SSIE 637 Advanced Topics in Healthcare
Assignment 04
Fall 2016

Instructor: *Daehan Won*

Due 11:59 PM on 12/05/2016

Note: Write down your answer briefly. Specify all parameters when you use the computationally implemented algorithms. The last columns in all used data set indicate the class information

Part 1 - Question 1 (5%)
Consider constructing a certain classifier in a situation with 500 features total. 50 of them are truly informative about class. Another 50 features are direct copies of the first 50 features. The remaining 400 features are not informative. Assume there is sufficient data to reliably assess how useful features are, and the feature selection methods are using good thresholds.
(1) How many features will be selected by mutual information filtering method?
(2) How many features will be selected by a wrapper method?

Part 1 - Question 2 (5%)
Consider k-fold cross-validation. Let's consider the trade-offs of larger or smaller k. With a higher number of folds, how the estimated error will be? specify your answer and supporting reason within 2 sentences.

Part 2- Question 1 (20 %)
The 'colon' data measures the transcript level of 2000 genes for patients with or without colon cancer. Feature selection can potentially provide some insight on what differentiates tissues that come from different tumor types, or what differentiates tumor from non-tumor tissue. Apply filter feature selection approach to select most informative features. Once you selected features, report the average training and testing accuracies, precision,recall and F-measure for 5-fold cross validation. You can use any kinds of classification method.

Part 2- Question 2 (10 %)
Let apply a wrapper method 'sequential forward election (SFS)' to the 'iris' data set.

- First, split the entire data set into training (70 %) and testing (30 %) randomly.
- Second, pick 3 features out 4 using SFS. Note that employ KNN (k=5) as a main classifier. In the selection procedure, you should use training accuracy obtained as a selection criterion.
- Third, calculate testing accuracy, precision, recall and F-measure.
- Fourth, repeat this process (First to Third) 5 times and report average accuracy, precision, recall, and F-measure.

Part 3- Question 1 (30 %)
Solve the staff scheduling problem in Lecture note 11 (page 27). Make a formulation of the problem with indicating decision variables, objective function and constraints. Also report the solution with sensitivity analysis.

Part 3- Question 2 (30 %)
An insurance company desires to enter the health-care market and offer its potential customers both a staff model Health Maintenance Organization (HMO) and commercial indemnity insurance. The company is

deciding how to allocate its marketing efforts between those options to maximize its profits. The analysts have estimated that the company will realize a profit of $1,100 per enrollee from the HMO, and $600 per enrollee from commercial plans. Furthermore, for the coming year the company is forced to rely on its present resources in terms of sales force. The administrative support of the HMO will take 200 hours, and the commercial administration will take, on average, 400 hours; currently, the company can allocate 1.6 million hours to sales. To break even, the insurance company requires that the contribution margins (contribution margin is sales revenue less variable costs; it is the amount available to pay for fixed costs and then provide any profit after variable costs have been paid) for enrollees must exceed $1.5 million. The estimated contribution margins are $400 and $300, for HMO and for commercial insurances enrollees, respectively. With a limited number of physicians participating in the staff model HMO at the present time, the number of enrollees that HMO can handle at most 5,000.
(1) Solve the problem as linear programming problem.
(2) If the solution is not integer, try to find the optimal integer solution.

Part 4 - Optional (20%)
Given 'ecoli_classification' dataset, consider following linear programming (LP) to build a linear discriminant function for data classification.

$$
\begin{aligned}
\text{minimize} \quad & \sum_{i=1}^{n} d_i \\
\text{subject to} \quad & w_0 + \sum_{j=1}^{m} w_j x_{ij} - d_i \leq -1, \forall i \in G_1 \\
& w_0 + \sum_{j=1}^{m} w_j x_{ij} + d_i \geq 1, \quad \forall i \in G_2 \\
& w_0, w_j \text{unrestricted} \qquad j = 1, ..., m \\
& d_i \geq 0, \qquad \forall i \in G_1 \cup G_2
\end{aligned}
$$

This model is a modification MSD(Minimize the Sum of Deviation) in order to prevent trivial solution. Suppose you have two classes; G_1 and G_2. x_{ij} is j-th feature of i-th sample and w_j is decision variable to construct linear discriminant function for future prediction. Once you solve the problem, you can calculate classification score c_t for future prediction of unseen sample $\mathbf{x}_t = \{x_{t1}, x_{t2}, ...x_{tj}\}$ such as $c_t = w_0^* + \sum_{j=1}^{m} w_j^* x_{tj}$ where w_0^* and w_j^* are obtained solution from the above LP. A new sample $\mathbf{x}_t$ is classified as G_1 if $c_t < 0$, as G_2 if $c_t > 0$ and not classified if $c_t = 0$. Apply 5-fold cross validation and report average accuracy, precision, recall, and F-measure.