Big heart data

Yu Lu

March 6, 2013

The Problem

Data

- 两组患有心梗的2581个病人，一组为安贞医院1214个，另一组为朝阳+301医院1367个
- X: 37个术前指标+ 36个术后指标
- Y: 病人“无复流”状态

Goal

- 分析73个指标对“无复流”状态的影响
- 将数据随机分成5份，4份train和validation参数, 1份test模型预测的准确率。

Methods

Step1 预处理数据：合并两组病人，加入一个变量表示所在组；同时去掉Y（无复流状态）缺失的7个样本

Step2 对每个指标分别做logistic回归，选出相对显著的13个变量

Step3 填补这13个变量的缺失数据，离散变量采用随机填补，连续变量用均值填补

Step4 在选出的13个变量中进一步选择变量用于SVM（支持向量机）学习和预测。

通过枚举得到4个变量：所在医院、随机血糖、PCI术中2b3a、PCI术中钙拮抗剂, 他们SVM的预测准确率最高，约为12.4%

Result

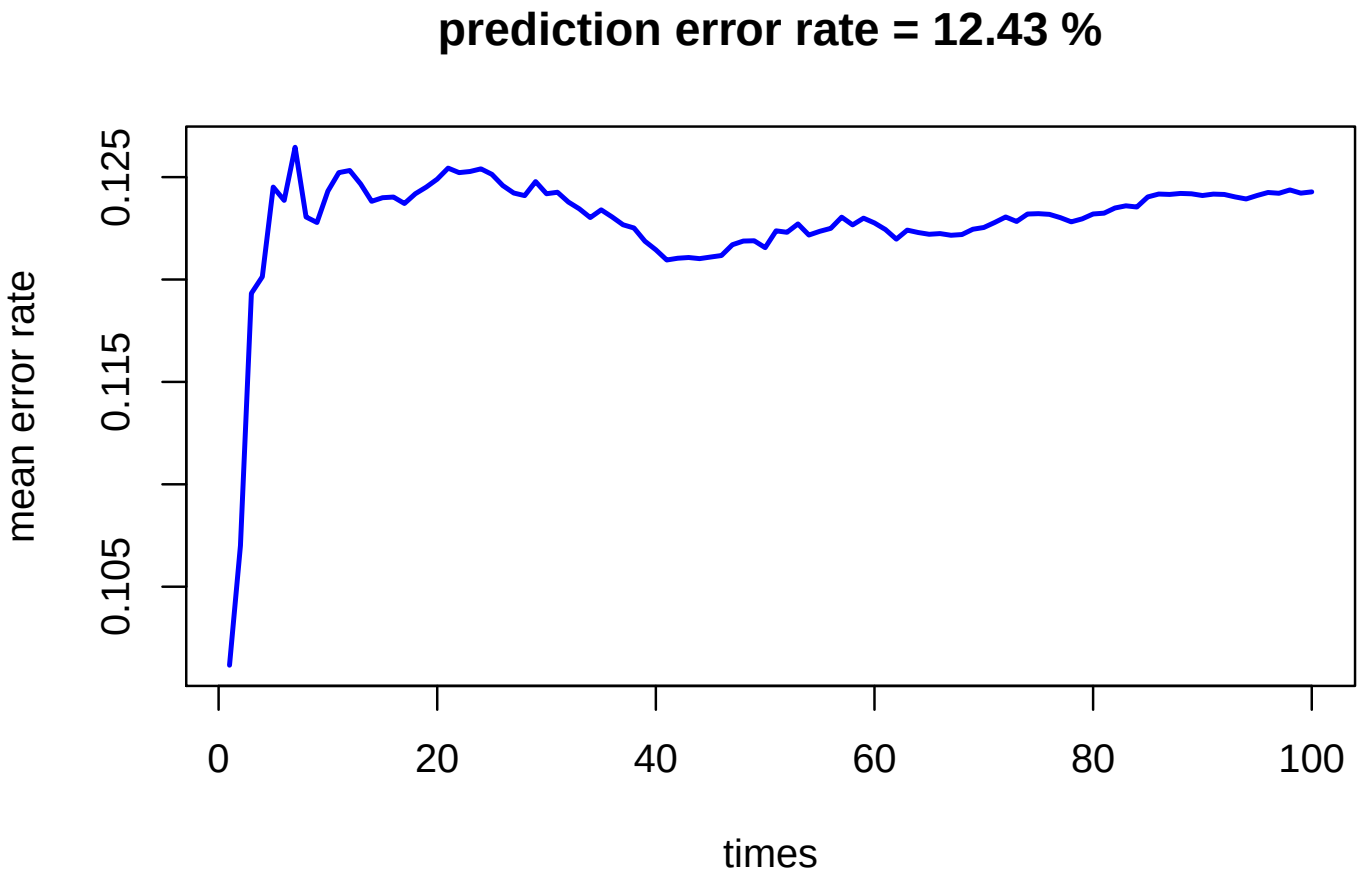

Figure: Result

Variable selection

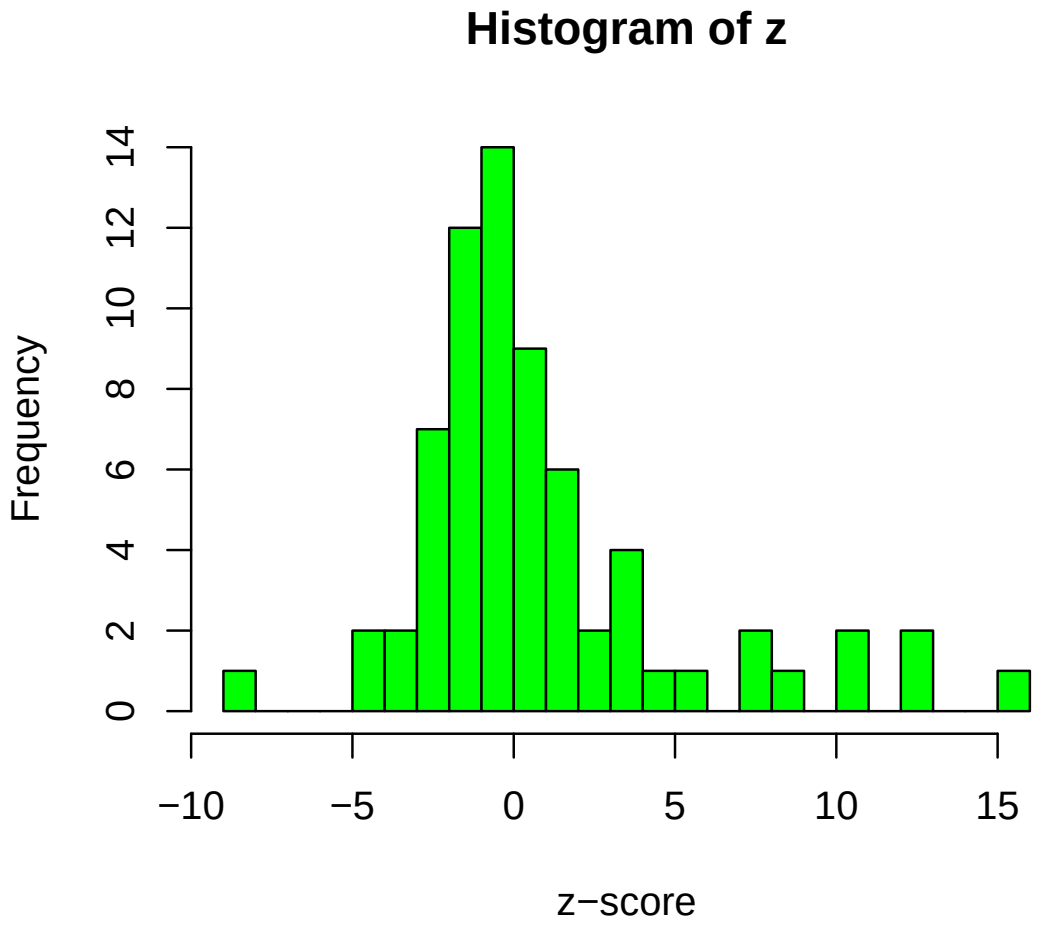

Figure: 边缘Logistic回归各变量的显著程度

Support Vector Machine

- Separable case: we can find a function f with $y_i f(x_i) > 0, \forall i$, and the optimization problem is

$$\max_{\beta,\beta_0,||\beta||=1} M, \quad \text{subject to } y_i(x_i^T \beta + \beta_0) \geq M$$

- Nonseparable case: the optimization problem is

$$\max_{\beta,\beta_0,||\beta||=1} M, \quad \text{subject to } \begin{cases} y_i(x_i^T \beta + \beta_0) \geq M(1 - \xi_i) \quad \forall i \\ \xi_i \geq 0, \ \sum \xi_i \leq \text{constant} \end{cases}$$

Support Vector Machine

- Nonlinear boundaries: 选一组基函数$h_m(x), m = 1, 2, \cdots, M$, SVM的输入变量为$h(x_i) = (h_1(x_i), h_2(x_i), \cdots, h_M(x_i))$, 相应的超平面是

$$\hat{f}(x) = h(x)^T \hat{\beta} + \hat{\beta}_0$$

在求解相应最优化问题中的lagrange dual function只和$h(x)$的内积有关，因此只需要给出一组内积。这里我采用了radial basis:

$$K(x, x') = \langle h(x), h(x') \rangle = exp(-\gamma ||x - x'||^2)$$

注：上述求解采用了SVM的R软件包‘e1071’

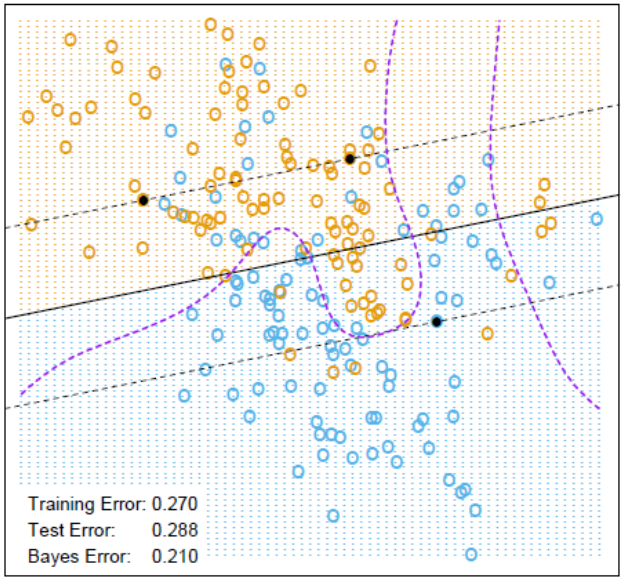

Figure: linear kernel

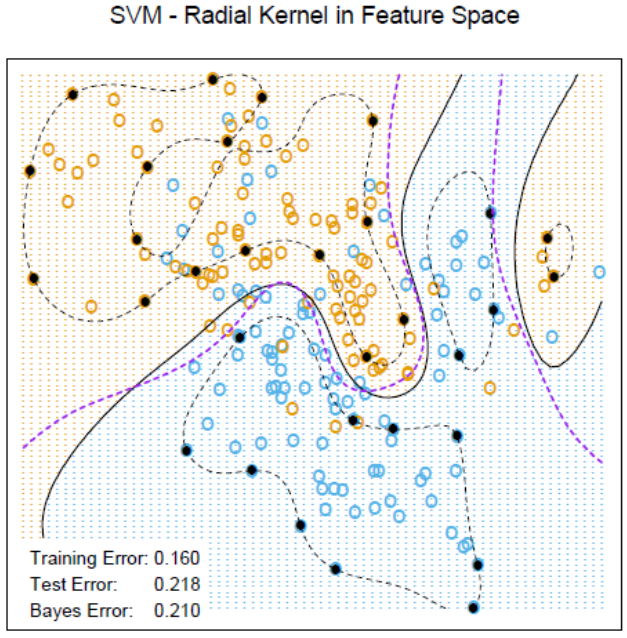

Figure: radial kernel