Explained N (Nitrogen content in soil) – Amount of nitrogen in the soil. P (Phosphorus content in soil) – Amount of phosphorus in the soil. K (Potassium content in soil) – Amount of potassium in the soil. Temperature (°C)- Air temperature at the location. Humidity (%)- Percentage of moisture in the air. pH - Acidity or alkalinity of the soil. Rainfall (mm) -Annual rainfall received. Crop (Target variable) - The recommended crop based on the given conditions. Open the CSV file and see the different features and the target variable Y also. We are going to create a model for crop reccommendation in a particular area based on the input features given.

  Do the necessary preprocessing of the dataset. Build a classifier based on the Softmax regression concept and print the accuracy of your model.