邻近点梯度法

王 尧

西安交通大学智能决策与机器学习中心
(Email: yao.s.wang@gmail.com)

2022. 5

Ridge回归

$$\hat{\beta}^{\text{ridge}} = \arg\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_2^2$$

- Close-form solution:

$$\hat{\beta}^{\text{ridge}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

- Gradient descent:

$$\mathbf{g}^t = \mathbf{X}^T(\mathbf{X}\beta^t - \mathbf{y}) + \lambda\beta^t$$

$$\beta^{t+1} = \beta^t - \mu^t\mathbf{g}^t$$

回顾：可分解函数

Suppose

$$f(x) = g(x) + h(x)$$

- g is convex, differentiable, $\mathrm{dom}(g) = \mathbb{R}^n$
- h is convex, not necessarily differentiable

If f were differentiable, then gradient descent update would be:

$$x^+ = x - t \cdot \nabla f(x)$$

Recall motivation: minimize quadratic approximation to f around x, replace $\nabla^2 f(x)$ by $\frac{1}{t} I$,

$$x^+ = \underset{z}{\mathrm{argmin}} \underbrace{f(x) + \nabla f(x)^T (z - x) + \frac{1}{2t} \|z - x\|_2^2}_{\bar{f}_t(z)}$$

典型例子：$\min_\beta \frac{1}{2} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1$

In our case f is not differentiable, but $f = g + h$, g differentiable. Why don't we make quadratic approximation to g, leave h alone?

That is, update

$$
\begin{aligned}
x^+ &= \operatorname*{argmin}_z \ \bar{g}_t(z) + h(z) \\
&= \operatorname*{argmin}_z \ g(x) + \nabla g(x)^T (z - x) + \frac{1}{2t}\|z - x\|_2^2 + h(z) \\
&= \operatorname*{argmin}_z \ \frac{1}{2t}\big\|z - \big(x - t\nabla g(x)\big)\big\|_2^2 + h(z)
\end{aligned}
$$

$\frac{1}{2t}\big\|z - \big(x - t\nabla g(x)\big)\big\|_2^2$ stay close to gradient update for g

$h(z)$ also make h small

邻近算子

Define the proximal operator

$$\text{prox}_h(\boldsymbol{x}) \;:=\; \arg\min_{\boldsymbol{z}} \left\{ \frac{1}{2} \|\boldsymbol{z} - \boldsymbol{x}\|_2^2 + h(\boldsymbol{z}) \right\}$$

for any convex function h

- well-defined under very general conditions (including nonsmooth convex functions)
- can be evaluated efficiently for many widely used functions (in particular, regularizers)

邻近算子与次梯度

定理

若h是适当的闭凸函数, 则 $u = \mathrm{prox}_h(x) \iff x - u \in \partial h(u)$

Proof.

若$u = \mathrm{prox}_h(x)$, 则由最优性条件得$0 \in \partial h(u) + (u - x)$, 因此有$x - u \in \partial h(u)$. 反之，若$x - u \in \partial h(u)$则由次梯度的定义可得到

$$h(v) \geqslant h(u) + (x - u)^{\mathrm{T}}(v - u), \quad \forall v \in \mathbf{dom}\ h$$

两边同时加$\frac{1}{2}\|v - x\|^2$, 即有

$$\begin{aligned} h(v) + \frac{1}{2}\|v - x\|^2 &\geqslant h(u) + (x - u)^{\mathrm{T}}(v - u) + \frac{1}{2}\|(v - u) - (x - u)\|^2 \\ &\geqslant h(u) + \frac{1}{2}\|u - x\|^2, \quad \forall v \in \mathbf{dom}\ h \end{aligned}$$

根据定义可得$u = \mathrm{prox}_h(x)$. □

实例

例：ℓ_1 范数

$h(x) = \|x\|_1, \quad \text{prox}_{th}(x) = \text{sign}(x)\max\{|x| - t, 0\}$

Proof.

邻近算子$u = \text{prox}_{th}(x)$ 的最优性条件为

$$x - u \in t\partial\|u\|_1 = \begin{cases} \{t\}, & u > 0 \\ [-t, t], & u = 0 \\ \{-t\}, & u < 0 \end{cases}$$

当$x > t$ 时，$u = x - t$；当$x < -t$ 时，$u = x + t$；当$x \in [-t, t]$ 时，$u = 0$，即有$u = \text{sign}(x)\max\{|x| - t, 0\}$. □

实例

例：二次函数(其中 A 对称正定)

$$h(x) = \frac{1}{2}x^{\mathrm{T}}Ax + b^{\mathrm{T}}x + c, \quad \mathrm{prox}_{th}(x) = (I + tA)^{-1}(x - tb)$$

例：负自然对数的和

$$h(x) = -\sum_{i=1}^{n} \ln x_i, \quad \mathrm{prox}_{th}(x)_i = \frac{x_i + \sqrt{x_i^2 + 4t}}{2}, \quad i = 1, 2, \cdots, n$$

课后作业：请完成上述邻近算子的证明

基本运算规则

- If $f(\boldsymbol{x}) = ag(\boldsymbol{x}) + b$ with $a > 0$, then

$$\text{prox}_f(\boldsymbol{x}) = \text{prox}_{ag}(\boldsymbol{x})$$

- **affine addition:** if $f(\boldsymbol{x}) = g(\boldsymbol{x}) + \boldsymbol{a}^\top \boldsymbol{x} + b$, then

$$\text{prox}_f(\boldsymbol{x}) = \text{prox}_g(\boldsymbol{x} - \boldsymbol{a})$$

- **quadratic addition:** if $f(\boldsymbol{x}) = g(\boldsymbol{x}) + \frac{\rho}{2}\|\boldsymbol{x} - \boldsymbol{a}\|_2^2$, then

$$\text{prox}_f(\boldsymbol{x}) = \text{prox}_{\frac{1}{1+\rho}g}\left(\frac{1}{1+\rho}\boldsymbol{x} + \frac{\rho}{1+\rho}\boldsymbol{a}\right)$$

- **scaling and translation:** if $f(\boldsymbol{x}) = g(a\boldsymbol{x} + \boldsymbol{b})$ with $a \neq 0$, then

$$\text{prox}_f(\boldsymbol{x}) = \frac{1}{a}\left(\text{prox}_{a^2 g}(a\boldsymbol{x} + \boldsymbol{b}) - \boldsymbol{b}\right)$$

课后作业：请完成上述运算规则的证明

Proof for quadratic addition

$$
\begin{aligned}
\mathrm{prox}_f(\boldsymbol{x}) &= \arg\min_{\boldsymbol{x}} \left\{ \frac{1}{2}\|\boldsymbol{z}-\boldsymbol{x}\|_2^2 + g(\boldsymbol{z}) + \frac{\rho}{2}\|\boldsymbol{z}-\boldsymbol{a}\|_2^2 \right\} \\
&= \arg\min_{\boldsymbol{x}} \left\{ \frac{1+\rho}{2}\|\boldsymbol{z}\|_2^2 - \langle \boldsymbol{z}, \boldsymbol{x}+\rho\boldsymbol{a}\rangle + g(\boldsymbol{z}) \right\} \\
&= \arg\min_{\boldsymbol{x}} \left\{ \frac{1}{2}\|\boldsymbol{z}\|_2^2 - \frac{1}{1+\rho}\langle \boldsymbol{z}, \boldsymbol{x}+\rho\boldsymbol{a}\rangle + \frac{1}{1+\rho} g(\boldsymbol{z}) \right\} \\
&= \arg\min_{\boldsymbol{x}} \left\{ \frac{1}{2}\left\| \boldsymbol{z} - \left(\frac{1}{1+\rho}\boldsymbol{x} + \frac{\rho}{1+\rho}\boldsymbol{a} \right) \right\|_2^2 + \frac{1}{1+\rho} g(\boldsymbol{z}) \right\} \\
&= \mathrm{prox}_{\frac{1}{1+\rho}g}\left(\frac{1}{1+\rho}\boldsymbol{x} + \frac{\rho}{1+\rho}\boldsymbol{a} \right)
\end{aligned}
$$

邻近点梯度下降

Proximal gradient descent: choose initialize $x^{(0)}$, repeat:

$$x^{(k)} = \mathrm{prox}_{h,t_k}\left(x^{(k-1)} - t_k \nabla g(x^{(k-1)})\right), \quad k = 1, 2, 3, \ldots$$

To make this update step look familiar, can rewrite it as

$$x^{(k)} = x^{(k-1)} - t_k \cdot G_{t_k}(x^{(k-1)})$$

where G_t is the generalized gradient of f,

$$G_t(x) = \frac{x - \mathrm{prox}_{h,t}\big(x - t\nabla g(x)\big)}{t}$$

回顾第4页： $\underset{z}{\mathrm{argmin}} \ \dfrac{1}{2t}\big\|z - \big(x - t\nabla g(x)\big)\big\|_2^2 + h(z)$

注记

Key point is that $\mathrm{prox}_{h,t}(\cdot)$ has a closed-form for many important functions h. Note:

- Mapping $\mathrm{prox}_{h,t}(\cdot)$ doesn't depend on g at all, only on h
- Smooth part g can be complicated, we only need to compute its gradients

Convergence analysis: will be in terms of the number of iterations, and each iteration evaluates $\mathrm{prox}_{h,t}(\cdot)$ once (this can be cheap or expensive, depending on h)

Lasso回归

Given $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, recall the lasso criterion:

$$f(\beta) = \underbrace{\frac{1}{2}\|y - X\beta\|_2^2}_{g(\beta)} + \underbrace{\lambda\|\beta\|_1}_{h(\beta)}$$

Proximal mapping is now

$$\begin{aligned} \text{prox}_t(\beta) &= \underset{z}{\text{argmin}} \ \frac{1}{2t}\|\beta - z\|_2^2 + \lambda\|z\|_1 \\ &= S_{\lambda t}(\beta) \end{aligned}$$

where $S_\lambda(\beta)$ is the soft-thresholding operator,

$$[S_\lambda(\beta)]_i = \begin{cases} \beta_i - \lambda & \text{if } \beta_i > \lambda \\ 0 & \text{if } -\lambda \leq \beta_i \leq \lambda \\ \beta_i + \lambda & \text{if } \beta_i < -\lambda \end{cases}, \quad i = 1, \ldots, n$$

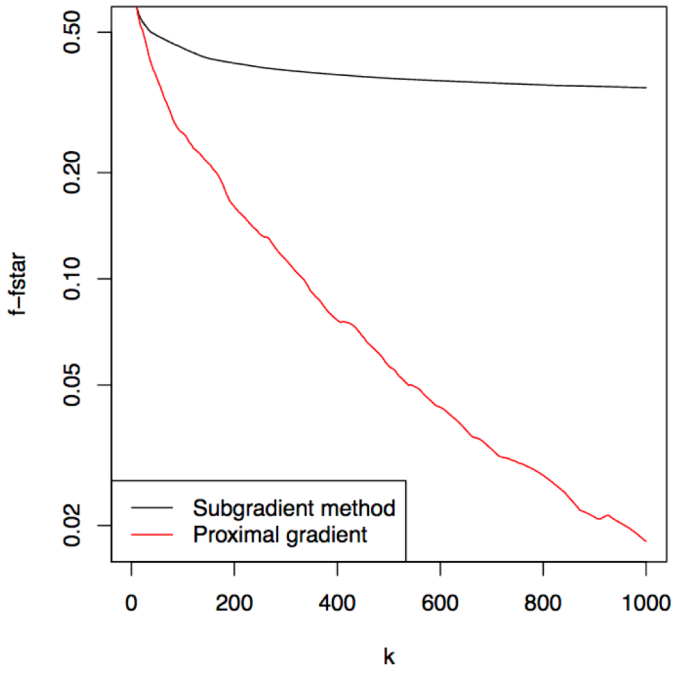
Lasso回归

Recall $\nabla g(\beta) = -X^T(y - X\beta)$, hence proximal gradient update is:

$$\beta^+ = S_{\lambda t}\big(\beta + tX^T(y - X\beta)\big)$$

Often called the iterative soft-thresholding algorithm (ISTA).

Example of proximal gradient (ISTA) vs. subgradient method convergence curves



收敛性

For criterion $f(x) = g(x) + h(x)$, we assume:

- g is convex, differentiable, $\mathrm{dom}(g) = \mathbb{R}^n$, and ∇g is Lipschitz continuous with constant $L > 0$
- h is convex, $\mathrm{prox}_t(x) = \mathrm{argmin}_z \{\|x - z\|_2^2/(2t) + h(z)\}$ can be evaluated

Theorem: Proximal gradient descent with fixed step size $t \leq 1/L$ satisfies

$$f(x^{(k)}) - f^\star \leq \frac{\|x^{(0)} - x^\star\|_2^2}{2tk}$$

Proximal gradient descent has convergence rate $O(1/k)$ or $O(1/\epsilon)$. Matches gradient descent rate! (But remember prox cost ...)

注：上述定理的证明请参见文再文版8.1.4节

收敛性

For criterion $f(x) = g(x) + h(x)$, we assume:

- g is convex, differentiable, $\mathrm{dom}(g) = \mathbb{R}^n$, and ∇g is Lipschitz continuous with constant $L > 0$
- h is convex, $\mathrm{prox}_t(x) = \mathrm{argmin}_z \{\|x - z\|_2^2/(2t) + h(z)\}$ can be evaluated

Theorem: Proximal gradient descent with fixed step size $t \leq 1/L$ satisfies

$$f(x^{(k)}) - f^\star \leq \frac{\|x^{(0)} - x^\star\|_2^2}{2tk}$$

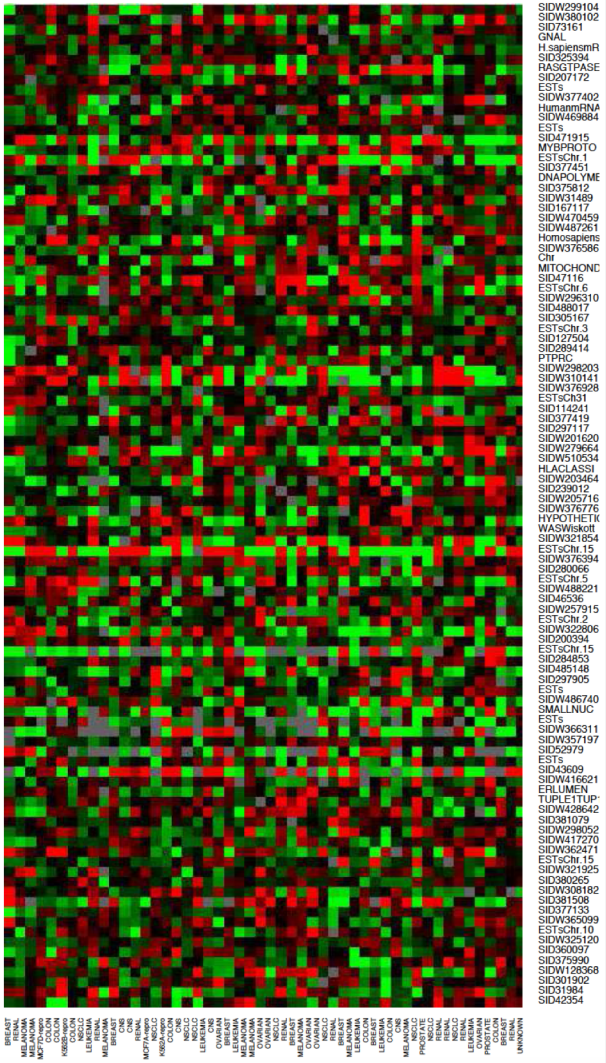
Proximal gradient descent has convergence rate $O(1/k)$ or $O(1/\epsilon)$. Matches gradient descent rate! (But remember prox cost ...)

注意：类似于梯度法，若g强凸，收敛率可提升到$O(\log \frac{1}{\epsilon})$

DNA微阵列数据

- Often tens of thousands of genes (features).
- Only tens of hundreds of samples.
- Data pre-processing:
  - Normalization.
  - Missing data imputation.
- Inference:
  - Which genes are significant?

$$n << p$$



Lasso与Ridge的几何对比

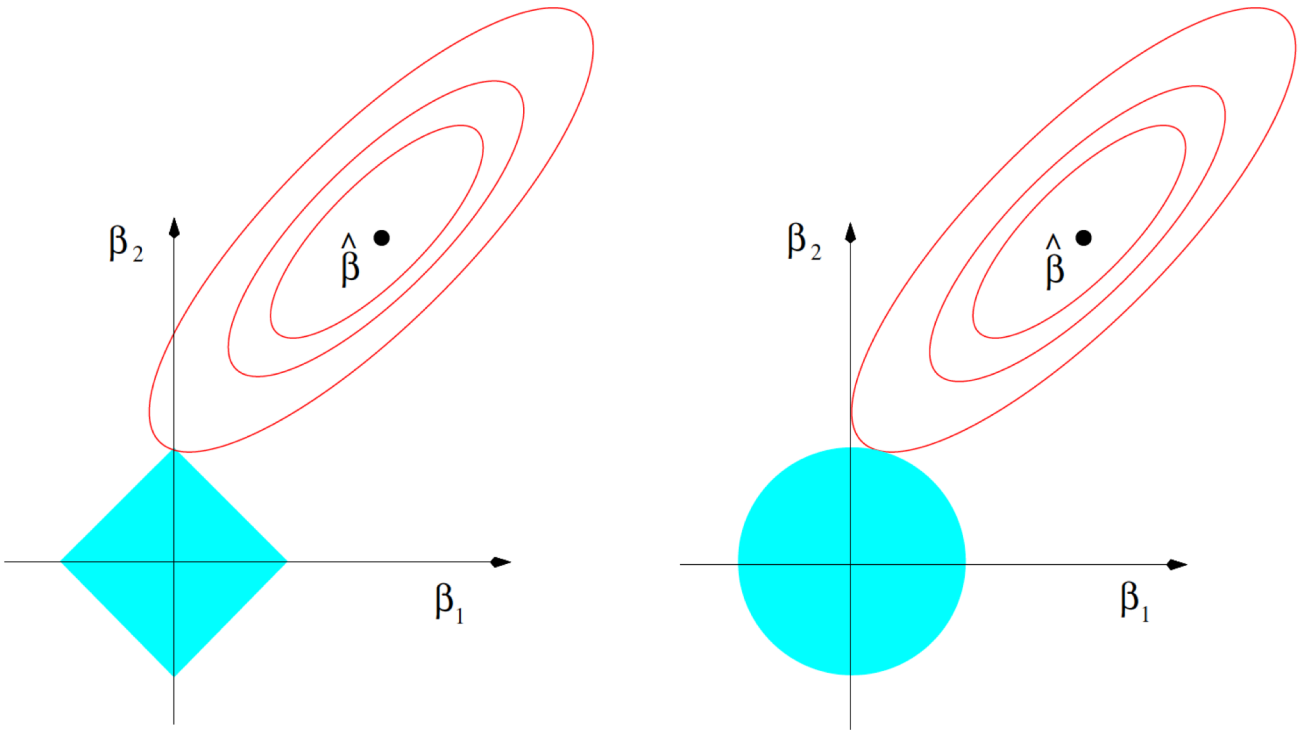
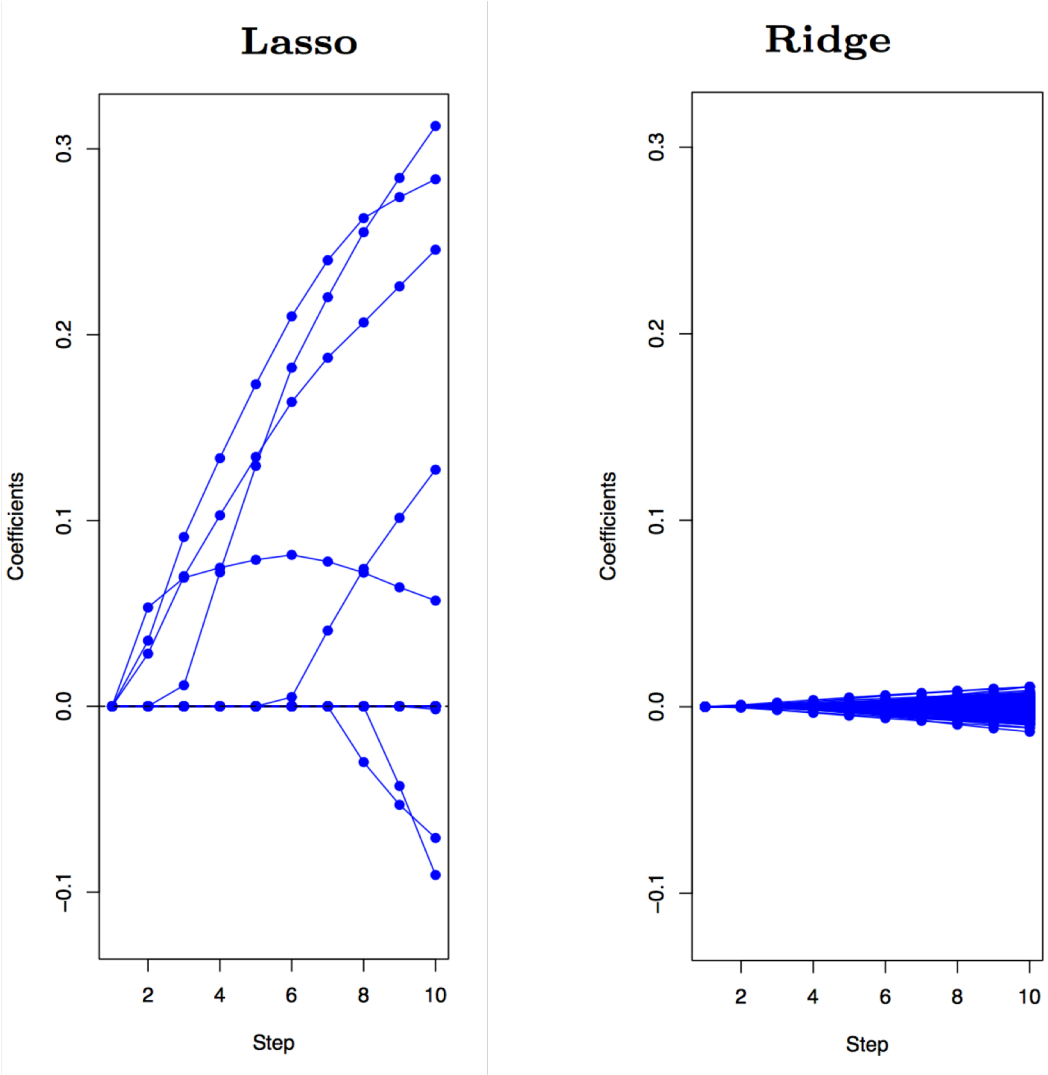


FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \leq t$ *and* $\beta_1^2 + \beta_2^2 \leq t^2$*, respectively, while the red ellipses are the contours of the least squares error function.*

图片来自于《The Elements of Statistical Learning》

Lasso与Ridge的实例对比



Leukemia Data, N=72, p=3571

Lasso与COVID-19

广州呼吸健康研究院 Guangzhou Institute of Respiratory Health

Calculation Tool For Predicting Critical-ill COVID-19 At Admission

Please answer the questions below to calculate.

1. X ray abnormality (平片异常)	No / Yes	7. Cancer history (肿瘤病史)	No / Yes
2. Age (年龄)	35	8. Neutrophil/Lymphocytes (NLR) (中性粒细胞/淋巴细胞) 0-80	5
3. Hemoptysis (咯血)	No / Yes	9. Lactate dehydrogenase (乳酸脱氢酶) 0-1500 U/L	75
4. Dyspnea (气促)	No / Yes	10. Direct Bilirubin (直接胆红素) 0-24 umol/L	2
5. Unconsciousness (意识丧失)	No / Yes		
6. Number of comorbidities (合并症数量)	0		

Total point (总分) : 37.9

Probability (概率) : 0.0073 (95% CI : 0.0049-0.0108)

Risk group (危险分层) : Low-risk (低危)

calculate (计算)

Note (备注) :Commorbidity includes Chronic Obstructive Pulmonary Disease,Hypertention,Diabetes,Coronary Heart Disease,Chronic Kindey Disease,Cancer,Cerebral Vascular Disease,Hepatitis B and Immunodeficiency. 共病包括：慢性阻塞性肺疾病、高血压、糖尿病、冠心病、慢性肾脏病、肿瘤、脑血管病、乙型肝炎和免疫缺陷。

Probability for Critical-ill events (invasive ventilation/ICU/death) : low-risk group 0.7% ; medium-risk group 7.3% ; high-risk group 59.3% . 发展为危重症 (插管/ICU/死亡) 总体概率：低危组0.7% ；中危组7.3% ；高危组59.3%。

Thank you for your attentions !