Assignment 4

Machine Learning, Summer term 2014, Ulrike von Luxburg

To be discussed in exercise groups on May 12-14

Exercise 1 (Rewriting the Fisher criterion for LDA, 2 points) Show that the Fisher criterion

$$J(w) = \frac{\langle w, m_+ - m_- \rangle^2}{\sigma_{w,+}^2 + \sigma_{w,-}^2}$$

can be rewritten as

$$J(w) = \frac{\langle w, C_B w \rangle}{\langle w, C_W w \rangle}.$$

See the lecture notes for the meaning of m_+, m_-, $\sigma_{w,+}^2$, $\sigma_{w,-}^2$, C_B and C_W.

Read `prepare04.pdf` (available on the course webpage) for an example of how to perform linear discriminant analysis (LDA) in MATLAB.

Exercise 2 (Digit classification with LDA, 1+2 points) We want to use LDA for written digit classification on the USPS dataset from Assignment 1 (do not forget to convert the data type from `uint8` to `double`).

(a) Use `ClassificationDiscriminant.fit` to train a LDA classifier for digit 2 versus 9. Use `imagesc` (see Assignment 1) to plot the learned weights in a 16x16 image. Here you do not need to change the colormap.

(b) Use your learned weights(`.Linear`) and the constant weight(`.Const`) to predict labels of the test data for digits 2 and 9. Report the 0-1 loss of your prediction and compare it with the performance of the kNN classifier for $k = 5$.

Multiclass classification: So far, all classification algorithms we have seen were designed to deal with two class (binary) classification problems. In many real world applications (e.g. digit recognition), our data has several classes. A common approach to solve such problems is to build a multiclass classifier from several binary classifiers. Assume we have k classes $C_1, \ldots, C_k$. Usually we follow one of the two strategies:

- One vs. All: A single classifier is trained per class to distinguish that class from all other classes. Prediction is then performed by predicting using each binary classifier, and choosing the prediction with the highest confidence score (in LDA, $w_i x + b_i$ is the score for class i).

- One vs. One: For each pair of classes we construct a binary classifier ($c(c-1)/2$ classifiers in total). Usually, classification of an unknown pattern is done according to the maximum voting, where each of $c(c-1)/2$ classifiers votes for one class.

Exercise 3 (Multiclass classification with LDA, 2+3 points)

(a) Implement the One vs. All strategy for LDA classifier to classify digits 1, 2, 3 and 4 from USPS dataset (use the complete dataset — available on the course webpage). Report the test error.

(b) Implement the One vs. One strategy for LDA classifier to classify digits 1, 2, 3 and 4 from USPS dataset (use the complete dataset — available on the course webpage). Report the test error.

Exercise 4 (Complexity of multiclass classification, 2+1 points) You have a multiclass classification problem with n training points and k classes. Assume that each class has the same number of training points. A binary classifier myClassifier is given which requires $c(m_1^2 + m_2^2)$ operations to learn the classifier (where m_i is the number of training points in class i and c is a constant).

(a) Compare the training time of multiclass classification with *One vs. All* and *One vs. One* strategies using myClassifier. Which strategy is faster?

(b) Does the situation change if myClassifier only requires $c(m_1 + m_2)$ operations for training?

Cross validation (m-fold): In many machine learning algorithms we need to choose parameters of our learning algorithm. The regularization parameter in ridge regression or the number of neighbors k in kNN classification are examples of such parameters.

In *m-fold cross-validation* the original sample is randomly partitioned into m (roughly) equally sized subsamples. Of the m subsamples, a single subsample is retained as the validation data for testing the model, and the remaining $m - 1$ subsamples are used as training data. The cross-validation process is then repeated m times (the folds), with each of the m subsamples used exactly once as the validation data. The m results from the folds then can be combined to produce a single estimation of the performance of the classifier under consideration.

Hence, we have the following scheme for choosing a parameter out of a finite set Λ of possible parameter values:

1. Split the data into m equally sized groups.

2. FOR $\in \Lambda$, $i = 1, \ldots, m$ DO

   (a) Select group i to be the validation set and all other $(m - 1)$ groups to be the training set.

   (b) Train the model with parameter on the training set, evaluate on the validation set and store the test error.

3. Select the parameter which performed best in the m folds (usually, one considers the average test error).

Exercise 5 (Parameter selection by the training error, 1 point) A naive approach for choosing the learning parameter would be to select the one which minimizes the training error. Explain why this is not a good idea in ridge regression and kNN classification.

Exercise 6 (Cross validation in ridge regression, 3 points) Use your ridge regression code (or the one available on the course webpage) and the training data in dataRidge.mat from Assignment 3. Choose the regularization parameter from $\in \{2^{-i} | i = -15, -14, \ldots, 7, 8\}$. Use 10-fold cross validation to select the regularization parameter with best performance (test error with L2 loss). To partition the data the MATLAB command cvpartition might be helpful.