Práctica 1: Classification

Data Processing
Academic year 2013–2014

6 de diciembre de 2013

Introduction

In this lab session we will study logistic regression and support vector machines using artificially generated data. After that, we will analyze the behavior of these classifiers using a real dataset.

1. Analysis of a bidimensional synthetic problem

Load the data contained in file `'datosP1.mat'`. This file contains two variables, `x` and `y`. The first one is an observation matrix. The second one contains the class label (0 or 1) corresponding to each observation.

1.1. Logistic regression. Linear classifier

First, we will analyze the behavior of a logistic regression model.

1. Visualize the observations on the plane, using a different marker for the data from each class.
2. Using function `glmfit`, fit a logistic regression model to the data.
3. Using function `contourf`, visualize the posterior probability map for class 1.
4. Determine the training and validation error rates.
5. Determine the training and validation data likelihoods for the estimated model.

1.2. Logistic regression. Nonlinear classifier

In the following, we will analyze the behavior of a non-lineal classifier based in a polynomial logistic regression model.

1. Using function `glmfit`, fit a logistic regression model with polynomial terms up to degrees 1, 2 and 3.
2. Using `contourf`, visualize the estimated posterior probability map for class 1.
3. Compute the training and validation error rates.
4. Determine the training and validation data likelihoods for the estimated model.

1.3. Máquina de vectores soporte

We will evaluate the performance of a classifier based on Support Vector Machines (SVM).

We will use an SVM with Gaussian kernels, taking kernel width $\sigma = 0{,}5$.

1. Using function `svmtrain`, train an SVM with the training data, visualizing the decision boundary and taking $C = 0.1$.
2. Exploring different values of parameter C, visualize the training and validation error rates, as a function of C, and choose a specific value according to the validation error.

1.4. Final evaluation

Choose, according to the validation error, the classifier providing the best performance, among all those analyzed in the sections above, and compute the test error rate.

2. Classification with real data

We will apply logistic regression an SVM to a real multidimensional dataset. The data visualization in the input space is no longer possible, but performance evaluation can be carried out using the error rates in any case.

We will work with the `cancer_dataset`.

2.1. Data preparation

In order to ensure that all variables have a similar scale, data normalization is generally advised.

The following code fragments carries a linear transformation making the observations in the training set have zero mean and unit sample variance. The same transformation must be applied to validation and test data:

```
mx = mean(xTrain); stdx = std(xTrain);
xTrain = (xTrain - ones(nTrain,1)*mx)./(ones(nTrain, 1)*stdx);
xVal = (xVal - ones(nVal,1)*mx)./(ones(nVal,1)*stdx);
xTest = (xTest - ones(nTest,1)*mx)./(ones(nTest,1)*stdx);
```

We will use the normalized data all along the rest of the lab exercise.

2.2. Classification with Support Vector Machines

In this section we will train a SVM with Gaussian kernels. We will use a validation dataset to determine the values of the hyperparameters σ and C.

1. Train the SVM for different values of σ ranging from -0.02 to 8 and values of C between 0.001 and 1000. Compute the training and validation eror rates in each case, and represent graphically (e.g., by means of function `contourf`) such error rates.

2. Compute the values of the kernel width, σ, and parameter C minimizing the validation error rate.
3. Evaluate the classifier performance using the test data, for the selected hyperparameter values.