CS 529: Assignment #3

Z. Berkay Celik
zcelik@purdue.edu
(Due Sunday 10/6/2019 11:59 PM)

Computer Science, Purdue University — September 25, 2019

Instructions

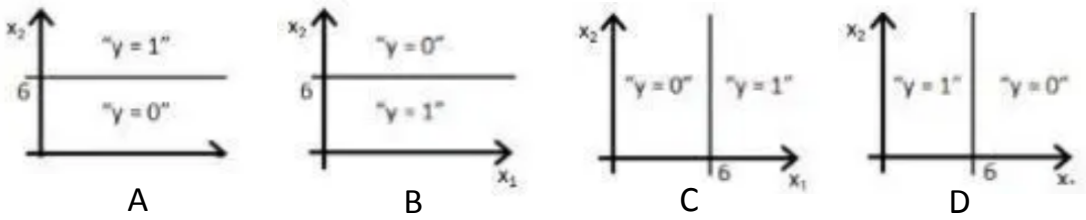
Info: This HW includes both theory and coding problems. Please read course policy before starting your HW.

- Your code must work with Python 3.5+ (you may install the Anaconda distribution of Python).
- You need to submit a report including solutions of theory problems (in pdf format), and a Jupyter notebook for programming problems that includes your source code.

1 Problem 1: Concepts [10pt]

Info: Please use at most four sentences per question. Q1.6 is 2.5 pt, others are 1.5 pt.

1. What would be the training error of nearest neighbors when k = 1? Justify your answer.
2. What are the differences between k-means clustering and Gaussian Mixture Model (GMM) clustering?
3. Compare Mahalanobis and Euclidean distance.
4. False or True (support your answer): The objective function optimized by the EM algorithm can also be optimized by a gradient descent algorithm which will find the global optimal solution, whereas EM finds its solution more quickly but may return only a locally optimal solution.
5. What is the principal assumption in the Naive Bayes' model, and when is this assumption useful?
6. Suppose you train a logistic regression classifier and your hypothesis function is $\mathtt{H} = \mathtt{g}(\mathtt{w}_0 + \mathtt{w}_1\mathtt{x}_1 + \mathtt{w}_2\mathtt{x}_2)$ where we found with MLE that $\mathtt{w}_0 = 6, \mathtt{w}_1 = 0, \mathtt{w}_2 = -1$. (1) Which plot below is the the decision boundary? (2) Plot the decision boundary when you replace coefficient of $\mathtt{x}_1$ with $\mathtt{x}_2$.



$$X = \left\{ \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \begin{bmatrix} 4 \\ 3 \end{bmatrix}, \begin{bmatrix} 5 \\ 4 \end{bmatrix} \right\}$$

2 Problem 2: Clustering [15pt]

You will cluster the data above. Using single link agglomerative hierarchical clustering and city block distance as the distance measure. Show each step of the clustering clearly.

1. What are the coordinates of the cluster centers (means) if you want to obtain two clusters? (**Hint:** Single link: distance between two clusters is the minimum of the pairwise distances between clusters. City block distance: sum of the absolute value of differences per dimension.)
2. Plot the Dendogram of the clusters.

3 Problem 3: Clustering [35pt]

Info: For this problem, you will implement your own algorithms.

For this problem, you will implement the k-means++ algorithm in Python. You will then use it to cluster Iris dataset from the UCI Machine Learning Repository. The data is contained the iris.data file under "Data Folder", while the file iris.names contains a description of the data. The features $\mathtt{x}$ are given as the first four comma-separated values in each row in the data file. The labels $\mathtt{y}$ are the last entry in each row, but you do NOT need the class label as it is unknown for clustering.

1. sepal length (cm)
2. sepal width (cm)
3. petal length (cm)
4. petal width (cm)
5. class: {Iris Setosa, Iris Versicolour, Iris Virginica}

You need to,

1. Create a new data set with two features by computing the ratio of raw features $\mathtt{x} = (\mathtt{x}_1, \mathtt{x}_2)$ where $\mathtt{x}_1$ = (sepal length/sepal width) and $\mathtt{x}_2$ = (petal length/petal width) and plot the data to observe the clusters in data by yourself (use class label to color the data points for better illustration of clusters).
2. Implement the k-means++ algorithm. You are provided with the skeleton of the code with main functions to be implemented (**Problem3.py** file in assignment directory). Submit the source code (documented!) of your implementation.
3. Cluster the modified Iris dataset with the two features explained above. Run your algorithm 50 times over the data with different values of clusters $\mathtt{k} = 1, 2, \ldots, 5$ and plot the accuracies ($\mathtt{x}$ and $\mathtt{y}$ axes should be the number of clusters and the clustering objective (accuracy, see `compute_objective()` in Problem3.py), respectively.
4. Based on the above plot, decide the number of final clusters and justify your answer. For the chosen number of clusters,
   - Create a plot showing how the accuracy changes with the number of iterations.
   - Create a plot with the data colored by assignment, and the cluster centers.

$$p(x \mid w_1) = N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}a & c\\ c & b\end{bmatrix}\right) \quad and \quad p(x \mid w_2) = N\left(\begin{bmatrix}d\\e\end{bmatrix}, \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}\right)$$

where $ab - c^2 = 1$.

4 Problem 4: Decision Boundaries [25pt]

Assume a two-class problem with equal a priori class probabilities and Gaussian class-conditional densities as above:

1. Find the equation of the decision boundary between these two classes in terms of the given parameters, after choosing a logarithmic discriminant function. **Hint:** The discriminant function using log likelihood is (this is similar to HW-2 related questions, $g_1(x) = g_2(x)$):

$$g_i(x) = \ln p\,(x|w_i) + \ln P\,(w_i)\,.$$

2. Determine the constraints on the values of `a`, `b`, `c`, `d` and `e`, such that the resulting discriminant function results with a linear decision boundary.

   Hints:

Hint1: If $\underline{\mathbf{x}} \sim \mathsf{N}_d\left(\underline{\boldsymbol{\mu}}, \boldsymbol{\Sigma}\right)$, then the pdf for x is given by:

$$p(\underline{\mathbf{x}}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}\left(\underline{\mathbf{x}} - \underline{\boldsymbol{\mu}}\right)^T \boldsymbol{\Sigma}^{-1}\left(\underline{\mathbf{x}} - \underline{\boldsymbol{\mu}}\right)\right]$$

Hint2:

For a 2×2 matrix,

$$\boldsymbol{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

the matrix inverse is

$$A^{-1} = \frac{1}{|A|}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix} = \frac{1}{ad - bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

5 Problem 5: Logistic Regression [15pt]

Info: For this question you will use the Optdigits data from UCI Machine Learning Repository. You need the files:

- optdigits.names // explanation of data
- optdigits.tra // training data
- optdigits.tes // test data

You can use scikit-learn to implement the logistic regression and regularized logistic regression. You will submit **problem5.pynb** for this questions that includes your code and answers.

Extra Credit [15pt]. If you implement algorithms through the steps presented in lecture slides without using scikit-learn, you will get extra credit. (**Hint:** We showed the derivation of logistic regression objective function for binary classification. You can now make logistic regression classifier work on multi-class classification problems with one versus all method. See Andrew Ng's explanation.)

1. Use the logistic regression (multi class) training algorithm and plot the training and test errors.
2. Use the logistic regression classifier with regularization so that you also penalize large weights ($\lambda\|\boldsymbol{w}\|_2^2$). Plot the average training and test errors for at least 5 different values of regularization parameter λ.