COMP24111 EX3 Report

Boyin Yang 10071239

Nov 27 2016

Part 1 Implementation for discrete input attribute values

1. In NB training function, I use *Attribute* and *LabelSet* as input, *px* and *pc* as output, which stand for probability of every different attributes with different labels, that is, conditional probabilities (eg. matrix rely on av7_c3 is 7*3*57) and probability of every different classes, that is prior probability respectively (eg. matrix rely on av7_c3 is 3*1). For the first matrix, I count the number of different attribute value in one condition with same label by 2300 turns, which dived the total number of attribute value with same label. For the second matrix, count the total number of attribute value with same label, which divide the total number of every attributes.

2. In NB test function, I use *px, pc, testAttributeSet* and *validLabel* as input, *predictLabel* and *accuracy* as output. First, get the predict probability of each class of each sample, using the max probability to define the sample class, that is predict label. Compare predict label and valid label, count the right number. Finally, calculate the accuracy.

Part 2 Bonus · Principal Component Analysis (PCA)

1. Benefit: data compression

2. Reduce the dimension of the data from n-D to k-D: Find a direction onto which to project the data so as to minimize the projection error.

3. Using PCA:
    - 3.1 Data preprocessing: mean normalization, if necessary, use feature scaling;
    - 3.2 Compute ‘covariance matrix’ Sigma;
    - 3.3 Compute ‘eigenvectors’ of matrix Sigma;

4. Reconstruction from compressed representation which may lose a little bit accuracy.