Project 2
Binary Classification

Data

Input to classifier is 2-D point, output is one of two classes (class 0 is called FALSE, class 1 is called TRUE)

Training data consist of 200 points, coordinates are given in xtrain.txt; these points are labelled with classes (in ctrain.txt)

Test data consist of 6831 points, coordinates are given in xtest.txt;

- For each test point, given its probability of appearance (ptest.txt)
- For each test point, given its probability of belonging to class 1 (c1test.txt)

Evaluation

Use the training data to learn a classifier, and then classify each test point as belonging to FALSE or TRUE

Calculate the following criterion

$$ErrorRate = \sum_{x \in \{x \text{ is classified as } FALSE\}} p(x)p(T|x) + \sum_{x \in \{x \text{ is classified as } TRUE\}} p(x)(1 - p(T|x))$$

where $p(x)$ is the probability of appearance (ptest.txt) and $p(T|x)$ is the probability of belonging to class 1 (c1test.txt)

Basic requirements (1)

Use linear regression with basis functions, you can choose any basis functions as you like

Calculate the least squares solution; then add quadratic regularization (i.e. ridge), test different choices of λ; present the test performance

Use cross validation to find out an optimal set of basis functions; for example, you can select from polynomials of different orders, and find out which order performs the best using cross validation; verify whether your choice is correct using the test data

Basic requirements (2)

Use logistic regression with basis functions, you can choose any basis functions

Calculate the maximum likelihood solution

Use the Bayesian information criterion to find out an optimal set of basis functions, verify whether your choice is correct using the test data

Hint: $p(\mathcal{D}|w_{MAP})$ can be estimated by

$$\prod_{x \in \{Training\ data\ set\}} p(x, c(x)|w)$$

where $c(x)$ is the labelled class (ctrain.txt); according to logistic regression, $p(x, c(x)|w) = \sigma(w^T \phi(x))$ if $c(x) = 1$ and $p(x, c(x)|w) = 1 - \sigma(w^T \phi(x))$ if $c(x) = 0$

Plus

If you fulfill the basic requirements, you can get at most 7/10

Try to do the following to get additional score (at most 3/10)

- Implement gradient descent or Newton-Raphson method by yourself, and use it in your experiments (+0.5~1)
- Use lasso regression (+0.5)
- For linear regression: Calculate model evidence in the Bayesian framework, and perform model selection accordingly, verify your choice using the test data (+1)
- Use Fisher's LDA (+0.5)
- Use Bayesian logistic regression (+0.5)
- Try to divide the training data into multiple classes (e.g. using k-means), and train a multi-class classifier accordingly, use it on the test data, and convert the multi-class results into binary classes; show the performance (+1.5)

Report

.pdf is best, .doc and .ppt are acceptable

English is encouraged

List all references

Source code is a plus