Application of Various Classification Techniques in Singapore Eye Study

Ming Gao
Prof: Shan Luo

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Contents

Singapore Eye Study

Around 3000 peoples 300 indexes:

- Basic information (age, height,...),
- Blood data (glucose, cholesterol,...),
- Eye data (myopia, blindness, sphere,...),
- Eye disease (cataract,...),
- Self-information (education, job, smoke, income,...),
- Main disease (heart attack).

Preprocessing

- Clean the data by removing or imputing missing entries,
- Split training and test data,
- Use One Hot Encoding to turn all the factors into dummy variables,
- Transform dependent variable=0 to represent a person got heart attack.

Dimension of training input:(2949, 339); dimension of test input:(327, 339);
length of training output:(2949); length of test output:(327).
Notation: y_i: a sample of dependent variable; x_i a sample of independent variable vector.

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

	czmi1	gender2	age	agegp2	agegp3	agegp4	agegp22	agegp23	agegp24	agegp25	agegp3	bpsys_f
887	1	1	58.55441478	1	0	0	1	0	0	0	4	99.5
1248	0	0	75.71800137	0	0	1	0	0	1	0	8	106
1918	0	0	54.00136893	1	0	0	1	0	0	0	3	107
3047	0	1	73.10335387	0	0	1	0	0	1	0	7	139.5
673	0	1	50.41204654	1	0	0	1	0	0	0	3	115.5
3013	0	0	78.65023956	0	0	1	0	0	1	0	8	180
3166	0	1	47.75633128	0	0	0	0	0	0	0	2	136.5
2211	1	0	66.69130732	0	1	0	0	1	0	0	6	136
2103	0	0	53.60711841	1	0	0	1	0	0	0	3	142
208	1	1	49.10882957	0	0	0	0	0	0	0	2	105
686	0	1	78.74606434	0	0	1	0	0	1	0	8	184
588	1	0	47.13483915	0	0	0	0	0	0	0	2	108
2296	1	1	58.88843258	1	0	0	1	0	0	0	4	135
1283	0	0	63.88227242	0	1	0	0	1	0	0	5	153

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

First Attempt

Logistic regression embracing all the predictors.
Suppose $y_i \sim \mathrm{Bin}(1, p_i)$, then $\mu_i \triangleq E(y_i) = p_i$.
Let $\mathbf{x_i}^T \beta = g(\mu_i) \triangleq \ln \frac{p_i}{1-p_i}$.
Hence

$$p_i = P(y_i = 1|\mathbf{x_i}) = \frac{1}{1 + e^{-\mathbf{x_i}^T \beta}} \tag{1}$$

Define

$$\begin{aligned} \mathrm{LOSS}(\beta) &= \frac{1}{n} \sum_{i=1}^{n} -y_i \ln p_i - (1 - y_i) \ln(1 - p_i) \\ &= \frac{1}{n} \sum_{i=1}^{n} -y_i(\mathbf{x_i}^T \beta) + \ln(1 + e^{\mathbf{x_i}^T \beta}) \end{aligned}$$

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Evaluation

MLE:

$$\hat{\beta} = \arg\min -\mathrm{Loglik}(\beta) \tag{2}$$

Prediction:

$$P(\hat{y}_i = 1|\mathbf{x_i}) = \frac{1}{1 + e^{-\mathbf{x_i}^T \hat{\beta}}} \tag{3}$$

Select a threshold c and forecast $\hat{y}_i = I(P(\hat{y}_i = 1|\mathbf{x_i}) > c)$.

Correct Classification Rate (CCR)

$$\mathrm{CCR} = P(y = \hat{y}) \tag{4}$$

after we set threshold $c = 0.5$.

CCR on the test dataset: 0.927 (NOT GOOD).

Why not work?

- Many predictors are noises,
- Time consuming,
- Overfitting.

Hence, we implement feature selection.

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Regularization Method

Estimator

$$\tilde{\beta} = \min_{\beta} \sum_{i=1}^{n} LOSS\left(y_i, f(\mathbf{x_i}, \beta)\right) + \lambda\phi(\beta) \tag{5}$$

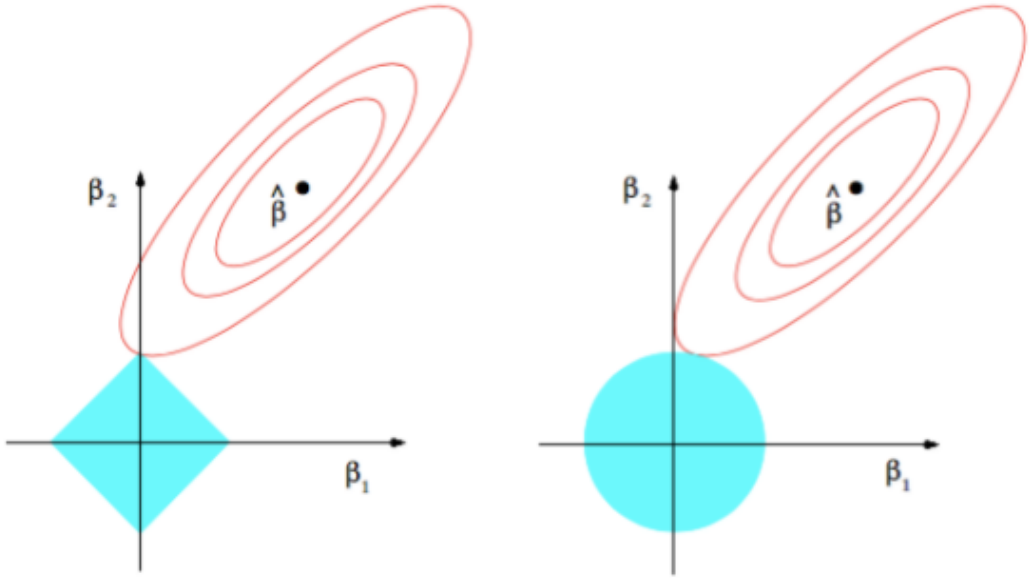
ϕ is the penalty function (regularization term). Its value is proportional to the complexity of the model.
$\lambda \geq 0$ is the tuning parameter. It is a hyper parameter, which should be determined before fitting the model.

Least Absolute Shrinkage and Selection Operator (LASSO)

$\phi(\beta) = |\beta|_1$

Suppose the design matrix X is orthonormal and y is under normality,

$$\beta_j^L = s[\hat{\beta}_j, \lambda] \triangleq sign(\hat{\beta}_j)(|\hat{\beta}_j| - \lambda)_+ \quad (6)$$



上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

Shortcoming

Predictors should not be too correlated, or else, LASSO tends to choose one of them and transform the others to zeros.
Severe problem!
Correlated predictors are very familiar due to one-hot-encoding.

Group LASSO (GL)

$\phi(\beta) = \sum_{j=1}^{J} \sqrt{p_j} |\beta_{G_j}|_2$

Here we divide {1...p} into J group $\{G_1, ..., G_J\}$. $p_j \triangleq \#\{G_j\}$.

$\beta_{G_j} = (\beta_k)_{k \in G_j}$

```
##   [1]   1   2   3   4   4   4   5   5   5   5   6   7   8   9  10  11  12
##  [18]  13  14  15  15  15  16  17  18  19  20  21  22  23  24  24  24  25
##  [35]  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40  41  42
##  [52]  43  43  44  44  45  45  46  46  47  47  48  48  49  49  50  50  51
##  [69]  51  52  53  54  55  55  55  55  55  56  56  56  56  56  57  57  57
##  [86]  57  57  58  58  58  58  58  59  60  61  62  63  64  65  66  67  68
## [103]  69  69  70  70  71  71  72  73  74  75  76  77  78  79  80  81  82
## [120]  83  84  85  86  87  88  89  90  91  92  93  94  95  96  97  97  97
## [137]  98  98  98  99 100 101 102 103 104 105 106 107 108 109 110 111 112
## [154] 113 114 115 116 116 117 117 118 119 120 121 122 123 124 125 126 126
## [171] 126 127 127 127 128 128 128 128 129 130 131 132 132 132 132 133 133
## [188] 133 134 135 136 137 138 138 139 140 141 142 143 144 145 146 147 148
## [205] 149 150 151 151 151 152 153 154 155 156 157 158 159 160 161 162 163
## [222] 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180
## [239] 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 196
## [256] 197 197 198 198 199 200 201 202 203 204 205 205 205 205 206 207 208
## [273] 209 210 211 212 213 214 214 214 214 215 215 215 215 216 217 218 219
## [290] 220 221 222 223 224 225 226 227 228 229 230 231 232 233 234 235 236
## [307] 237 238 239 240 241 242 243 244 244 245 246 247 247 248 249 250 251
## [324] 251 252 253 254 254 255 256 257 258 259 260 261 262 263 264 265
```

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Why it works?

Group LASSO is an intermediate between LASSO and l_2 penalty. Consider a case where there are two factors: a bi-dimension vector $\beta_1 = (\beta_{11}, \beta_{12})^T$ and a scalar β_2. We combine β_{11}, β_{12} in a group and β_2 another group.

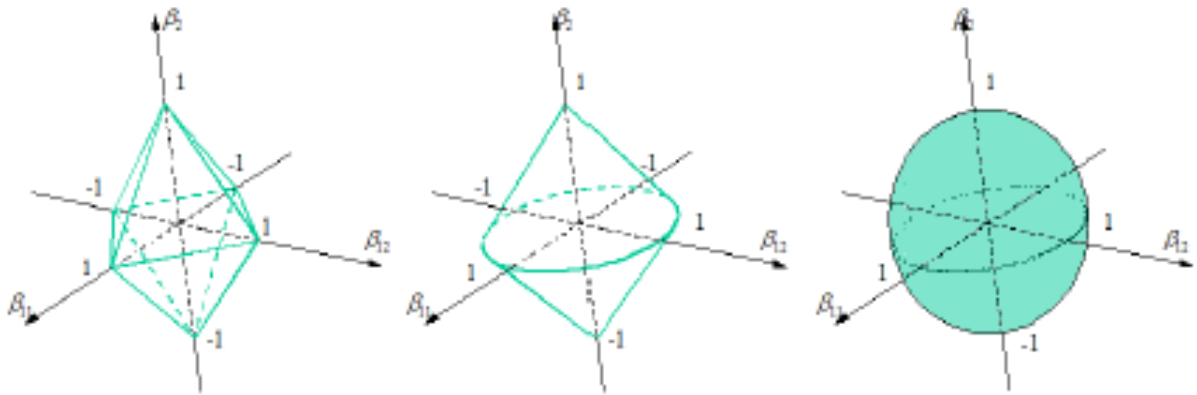


Figure: The l_1 penalty (left panels), GL penalty (central panels) and l_2 penalty (right panels).

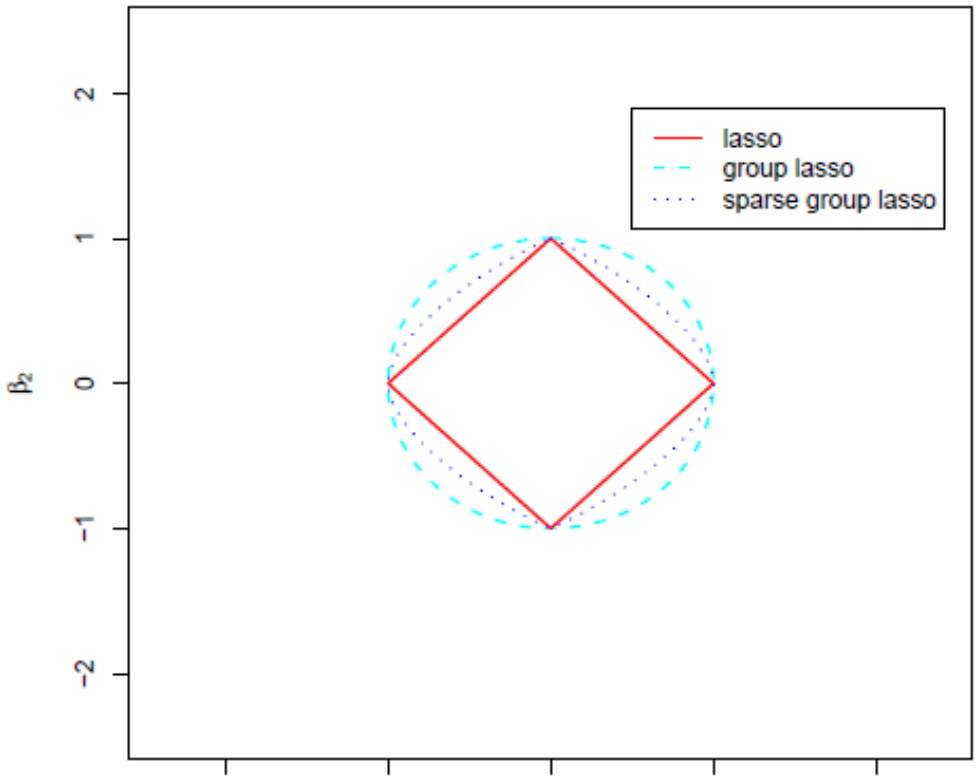
Shortcoming

The group lasso does not yield sparsity within a group. If a group of parameters is non-zero, they will all be non-zero, which may select too many predictors.

Sparse Group LASSO (SGL)

$$\phi(\beta) = \{(1-\alpha)\sum_{j=1}^{J}\sqrt{p_j}|\beta_{G_j}|_2 + \alpha|\beta|_1\}$$

Here $0 \leq \alpha \leq 1$



Selection of Tuning Parameter

We try to get the best model (λ) based on some criteria:

Hyper Parameter tuning

$$\lambda^* = \arg\min_{\beta} \mathrm{CRITERION}(\lambda) \tag{7}$$

Cross Validation through Deviance

In K-fold cross-validation, the original observations are randomly partitioned into K equal sized subsamples.
We fit the model with parameter λ to the other $K-1$ parts, giving $\tilde{\beta}^{-k}(\lambda)$ and compute its loss $\mathrm{LOSS}_k(\lambda)$ in predicting the k^{th} part.

CV_{dev}

$$\mathrm{CV}_{\mathrm{dev}}(\lambda) = \frac{1}{K}\sum_{k=1}^{K}\mathrm{dev}_k(\lambda)$$

$$\mathrm{dev}_k(\lambda) = \frac{-2}{n/k}\mathrm{Loglik}(\tilde{\beta}^{-k}(\lambda))$$

Cross Validation through Misclassification Error

CV_{ME}

$$\mathrm{CV}_{\mathrm{ME}}(\lambda) = \frac{1}{K}\sum_{k=1}^{K}\mathrm{ME}_k(\lambda)$$

$$\mathrm{ME} = \frac{1}{n/k}\{\#_i(p_i > 0.5 \ \& \ y_i = 0) + \#_i(p_i < 0.5 \ \& \ y_i = 1)\}$$

Shortcoming

- Time-consuming because we need to fit K models,
- Too generous.

Bayesian Information Criterion (BIC) & Extended BIC

BIC

$$\mathrm{BIC}(\lambda) = -2 \ln \mathrm{Loglik}(\hat{\beta}(\lambda)) + \mathrm{df}(\lambda) \ln n \tag{8}$$

EBIC

$$\mathrm{EBIC}_\gamma(\lambda) = -2 \ln \mathrm{Loglik}(\hat{\beta}(\lambda)) + \mathrm{df}(\lambda) \ln n + 2\gamma \mathrm{df}(\lambda) \ln p \tag{9}$$

The set of non-zero elements (support, or active set) is a unbiased estimator of df.

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Correct Classification Rate (CCR)

CCR

$$\text{CCR} = P(y = \hat{y}) \tag{10}$$

$$CCR = \sum_{m=1}^{M} P(y = m)P(\hat{y} = m|y = m) \triangleq \sum_{m=1}^{M} \rho_m CCR_m \tag{11}$$

Area under Receiver Operating Characteristic Curve (AUC)

Evaluating the tradeoffs between true positive rate (TPR, sensitivity) and false positive rate (FPR, 1-specificity), where $\mathrm{FPR} = P(\hat{y} > c | y = 0)$ and $\mathrm{TPR} = P(\hat{y} > c | y = 1)$.

AUC

$$\mathrm{AUC} = \int \mathrm{TPR}(c)\, \mathrm{dFPR}(c) = P(\hat{y}|_{y=1} > \hat{y}|_{y=0}) \tag{12}$$

Comparison

Table 6–1 A table of statistics of 12 models. Active represents the number of non zero predictors.

	λ	Active	CCR_train	AUC_train	CCR_test	AUC_test
CV_dev_L	0.004	63	0.968	0.937	0.960	0.846
CV_ME_L	0.011	14	0.964	0.911	0.963	0.852
BIC_L	0.010	15	0.964	0.913	0.963	0.851
EBIC_L	0.016	10	0.963	0.900	0.963	0.846
CV_dev_GL	0.006	37	0.965	0.920	0.960	0.839
CV_ME_GL	0.008	19	0.964	0.915	0.963	0.832
BIC_GL	0.009	17	0.964	0.913	0.963	0.834
EBIC_GL	0.015	9	0.963	0.904	0.963	0.841
CV_dev_SGL	0.003	74	0.967	0.940	0.960	0.840
CV_ME_SGL	0.006	38	0.966	0.925	0.960	0.846
BIC_SGL	0.013	10	0.964	0.907	0.963	0.840
EBIC_SGL	0.015	9	0.963	0.904	0.963	0.841

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Important Predictors

We choose SGL with EBIC as the final model. Here, p is the predicted probability that one doesn't get heart attack.

$$\ln \frac{p}{1-p} = (-1.473) + (0.067) * gender2 + (-0.078) * anti\_ht1 + (-0.467) * anti\_chol1 + (-0.185) * drugs\_others1 + (0.146) * chol + (0.006) * GFR\_EPI + (0.172) * smkyn2 + (0.904) * ang2$$

Table 6–3 Meanings of predictors selected. Bi represents binary variable, and con represents continuous variable.

	Variable's code	Meaning	Type	Range
1	gender2	gender	bi	1:female
2	anti_ht1	Anti-hyperstensive drugs	bi	1:yes
3	anti_chol1	Anti-cholesterol drugs	bi	1:yes
4	drugs_others1	Drugs - Others	binary	1:yes
5	chol	Blood Total Cholesterol 血总胆固醇	con	/
6	GFR_EPI	Glomerular Filtration Rate 肾小球	con	/
7	smkyn2	Have you ever smoked?	binary	1:no
8	ang2	Angina (self-reported history) 心绞痛	bi	1:no

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

Thank You!

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY