Machine Learning, Fall 2020: Project 1

Your name here

Header: Operating System: Mac OS Programming Language: Resources citation: Machine Learning by Stanford University, Coursera.
List the major resources you used to complete this project and the programming language you used.

You may use any programming language you like (Matlab, C++, C, Java...). All programming must be done individually from first principles. You are only permitted to use existing tools for simple linear algebra such as matrix multiplication/inversion. **Cite any resources that were used.**

In this project you will practice the basics of Machine Learning Classification by creating a K-NN classifier for classification on two data sets, using the Perceptron for performing classification on a data set, and by creating a Linear Regression model for performing regression on a data set. You will also learn good practices for how to describe, evaluate, and write up a report on the model's performance.

It is expected that your project report may require 2 pages per data set if you are good about making interesting figures and making them not too large, or 3-4 pages if your figures are big. The LaTex that generated this page is available here: https://www.overleaf.com/read/tvsnpcfgyxfc. **Please submit a pdf file created using this LaTex template and your code for project 1.**

Data sets: The project will explore four data sets, the famous MNIST data set of pictures of handwritten numbers, a data set that explores the prevalance of diabetes in a Native American tribe named the Pima, a data set concerning flower type recognition, and a data set that examines student achievement in secondary education in two Portuguese schools. You can access the data sets here:

1. https://www.kaggle.com/c/digit-recognizer/data
2. https://www.kaggle.com/uciml/pima-indians-diabetes-database
3. https://archive.ics.uci.edu/ml/machine-learning-databases/iris/
4. https://archive.ics.uci.edu/ml/machine-learning-databases/00320

1 K-Nearest Neighbors: MNIST and Pima Indians

For the MNIST data set and the Pima data set, you must create a K-NN classifier from scratch that uses the training data to build a classifier, and evaluate and report on the classifier performance. **Do NOT use machine learning packages for the KNN portion of the assignment. You are only permitted to use existing tools for simple linear algebra.**

1. **(10 points)** Describe the data and some simple visualizations (for images, a few examples from each category; for other data, perhaps some scatter plots or histograms that show a big picture of the data). Describe your training/test split for K-NN and justify your choices.

2. **(5 points)** K-NN is a very clear algorithm, so here describe any data pre-processing, feature scaling, distance metrics, or otherwise that you did.

3. **(10 points)** Show the accuracy of your algorithm — in the case of the Pima data set, show accuracy with tables showing false positive, false negative, true positive and true negatives. For the Pima data set, use three different distance metrics and compare the results.

   In the case of the MNIST digits show the complete confusion matrix. Choose a single digit to measure accuracy and show how that number varies as a function of K.

4. **(5 points)** Describe the run-time of your algorithm and also share the actual "wall-clock" time that it took to compute your results.

2 Perceptron: Iris flower classification

This version of the iris data set contains 2 classes of 50 instances each, where each class refers to a type of iris plant. The task is to select if a given flower is Iris-setosa or Iris-virginica. Evaluate your Perceptron implementation on this version of the iris data set that is linked on the syllabus with 10-fold-stratified-cross-validation. **Do NOT use machine learning packages for the Perceptron portion of the assignment. You are only permitted to use existing tools for simple linear algebra.**

1. **(5 points)** What happens when the learning rate is 0.00005, 0.001, and 0.005?
2. **(10 points)** Does the algorithm converge? Plot the classification accuracy for each learning rate from 1 to 20 training epochs.
3. **(10 points)** Come up with a confidence metric in your classification. (For example come up with an activation function that might correspond to confidence.) Create a scatter plot for confidence vs classification result for all instances with learning rate 0.00005.
4. **(10 points)** Is this data set linearly separable? Justify your answer with a scatterplot. Explain why and how you created this scatterplot.

3 Linear Regression: Student Performance in Portuguese class

Please use the Portuguese data set (`student-por.csv`) in the provided link for this assignment. This data set contains 649 instances and 30 features. Write a linear regression model from scratch and use it on this data set to predict the value for the final variable `G3`, the final grade for each student. **Do NOT use machine learning packages for the Linear Regression portion of the assignment. You are only permitted to use existing tools for simple linear algebra.**

1. **(10 points**) Some of the variables in this data set are categorical and some of them are numeric. How can we encode the categorical variables for the linear regression process? Please describe your approach to encoding categorical values and apply it to the data set in your code.

2. **(5 points)** Experiment by using different groups of features during training. What features work well in predicting a student's final score? What features work poorly? Why might you use or not use certain features? Calculate mean squared error scores for your linear regression model using at least three different groups of features, and compare the performance of the feature groups with each other.

3. **(5 points)** Perform linear regression using all available features. Use mean squared error to report the ability of your model to fit to the data. How does this approach compare to the groups of features you selected?

4 Parameter Estimation: MLE and MAP estimates

If X (e.g. packet arrival density) is Poisson distributed, then it has pmf

$$P(X|\lambda) = \frac{\lambda^X e^{-\lambda}}{X!}$$

1. **(5 points)** Show that $\hat{\lambda} = \frac{1}{n}\sum_i X_i$ is the maximum likelihood estimate of λ and that it is unbiased (that is, show that $\mathbb{E}[\hat{\lambda}] - \lambda = 0$).

2. **(5 points)** Recall that the Gamma distribution has pdf:

$$p(\lambda|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}, \quad \lambda > 0$$

   Assuming that λ is distributed according to $\Gamma(\lambda|\alpha,\beta)$, compute the posterior distribution over λ.

3. **(5 points)** Derive an analytic expression for the maximum a posteriori (MAP) estimate of λ under a $\Gamma(\alpha,\beta)$ prior.

5 Extra Credit: Comparison KNN and the Perceptron

Run your implementation of KNN and the Perceptron on a classification data set of your choosing and compare the results. A variety of sample data sets are available at the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/index.php).

1. **(10 points)** Describe the data set you used and compare the accuracy of the algorithms - display the accuracy for both models with tables showing false positive, false negative, true positive and true negatives. Which algorithm is better suited for this data set? Why might you choose one over the other?