Optimal Screening methods in classification of pairs of covariates

• Consider a pair of covariates \( X = (X_1, X_2) \), each observation \( x \) having a true class label in 0, 1.

• Let \( Y \) be a binary random variable assuming value 1 if \( X \) has class 1 and value 0 otherwise.

• The optimal classification region of size \( \alpha \) is

\[
C_\alpha = \left\{ x \in \mathbb{R}^2 : \frac{P(Y = 1|D)p(x|Y = 1, D)}{P(Y = 1|D)p(x|Y = 1, D) + P(Y = 0|D)p(x|Y = 0, D)} > k \right\}
\]

where \( k \) such that \( P(x) \in C_\alpha = \ldots \).

• Consider \( Y \sim \text{Ber}() \), \( \theta \in (0, 1) \) and assume a Beta priori distribution for \( \theta \) \( \sim \text{Beta}(a, b) \). The posterior distribution of \( \theta \) is \( \theta | D \sim \text{Beta}(a + n_1, b + n_2) \).

• The predictive density of a randomly selected individual to be a success (\( Y = 1 \)) and the predictive probability of a randomly selected individual not to be a success (\( Y = 0 \)) are, respectively

\[
P(Y = 1|D) = \frac{n_1 + a}{n_1 + a + b} \quad \text{and} \quad P(Y = 0|D) = \frac{n_2 + b}{n_1 + a + b}
\]

with \( n \) is the number of individuals in the sample for which \( Y = i, i = 0, 1 \).

• The predictive density of a future observation in class \( Y = i, p(x|Y = i, D) \), is estimated using a Bayesian Nonparametric method by considering a bivariate mixture of Poisson trees prior (BMPT). Denote the bivariate data from the success group and from the success group by \( x_{11}, x_{12}, x_{13} \sim G_1 \) and \( x_{21}, x_{22}, x_{23} \sim G_2 \) respectively. The data for each group \( i, i = 0, 1 \) are modelled as

\[
x_{11}, x_{12}, x_{13} \sim G_1 \quad \text{and} \quad x_{21}, x_{22}, x_{23} \sim G_2
\]

where \( M \) is the maximum level of the partition to be updated, \( \pi^0, \pi^1, \pi^2 \) is a set of partitions of \( \mathbb{R}^2 \) indexed by \( \Sigma \) and \( \Sigma, A^1 \) is a family of non-negative vectors controlling the variability of the process indexed by \( \Sigma \). The \( PT \) is centered around a \( \Delta_0(\mu, \Sigma) \) distribution.

• The predictive densities \( \delta = \{P(Y = 1|X = x, D) \} \), \( \epsilon = \{P(Y = 1|X = x, D) \} \), \( \gamma = \{P(Y = 1|D) \} \) and \( \nu = \{P(x \in C_\alpha|D) \} \) are called operating characteristics (OC) of the screening problem.

• Computations

\( C_\alpha \) does not have a closed form, so the screening region boundaries must be approximated.

The computation implemented was the following:

1. Build a fine grid \( W = \{(x_1, x_2) : x_1, x_2 \in \mathbb{R} \} \) such that \( P(X_1, X_2) \leq W = 1 \).

2. For each \( (x_1, x_2) \in W \) calculate \( P(Y = 1|X = (x_1, x_2), D) \).

3. For several values of \( \alpha \) form the sets \( C_\alpha = \{(x_1, x_2) : P(Y = 1|X = (x_1, x_2), D) > \alpha \} \).

4. Fit to the boundaries of each \( C_\alpha \) to approximate \( C_\alpha \) by \( \{x_1, x_2 : x_1, x_2 \in R \} \), where \( \lambda_1 \) is an interval of the form \( -\infty, l(x_1, x_2) \), \( l(x_1, x_2) \) depending on the shape of the \( C_\alpha \).

5. After each region is obtained, the OC are calculated numerically.

Method

Classification procedure (continuation)

• Consider a new individual, where \( P \) has \( m \) pairs.

• Let \( C_\alpha \) be the probability of success given that the \( j \)th profile belongs to \( C_\alpha \).

• The final classification of the individual is given by: \( \{C_\alpha \} \) (the indicator function)

\[
C = \{C_\alpha \}
\]

Applications

Data sets

• Prostate study: 49 prostate tumor samples (T) and 43 non-tumor samples (NT); one pair of genes considered as input for the screening classifier (\( m = 1 \)).

• Leukemia study: 47 ALL samples and 25 AML samples; three pairs of genes considered as input for the screening classifier (\( m = 3 \)).

• Breast study: 18 women who did not experience recurrence of the tumor (NR) and 34 who experienced the recurrence of the tumor (R); one pair of genes considered as input of the screening classifier (\( m = 1 \)).

Classifications results

The computation implemented was based on the marginalized version of the model the random probability measure \( G \) is integrable out. We used the function PTdensity implement in R package DPpackage. Here \( J = 7 \) and for \( i = 0, 1 \) we set \( a_0 = b_0 = 0.001 \) to yield diffuse prior distributions.

BMPT method

<table>
<thead>
<tr>
<th>Study</th>
<th>T</th>
<th>BT</th>
<th>S</th>
<th>E</th>
<th>E (BP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prostate</td>
<td>0.67</td>
<td>0.5319</td>
<td>0.5103</td>
<td>0.6735</td>
<td>0.7098</td>
</tr>
<tr>
<td>Leukemia</td>
<td>0.85</td>
<td>0.6486</td>
<td>0.5307</td>
<td>0.9734</td>
<td>0.1351</td>
</tr>
<tr>
<td>Breast</td>
<td>0.45</td>
<td>0.3519</td>
<td>0.2920</td>
<td>0.8149</td>
<td>0.1111</td>
</tr>
</tbody>
</table>

BOSc method

<table>
<thead>
<tr>
<th>Study</th>
<th>T</th>
<th>BT</th>
<th>S</th>
<th>E</th>
<th>E (BP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prostate</td>
<td>0.63</td>
<td>0.5267</td>
<td>0.8956</td>
<td>0.0931</td>
<td></td>
</tr>
<tr>
<td>Leukemia</td>
<td>0.86</td>
<td>0.6486</td>
<td>0.5675</td>
<td>0.9918</td>
<td>0.1985</td>
</tr>
<tr>
<td>Breast</td>
<td>0.42</td>
<td>0.3519</td>
<td>0.3128</td>
<td>0.8459</td>
<td>0.1269</td>
</tr>
</tbody>
</table>

Conclusion

The boundaries of screening region can be estimated by an appropriately chosen parametric function (linear, polynomial, spline), depending on the shape of the region. In this work, quadratic functions were considered.

The non parametric classifier achieves classification rates and OC similar to those obtained with the parametric approach.

References
