Computer Vision

SSI Pattern Recognition

Homework 2

Désiré Sidibé
dro-desire.sidibe@u-bourgogne.fr

Linear Classification

I. Introduction

In this homework, we will implement logistic regression and linear discriminant analysis.
You shall submit a clearly written and commented report as well as your own code.

II. F*!& Spams !

We want to build a classifier to filter spam emails. We will use a dataset which contains a training set and a test set of, respectively, 3065 and 1536 emails. The data is given is the file `spamData.zip`. Ecah email has been processed and a set of 57 features were extrated as follows:

- 48 features, in $[0, 100]$, giving the percentatge of words in a given message which match a given word on a predifined list (called vocubulary). The list contains words such as "business", "free", "george", etc.
- 6 features, in $[0, 100]$, giving the percentatge of characters in the email that match a given character on the list. The characters are ; ([! $ #.
- Feature 55: the average length of an uninterrupted sequence of capital letters.
- Feature 56: the length of the longest uninterrupted sequence of capital letters.
- Feature 57: the sum of the lengths of uninterrupted sequence of capital letters.

1. What are the *max* and *mean* of the average length of uninterrupted sequences of capital letters in the training set?
2. What are the *max* and *mean* of the lengths of the longest uninterrupted sequences of capital letters in the training set?
3. Before training a classifier, we can apply several preprocessing methods to this data. We will try the following ones:
   (a) Standardize the columns so they all have mean 0 and unit variance
   (b) Transform the features using $\log(x_{ij} + 0.1)$, i.e. add 0.1 to each feature for every example and take the $\log$. We add a small number to avoid taking $\log$ of zero !
   (c) Binarize the features using $\mathbb{I}(x_{ij} > 0)$, i.e. make every feature vector a binary vector.

   For each version of the data, i.e. using a different preprocessing, fit a logistic regression model.

- use cross validation to choose the regularization parameter.
- report the mean error rate on the training ad test sets.
- what is the best preprocessing strategy? why?

4. Compare with the results of a Naive Bayes classifier.

III. Cats & Dogs !

In this part, we want to recognize dogs and cats. The files `catData.mat` and `dogData` contains respectively 80 examples of cat and dog images. Figure 1 show two images of a cat and a dog.

Figure 1: Example of a cat and a dog image.

In the previous task, digits recognition, we only had binary images. Now, we have more content in the images and we shal use this in building a classifier.
Your task is to

1. Represent each training example using a 'useful' feature descriptor. You should explain your choice for a descriptor.
2. Train a SVM to distinguish cats from dogs using your selected features. You should apply cross-validation.
3. Submit a report describing each step and your code.

Important: Clearly explain everything!