FMAN-45 Machine Learning, Fall 2016
Assignment 3

Solve the problems and write down the solutions. If the assignment involves programming, download the code we provide and do the additional, required programming. Write a detailed report. All solutions, plots and figures should be in one *pdf. It should be possible to understand all material presented in the report without running any code.* Submit your solutions and code using your individual Moodle account *as two files (a pdf and a single archive with all the code) at* `http://moodle.maths.lth.se/course/` *by the deadline (note that there will be no extensions).*

1 Lagrange Multipliers in Regularized Optimization (10 points)

Using the technique of Lagrange multipliers, show that minimization of the regularized error function

$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_n)\}^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^q \tag{1}$$

where $q \geq 1$, is equivalent to minimizing the unregularized sum-of-squares error

$$E_D(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_n)\}^2 \tag{2}$$

subject to the constraint $\sum_{j=1}^{M}|w_j|^q \leq \eta$. That is, show that solving (1) for a fixed λ is equivalent to solving (2) for some η. Also discuss the relationship between the parameters η and λ, i.e., find a relationship $\eta = f(\mathbf{w}^*, \lambda)$, where $\mathbf{w}^*$ is a minimizer to (1) for a fixed λ.

2 Fisher Linear Discriminat vs. Logistic Regression (90 points)

The class separation criterion of Fisher's linear discriminant is given by

$$J(\mathbf{w}) = \frac{\mathbf{w}^\top \boldsymbol{S}_B \mathbf{w}}{\mathbf{w}^\top \boldsymbol{S}_W \mathbf{w}}, \tag{3}$$

where

$$\boldsymbol{S}_B = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^\top \tag{4}$$

$$\boldsymbol{S}_W = \sum_{i=1}^{2}\sum_{n\in C_i}(\mathbf{x}_n - \mathbf{m}_i)(\mathbf{x}_n - \mathbf{m}_i)^\top \tag{5}$$

where $\mathbf{m}_i$ is the mean vector of class i (C_1), $\mathbf{x}_n$ are samples (feature vectors) from the classes, and $\mathbf{w}$ is the weights vector.

For this assignment you are provided a MATLAB data file `data.mat` containing 4 datasets, all corresponding to binary classification tasks. Each dataset consists of a training set `train`i and a test set `test`i, as follows:

- `train1, train1_2, test1` Data generated from two bivariate Gaussians with identical covariances
- `train2, test2` Data generated from two bivariate Gaussians with different covariances
- `train3, test3` A digit classification task. The vectors representing digits are projected to a plane defined by two dimensions that capture most of the variability.
- `train4, test4` The same digit classificaction task as in the 3rd dataset, but now digit images are 64-dimensional vectors of 1s or 0s pixels in a 8×8 bitmap.

The 3rd dataset is derived from this representation by projecting onto a plane. The `.X` field of each variable is the representation of the points (each row is one datapoint), while the `y` field contains the class labels (0 or 1). You are provided the following MATLAB functions that implement both Fisher discriminant classification and logistic regression:

- `plotdata(train`i`)` For 2D data, plots the data and associated labels.
- `w = fisherdiscriminant(traini.X, traini.y)` Trains a Fisher discriminant linear classifiers and returns its parameters.
- `w = logisticreg(traini.X, traini.y)` Trains a logistic regression classifier by maximizing the likelihood using Newton's method, and returns its parameters.
- `boundary([w1 w2 ...], test`i`)` For 2D data, plots the data and the decisions boundaries of several logistic regression or Fisher discriminant sets of parameters in a single figure.
- `errorrate(w, test`i` )` Computes the error rate of a Fisher discriminant or logistic regression model.

Given the above, solve the following:

1. (10 points) Consider a simpler class separation criterion $\tilde{J}(\mathbf{w}) = \mathbf{w}^\top(\mathbf{m}_2 - \mathbf{m}_1)$ with the constraint that $\mathbf{w}^\top\mathbf{w} = 1$. Show that minimizing $\tilde{J}$ with respect to $\mathbf{w}$, using a Lagrange multiplier to enforce the constraint $\mathbf{w}^\top\mathbf{w} = 1$, leads to the result that $\mathbf{w} \propto (\mathbf{m}_2 - \mathbf{m}_1)$. When do you expect J in (3) to yield better class separation than $\tilde{J}$?
2. (20 points) Train logistic regression and Fisher discriminant classifiers on each of the first 3 training sets and report the test error on the corresponding test set (6 numbers). For each dataset plot on the same graph the test data and the decision boundaries corresponding to logistic regression and Fisher discriminant methods (3 plots).
3. (20 points) The performance of the two classifiers is similar on the data sets in the previous task. Would you expect the performances to be similar for all datasets sampled from class conditional distributions that are Gaussians with equal covariances?

4. (20 points) Training set `train1_2` is identical to `train1` except that it has an extra training point. Train the logistic regression and the Fisher discriminant on `train1_2` and compare the error rates on `test1` with those achieved by models trained on `train1`. For each method explain why the error rates change or not change with the addition of a single training point.

5. (20 points) Train logistic regression and the Fisher discriminant on the full 64-dimensional representation of the digits and report the error rates (`train4` and `test4`). Compare the error rates with those achieved on the reduced 2D representation (3rd dataset). Is the multivariate Gaussian assumption reasonable for the full representation of the digits? How sensitive are logistic regression and Fisher discriminant classification to a Gaussian assumption for the data in the respective classes?