MACHINE LEARNING

អត្ថបទសង្ខេបជាភាសាខ្មែរ

លឹម ម៉េងសាយ

Mengsay's NOTES

Table of Contents

សេចក្តីផ្តើមអំពី Machine Learning

Machine Learning គឺសំដៅដល់បច្ចេកវិទ្យាដែលមានគោលដៅផ្តល់ឱ្យម៉ាស៊ីន(កុំព្យូទ័រ)នូវសមត្ថភាពដោះស្រាយបញ្ហាបានដោយខ្លួនឯង ពោលគឺមនុស្សមិនផ្តល់នូវបែបបទជាក់លាក់អំពីរបៀបដោះស្រាយណាមួយឡើយ។ ការរៀនពីរបៀបដោះស្រាយបញ្ហាក្នុងMachine Learningគឺធ្វើឡើងតាមរយៈការវិភាគលើទិន្នន័យជាគម្រូដែលផ្តល់ឱ្យដោយមនុស្ស។

បើធ្វើចំណាត់ថ្នាក់ក្រុមដោយផ្អែកលើប្រភេទគម្រូដែលផ្តល់ឱ្យម៉ាស៊ីនដើម្បីសង្កេត នោះគេអាចបែងចែកជា Supervised learning និង Unsupervised learning។ ចំពោះ Supervised learning ទិន្នន័យដែលផ្តល់ឱ្យម៉ាស៊ីនមានធាតុចូល(input)និងចម្លើយនៃបញ្ហាភ្ជាប់ជាមួយ។ ករណីនេះប្រៀបដូចជាករណីគ្រូផ្តល់លំហាត់និងដំណោះស្រាយជាមួយគ្នាឱ្យសិស្សរៀន។ផ្ទុយទៅវិញ Unsupervised learning ទិន្នន័យដែលផ្តល់ឱ្យម៉ាស៊ីនមានតែធាតុចូល(input)។ ករណីនេះអាចប្រៀបដូចជាករណីគ្រូផ្តល់តែលំហាត់ឱ្យសិស្សរៀនដោះស្រាយ សង្កេតដោយខ្លួនឯង។

1. Supervised Learning

ចំពោះSupervised Learning, គម្រូនៃធាតុចូល $\boldsymbol{x}$ និង ចម្លើយនៃបញ្ហា y ជាច្រើន $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ ត្រូវបានផ្តល់ឱ្យ។អ្វីដែលMachine Learningធ្វើគឺចង់កំណត់នូវទំនាក់ទំនងរវាង $\boldsymbol{x}$ និង y ។ ដោយផ្អែកលើប្រភេទនៃតម្លៃyនោះគេអាចបែងចែកជា

- ចំណោទតម្រែតម្រង់ Regression problem (y ជាអថេរជាប់ ចំនួនពិត)
  - Ex: ប៉ាន់ស្មានតម្លៃអចលនទ្រព្យ $y = 120000, 498302.25$ ...
- ចំណោទចំណាត់ថ្នាក់ទិន្នន័យDiscrimination problem/ Classification (y ជាអថេរដាច់)
  - Ex: ចំណាត់ថ្នាក់ប្រភេទសារ $\text{spam}(y = 1)$, $\text{not spam}(y = 0)$ ($\text{input } x: \textit{mail text}$)
  - Ex: កំណត់លេខសរសេរដៃ $y = 0,1,2, \dots,9.$ ($\text{input } x:$ *រូបភាព*)

2. Unsuperivsed Learning

ចំពោះUnupervised Learning, គម្រូនៃធាតុចូល $\boldsymbol{x}$ ជាច្រើន $\{\boldsymbol{x}\_i\}\_(i=1)\hat{}n$ ត្រូវបានផ្តល់ឱ្យដោយគ្មានគូចម្លើយនៃបញ្ហា y ។អ្វីដែលMachine Learningធ្វើគឺចង់ទាញរកនូវលក្ខណៈឬទម្រង់ពិសេសពីទិន្នន័យ $\boldsymbol{x}$ ។ បើនិយាយពីMachine learningបែបស្ថិតិវិទ្យា បញ្ហាប្រភេទនេះមានដូចជា

Dimensionality reduction : ការបង្ហាញទិន្នន័យដែលមានវិមាត្រខ្ពស់មកជាវិមាត្រទាបដោយរក្សាលក្ខណៈពិសេសនៃទិន្នន័យ

Feature selection : ការកំណត់នូវធាតុសំខាន់ដែលមានឥទ្ធិពលលើការប៉ាន់ស្មានអ្វីមួយ

Clustering : ការធ្វើចំណាត់ក្រុមដោយស្វ័យប្រវត្តិដោយផ្អែកលើលក្ខណៈនៃទិន្នន័យដែលមាន

ម៉ូឌែលតម្រែតម្រង់លីនេអ៊ែរ(Linear Regression Model)

ការសិក្សាលើម៉ូឌែលនៃបាតុភូត អាចឱ្យគេរកឃើញពីមូលហេតុឬកត្តាទាក់ទងនៃបាតុភូតនោះបាន។ លើសពីនេះគេក៏អាចប្រើប្រាស់ម៉ូឌែលនោះសម្រាប់ការទស្សន៍ទាយសម្រាប់អនាគតឬការប៉ាន់ស្មានលើទិន្នន័យដែលមិនមានក្នុងដៃ។

ជាទូទៅម៉ូឌែលដែលសិក្សាអំពីទំនាក់ទំនងរវាងលទ្ធផលនៃបាតុភូតមួយនិងកត្តាដែលអាចគិតបានថាជាកត្តាជះឥទ្ធិពលលើលទ្ធផលនោះ ហៅថា ម៉ូឌែលតម្រែតម្រង់ (Regression Model) ។ ក្នុងសប្តាហ៍នេះយើងនឹងណែនាំអំពីម៉ូឌែលមានទម្រង់ជាលីនេអ៊ែរដែលជាមូលដ្ឋាននៃតម្រែតម្រង់។

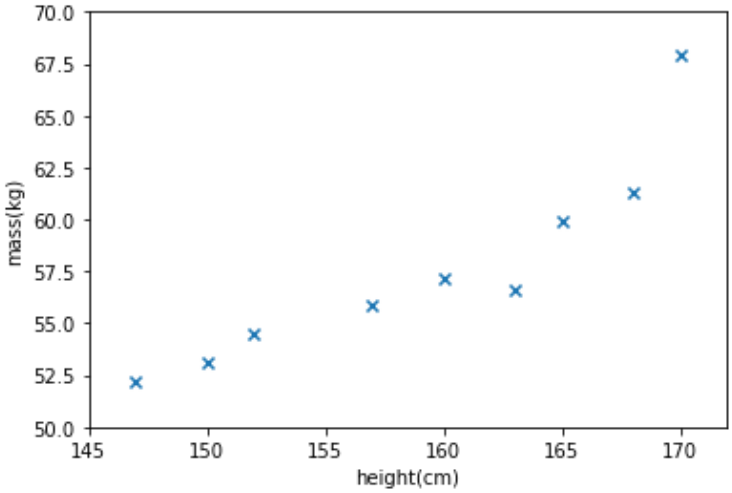
1. ការសិក្សាលើទំនាក់ទំនងរវាង២អថេរ

ជាឧទាហរណ៍យើងលើកយកការសិក្សារវាងទំនាក់ទំនងរវាងកម្ពស់(cm)និងម៉ាស(kg) តាមរយៈម៉ូឌែលតម្រែតម្រង់លីនេអ៊ែរ។

1.1. ទិន្នន័យនិងម៉ូឌែល

តារាងទី១ ទិន្នន័យកម្ពស់(cm)និងម៉ាស(kg)មនុស្ស៩នាក់

កម្ពស់(cm)	152	157	160	163	150	147	165	168	178
ម៉ាស(kg)	54.48	55.84	57.20	58.57	53.12	52.21	59.93	61.29	69.92



រូបទី១ របាយទិន្នន័យពីតារាងទី១

ក្នុងតារាងទី១ ទិន្នន័យអំពីកម្ពស់គិតជាសង់ទីម៉ែត្រ(cm)និងម៉ាសគិតជាគីឡូក្រាម(kg)របស់មនុស្ស ៩នាក់។ ពីទិន្នន័យនេះ យើងចង់សិក្សាពីទំនាក់ទំនងរវាងកម្ពស់(x)និងម៉ាស(y)។ នៅទីនេះយើងសន្មតថា តម្លៃម៉ាស(y)គឺជាអនុគមន៍នៃតម្លៃកម្ពស់(x): $y = f(x)$។

ដើម្បីងាយស្រួល ជាដំបូងយើងឧបមាថាទំនាក់ទំនងដើមរវាង x, y កំណត់ដោយអនុគមន៍ដូចខាងក្រោមដែលយើងហៅថា **ម៉ូឌែលលីនេអ៊ែរ**។

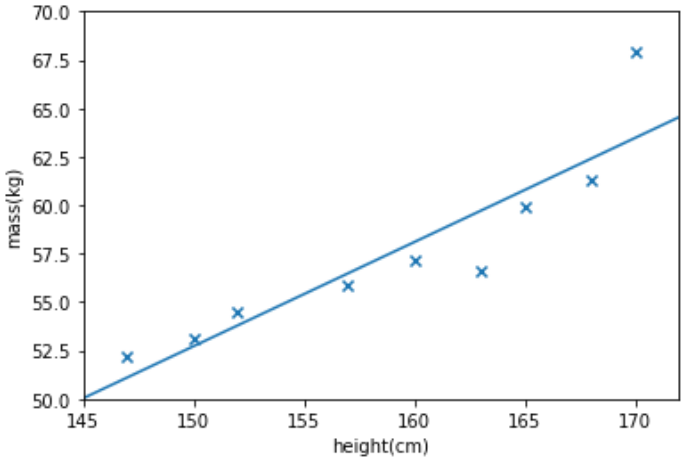
$$y = f(x) = \beta_0 + \beta_1 x$$

ក្រោមឧបមានេះទិន្នន័យដែលមានក្នុងតារាង១ ត្រូវធ្វើងផ្ទាត់ទំនាក់ទំនងខាងលើនេះ។ ប៉ុន្តែក្នុងដំណើរការវាស់វែង ឬ ស្រង់ទិន្នន័យលម្អៀងឬគម្លាតរវាងតម្លៃតាមម៉ូឌែលដែលឧបមាខាងលើនិងតម្លៃទិន្នន័យជាក់ស្តែងតែងតែកើតឡើង។ ហេតុនេះ យើងសិក្សាករណីដូចខាងក្រោម។

$$\text{ម៉ាស } (y) = \beta_0 + \beta_1 \times \text{កម្ពស់ } (x) + \text{ លម្អៀង}(\epsilon)$$

ក្នុងពេលនេះទិន្នន័យដែលមានអាចសរសេរជាទំនាក់ទំនងដូចខាងក្រោម។

$$54.58 = \beta_0 + \beta_1 \times 152 + \epsilon \text{ , } \ldots\ldots\ldots$$



រូបទី២ របាយទិន្នន័យនិងបន្ទាត់តម្រែតម្រង់លីនេអ៊ែរ(ម៉ូឌែល)

ជាទូទៅចំពោះទិន្នន័យចំនួន N: $\{(x_i, y_i)\}_{i=1}^{N}$ ទិន្នន័យនិមួយៗអាចបង្ហាញដោយទំនាក់ទំនងដូចខាងក្រោម។

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \ \ (i = 1,2, \ldots, N)$$

នៅទីនេះ β_0, β_1 ហៅថាមេគុណតម្រែតម្រង់(regression coefficient) ϵ គឺតម្លៃលម្អៀងរវាងទិន្នន័យពីការវាស់វែងនិងតម្លៃពិតតាមការសន្មត។ តម្លៃនៃម៉ាស y ហៅថា អថេរគោលដៅ(subjective variable) x ហៅថា អថេរពន្យល់ឬអថេរឯករាជ្យ(explanatory variable)។

គោលដៅរបស់យើងនៅទីនេះគឺការកំណត់តម្លៃនៃមេគុណតម្រែតម្រង់ ដែលធ្វើអោយម៉ូឌែលដែលសន្មតខាងលើអាចបង្ហាញទំនាក់ទំនងរវាងអថេរពន្យល់និងអថេរគោលដៅបានល្អប្រសើរ។

មានវិធីជាច្រើនដែលអាចឱ្យយើងកំណត់តម្លៃនៃមេគុណតម្រែតម្រង់ដែលល្អប្រសើរសម្រាប់ទិន្នន័យដែលមាន។ នៅក្នុងអត្ថបទនេះ យើងនឹងណែនាំអំពីវិធីសាស្ត្រងាយនិងពេញនិយម Least Square Error ។

1.2. ការប៉ាន់ស្មានតម្លៃប៉ារ៉ាម៉ែត្រដោយLeast Square Error

តាមការសន្មតនៃម៉ូឌែលខាងលើ តម្លៃនៃលម្អៀងរវាងទិន្នន័យនិមួយៗនិងតម្លៃពិតតាមម៉ូឌែលអាចកំណត់បានដូចខាងក្រោម។

$$\epsilon_i = y_i - (\beta_0 + \beta_1 x_i) \ , \qquad (i = 1,2, \dots, N)$$

ក្នុងវិធីសាស្ត្រLeast Square Error យើងសិក្សាលើផលបូកនៃការេរបស់តម្លៃលម្អៀងទាំងអស់របស់ទិន្នន័យដែលមាន ពោលគឺ $\epsilon_1^2 + \epsilon_2^2 + \cdots + \epsilon_N^2$ ។ គំនិតក្នុងវិធីសាស្ត្រនេះគឺងាយស្រួល ។ម៉ូឌែលដែលអាចពន្យល់ទំនាក់ទំនងរវាងអថេរទាំង២បានល្អប្រសើរ អាចត្រូវបាននិយាយបានថាជាម៉ូឌែលដែលមានតម្លៃនៃកម្រិតលម្អៀងតូចបំផុត។ ហេតុនេះ យើងនឹងធ្វើការកំណត់តម្លៃមេគុណតម្រែតម្រង់(ប៉ារ៉ាម៉ែត្រ)ណាដែលធ្វើឱ្យតម្លៃនៃផលបូកនៃការេរបស់តម្លៃលម្អៀងទាំងអស់របស់ទិន្នន័យ $E(\beta_0, \beta_1)$មានតម្លៃតូចបំផុត។

$$E(\beta_0, \beta_1) = \sum_{i=1}^{N} \epsilon_i^2 = \sum_{i=1}^{N} \{y_i - (\beta_0 + \beta_1 x_i)\}^2$$

អ្នកដែលបានសិក្សាគណិតវិទ្យា អាចមើលឃើញយ៉ាងងាយថា ពេលនេះវាបានក្លាយជាបញ្ហាបរមាកម្ម លើតម្លៃ $E(\beta_0, \beta_1)$ ដោយយកβ_0, β_1 ជាអថេរ។ យើងអាចដោះស្រាយបញ្ហានេះបានដោយងាយដោយប្រើចំណេះដឹងផ្នែកវិភាគមូលដ្ឋានដូចជាដេរីវេដោយផ្នែក។

នៅទីនេះ ដើម្បីមានភាពងាយស្រួលក្នុងការសិក្សាលើករណីច្រើនអថេរពន្យល់ យើងនឹងណែនាំការដោះស្រាយបញ្ហាខាងលើដោយប្រើវ៉ិចទ័រនិងម៉ាទ្រីស។

យើងកំណត់សរសេរម៉ាទ្រីសនិងវ៉ិចទ័រដូចខាងក្រោម។ X ពេលខ្លះត្រូវបានហៅថាម៉ាទ្រីសផែនការ។

$$X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_N \end{pmatrix} \in \mathbb{R}^{N \times 2} \ , \boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} \in \mathbb{R}^N \ , \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_N \end{pmatrix} \in \mathbb{R}^N \ , \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \in \mathbb{R}^2$$

ពេលនេះម៉ូឌែលនិងផលបូកតម្លៃការេនៃលម្អៀងខាងលើអាចសរសេរដូចខាងក្រោម។

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$
$$E(\boldsymbol{\beta}) = \sum_{i=1}^{N} \epsilon_i^2 = (\boldsymbol{y} - X\boldsymbol{\beta})^\top (\boldsymbol{y} - X\boldsymbol{\beta})$$

ត្រលប់ទៅកាន់បញ្ហារបស់យើងវិញ។ គោលដៅរបស់យើងគឺកំណត់តម្លៃ $\boldsymbol{\beta} = \begin{pmatrix}\beta_0 \\ \beta_1\end{pmatrix} \in \mathbb{R}^2$ដែលធ្វើឱ្យតម្លៃនៃ $E(\boldsymbol{\beta})$តូចបំផុត។ នៅទីនេះដូចដែលបានឃើញស្រាប់ អនុគមន៍ $E(\boldsymbol{\beta})$ ជាអនុគមន៍ប៉ោង ហេតុនេះយើងអាចកំណត់តម្លៃអប្បបរមារបស់វាបានងាយដោយគ្រាន់តែគណនាដេរីវេធៀបនឹងប៉ារ៉ាម៉ែត្រ $\boldsymbol{\beta}$ដូចខាងក្រោម។

$$E(\boldsymbol{\beta}) = (\boldsymbol{y} - X\boldsymbol{\beta})^\top (\boldsymbol{y} - X\boldsymbol{\beta}) = \boldsymbol{y}^\top \boldsymbol{y} - 2\boldsymbol{y}^\top X\boldsymbol{\beta} + \boldsymbol{\beta}^\top X^\top X\boldsymbol{\beta}$$

$$\frac{\partial}{\partial \boldsymbol{\beta}} E(\boldsymbol{\beta}) = -\mathbf{2}X^\top \boldsymbol{y} - \mathbf{2}X^\top X\boldsymbol{\beta}$$

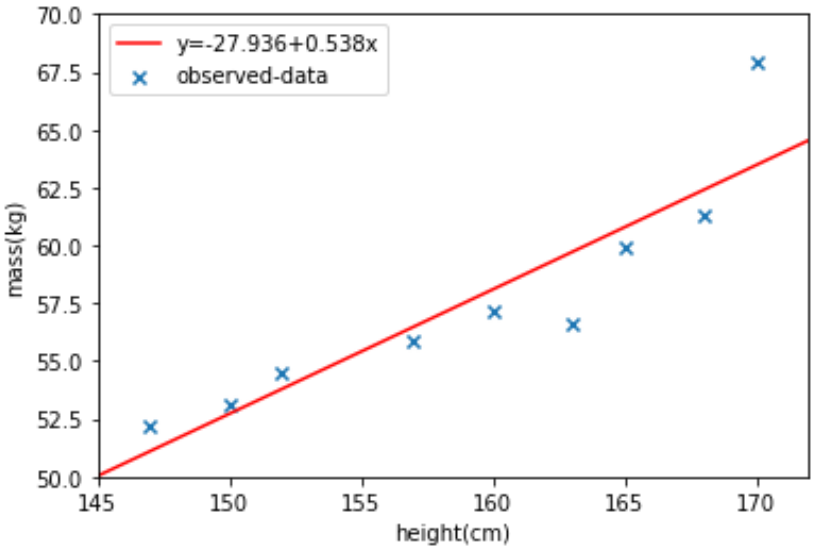
ដោយដោះស្រាយសមីការ $\frac{\partial}{\partial \boldsymbol{\beta}} E(\boldsymbol{\beta}) = \mathbf{0}$ យើងបាន

$$\widehat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^T \boldsymbol{y}$$

ដោយជំនួសតម្លៃដែលគណនាបាននេះទៅក្នុងម៉ូឌែលដើម យើងអាចគណនាតម្លៃទស្សន៍ទាយនៃអថេរគោលដៅ $\widehat{\boldsymbol{y}}$ នៅពេលស្គាល់តម្លៃអថេរពន្យល់ $\boldsymbol{x}$ បានដូចខាងក្រោម។

$$\widehat{\boldsymbol{y}} = \boldsymbol{x}\widehat{\boldsymbol{\beta}}$$

ក្នុងករណីទិន្នន័យក្នុងតារាងទី១ខាងលើ យើងអាចទទួលបានតម្លៃនៃមេគុណតម្រែតម្រង់និងបន្ទាត់តាងម៉ូឌែលតម្រែតម្រង់លីនេអ៊ែរដូចខាងក្រោម។



រូបទី៣ របាយទិន្នន័យនិងបន្ទាត់តម្រែតម្រង់លីនេអ៊ែរតាមLeast Square Error

1.3 ឯករាជភាពនៃតម្លៃម្លៀង Independence of errors

យើងពិនិត្យលើទំនាក់ទំនងរវាងតម្លៃនៃលម្អៀងនិងអថេរគោលដៅនិងអថេរពន្យល់។

បើសង្កេតលើតម្លៃនៃកូវ៉ារ្យង់រវាងតម្លៃលម្អៀងϵ និងតម្លៃទស្សន៍ទាយនៃអថេរគោលដៅ$\hat{y}$ ឬ តម្លៃនៃកូវ៉ារ្យង់រវាងតម្លៃលម្អៀងϵ និងតម្លៃទស្សន៍ទាយនៃអថេរពន្យល់xយើងបានលទ្ធផលដូចខាងក្រោម។*(សម្រាយបញ្ជាក់ទុកជាលំហាត់សម្រាប់អ្នកអាន)*

$$Cov[\hat{y}, \hat{\epsilon}] = 0$$
$$Cov[x, \hat{\epsilon}] = 0$$

លទ្ធផលនេះបង្ហាញពីឯករាជភាពនៃតម្លៃលម្អៀងដែលបង្កើតដោយម៉ូឌែលនិងតម្លៃទស្សន៍ទាយនៃអថេរគោលដៅ ឬ អថេរពន្យល់។

1.4 Coefficient of determination (R^2)

ដើម្បីបង្ហាញពីកម្រិតនៃការពន្យល់របស់ម៉ូឌែលទៅលើទំនាក់ទំនងរវាងទិន្នន័យដែលមាន Coefficient of determination (R^2) ត្រូវបានប្រើ។ តម្លៃ R^2 កំណត់ដោយផលធៀបរវាងវ៉ារ្យង់នៃតម្លៃទស្សន៍ទាយរបស់អថេរគោលដៅ និងវ៉ារ្យង់នៃតម្លៃអថេរគោលដៅពិត។

$$R^2 = \frac{Var[\hat{y}]}{Var[y]} = 1 - \frac{Var[\hat{\epsilon}]}{Var[y]}$$

តម្លៃនេះយកតម្លៃលើចន្លោះ $[0,1]$ ដែលតម្លៃខិតជិត១បង្ហាញពីភាពល្អប្រសើរនៃការពន្យល់របស់ម៉ូឌែលទៅលើទិន្នន័យ។

2. ការសិក្សាលើទំនាក់ទំនងរវាងច្រើនអថេរ

ក្នុងការសិក្សាលើទំនាក់ទំនងរវាងអថេរច្រើន ចំនួនអថេរពន្យល់អាចមានលើសពី១។ ក្នុងករណីនេះយើងសន្មតម៉ូឌែលដូចខាងក្រោម។

$$y = f(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_d x_d$$

ជាឧទាហរណ៍ដូចជាការសិក្សាទំនាក់ទំនងរវាងម៉ាស(kg)ជាអថេរគោលដៅ និង កម្ពស់(cm) កេទ(0/1) ទំហំចង្កេះ(cm) ជាអថេរពន្យល់ជាដើម។

ទោះបីជាចំនួននៃអថេរពន្យល់មានការកើនឡើងក្តី ការវិភាគដោយប្រើម៉ូឌែលតម្រែតម្រង់លីនេអ៊ែរមិនមានអ្វីប្រែប្រួលជាធំដុំនោះឡើយ។ អ្នកអាចបង្ហាញម៉ូឌែលខាងលើជាទម្រង់វ៉ិចទ័រនិងម៉ាទ្រីសរួចសិក្សាដូចគ្នា។

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1d} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{N1} & \cdots & x_{Nd} \end{pmatrix} \in \mathbb{R}^{N\times(d+1)}, \boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} \in \mathbb{R}^N, \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_N \end{pmatrix} \in \mathbb{R}^N, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_d \end{pmatrix} \in \mathbb{R}^{d+1}$$

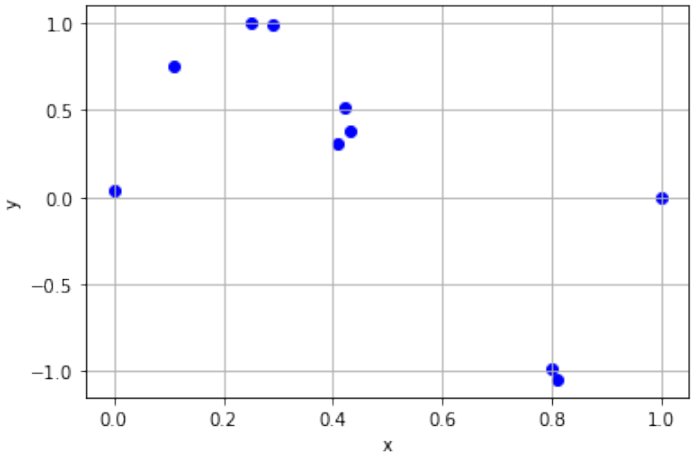
ពេលនេះម៉ូឌែលនិងផលបូកតម្លៃការេនៃលម្អៀងខាងលើអាចសរសេរដូចខាងក្រោម។

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}\ , \qquad E(\boldsymbol{\beta}) = \sum_{i=1}^{N} \epsilon_i^2 = (\boldsymbol{y} - X\boldsymbol{\beta})^\top (\boldsymbol{y} - X\boldsymbol{\beta})$$

ដោយប្រើវិធីសាស្ត្រLeast Square Error ដូចខាងលើយើងបានលទ្ធផលដូចគ្នា។

$$\widehat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^T \boldsymbol{y}$$

ក្នុងករណីខ្លះការប្រើប្រាស់តម្លៃផ្ទាល់នៃអថេរពន្យល់មិនអាចពណ៌នាទំនាក់ទំនងរវាងអថេរគោលដៅនិងអថេរពន្យល់បានល្អឡើយ។ ហេតុនេះការបម្លែងតម្លៃអថេរពន្យល់ត្រូវបានអនុវត្ត។



រូបទី៤ របាយទិន្នន័យដែលមិនអាចពន្យល់ដោយតម្លៃអថេរពន្យល់ផ្ទាល់

ជាឧទាហរណ៍ដូចជាការប្រើអនុគមន៍ដឺក្រេទី២ ឬខ្ពស់ជាងនេះ ឬការប្រើអនុគមន៍មិនមែនលីនេអ៊ែរជាដើម។ ក្នុងករណីនេះម៉ូឌែលតម្រែតម្រង់លីនេអ៊ែរអាចបង្ហាញដូចខាងក្រោម។នៅទីនេះទោះបីជាអនុគមន៍ $\phi_i(x)$ ជាទម្រង់មិនលីនេអ៊ែរក្តី ការហៅថាម៉ូឌែលតម្រែតម្រង់លីនេអ៊ែរ ព្រោះចង់សង្កត់ធ្ងន់លើផលបូកជាទម្រង់លីនេអ៊ែរនៃអនុគមន៍ទាំងអស់នោះ។

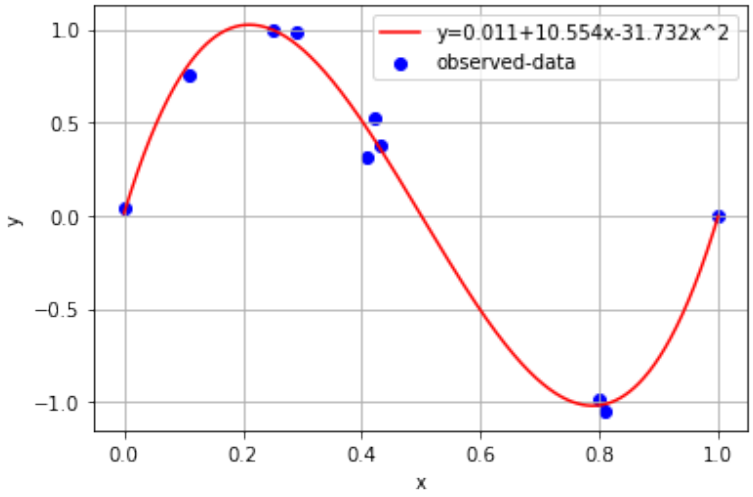
$$y = f(x) = \beta_0\phi_0(x) + \beta_1\phi_1(x) + \cdots + \beta_d\phi_d(x)$$

ករណីប្រើពហុធាដឺក្រេទី៣យើងបាន

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$$

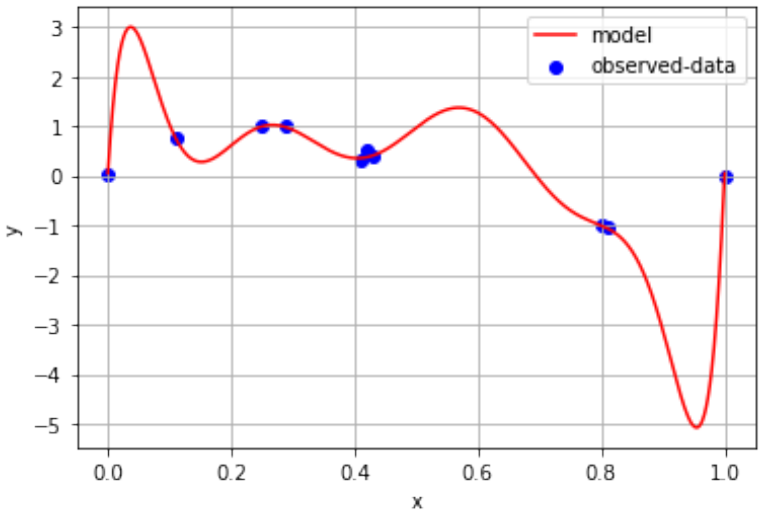
អ្នកអាចបង្ហាញម៉ូឌែលខាងលើជាទម្រង់វ៉ិចទ័រនិងម៉ាទ្រីសរួចសិក្សាដូចគ្នា។

$$X = \begin{pmatrix} 1 & x_1 & x_1^2 & x_1^3 \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_N & x_N^2 & x_N^3 \end{pmatrix} \in \mathbb{R}^{N\times 4}, \boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} \in \mathbb{R}^N, \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_N \end{pmatrix} \in \mathbb{R}^N, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_3 \end{pmatrix} \in \mathbb{R}^4$$



រូបទី៥ របាយទិន្នន័យនិងខ្សែកោងម៉ូឌែលតម្រែតម្រង់ជាពហុធាដឺក្រេទី៣

តាមការសង្កេត អ្នកអាចនឹងមើលឃើញថាប្រសិនបើយើងតម្លើងដឺក្រេនៃពហុធានោះកម្រិតនៃការពណ៌នារបស់ម៉ូឌែលលើទិន្នន័យនឹងមានការកើនឡើង។ ប៉ុន្តែការបង្ហាញម៉ូឌែលដែលមានភាពស្មុគស្មាញពេកអាចត្រឹមតែពណ៌នាលើទិន្នន័យដែលមានតែប៉ុណ្ណោះ ផ្ទុយទៅវិញវាមិនអាចទស្សន៍ទាយបានល្អឡើយចំពោះទិន្នន័យដែលមិនមានក្នុងដៃ ។រូបភាពទី៥បង្ហាញករណីនេះ។



រូបទី៦ របាយទិន្នន័យនិងខ្សែកោងម៉ូឌែលតម្រែតម្រង់ជាពហុធាដឺក្រេទី៩

ក្នុងរូបទី៤នេះ បើយើងគណនា Coefficient of determination (R^2)យើងនឹងបានតម្លៃ 1 ដែលជាតម្លៃយ៉ាងល្អក្នុងការពណ៌នាទិន្នន័យដែលមានក្នុងដៃ។ ប៉ុន្តែបើយើងសង្កេតលើគម្រូទិន្នន័យជាក់ស្តែងនិងក្រាបនៃម៉ូឌែល តម្លៃនៃការទស្សន៍ទាយត្រង់តំបន់ដែលគ្មានទិន្នន័យ គឺគ្មានលំនឹង និងចេញផុតឆ្ងាយពីដែននៃទិន្នន័យដែលមានក្នុងដៃ។

ហេតុនេះក្នុងការជ្រើសរើសម៉ូឌែល អ្នកគួរសង្កេតលើចរិតលក្ខណៈនៃទិន្នន័យព្រមទាំងភាពល្អប្រសើរនៃការពន្យល់របស់វាចំពោះទាំងទិន្នន័យដែលមានក្នុងដៃស្រាប់ និងទិន្នន័យដែលមិនមានពេលគឺភាពប្រសើរក្នុងការទស្សន៍ទាយឬប៉ាន់ស្មាននាពេលអនាគត។

គណនាមេគុណតម្រែតម្រង់ដោយប្រើម៉ូឌែលជាពហុធាដឺក្រេទីk និងការទស្សន៍ទាយតម្លៃអថេរ គោលដៅជាមួយPython

```
import numpy as np
```

```
def fit(x,y,k):
  X_ = np.zeros((len(x),k+1))
  for i in range(k+1):
    X_[:,i] = x**i
  w = np.linalg.inv(X_.T@X_)@X_.T@y
  return w
```

```
def predict(x,w,k):
  X_ = np.zeros((len(x),k+1))
  for i in range(k+1):
    X_[:,i] = x**i
  return X_@w
```

វិធីសាស្ត្របរមាកម្មតាមរយៈ SGD

(SGD: Stochastic Gradient Descent)

ក្នុងអត្ថបទមុនយើងបានណែនាំអំពីម៉ូឌែលតម្រែតម្រង់លីនេអ៊ែរ ដែលត្រូវបានប្រើប្រាស់សម្រាប់សិក្សាពីការទំនាក់ទំនងរវាងអថេរពន្យល់និងអថេរគោលដៅ។
ក្នុងការកំណត់តម្លៃប៉ារ៉ាម៉ែត្រនៃម៉ូឌែល(មេគុណតម្រែតម្រង់)
យើងបានដោះស្រាយតាមរយៈវិធីសាស្ត្រជាមូលដ្ឋាននៃគណិតវិទ្យាវិភាគដូចខាងក្រោមនេះ។

$$\textit{ម៉ូឌែល } \boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

$$\textit{មេគុណតម្រែតម្រង់ } \widehat{\boldsymbol{\beta}} = (X^{\top}X)^{-1}X^{T}\boldsymbol{y}$$

ប៉ុន្តែក្នុងជីវភាពរស់នៅ ករណីភាគច្រើនចំនួននៃអថេរពន្យល់
មានចំនួនច្រើនលើសលប់ ដែលធ្វើឱ្យវិមាត្រនៃម៉ាទ្រីសផែនការមានការកើនឡើងខ្ពស់។ ហេតុនេះ វាមានការលំបាកក្នុងការគណនាម៉ាទ្រីសច្រាស់
ដូចក្នុងរបៀបខាងលើទោះបីប្រើប្រាស់ម៉ាស៊ីនកុំព្យូទ័រក្តី។

ក្នុងអត្ថបទនេះ យើងនឹងណែនាំវិធីសាស្ត្រកំណត់តម្លៃប៉ាន់ស្មាននៃមេគុណតម្រែតម្រង់ដោយវិធីគណនាដដែលៗលើតម្លៃលេខតាមប្រមាណវិធីងាយៗគឺ Stochastic Gradient Descent (SGD) ។
ដើម្បីងាយស្រួលស្វែងយល់អំពីSGD ជាដំបូងយើងនឹងណែនាំអំពីគំនិត និងការគណនាក្នុងវិធីសាស្ត្រ Gradient Descent ជាមុន។

1. Gradient Descent

ដូចដែលបានបង្ហាញក្នុងអត្ថបទមុន យើងចង់កំណត់យកមេគុណតម្រែតម្រង់ណាដែលធ្វើឱ្យតម្លៃផលបូកការេនៃកម្រិតលម្អៀងតូចបំផុត។ គោលគំនិតក្នុងGradient Descent គឺផ្លាស់ប្តូរតម្លៃនៃមេគុណតម្រែតម្រង់(ប៉ារ៉ាម៉ែត្រ)
បន្តិចម្តងៗ ទៅតាមទិសដៅដែលធ្វើឱ្យតម្លៃផលបូកការេនៃកម្រិតលម្អៀងមានការថយចុះ
។ អ្នកអាចប្រដូចវិធីនេះទៅនឹងការចុះជំរាលឬចុះពីទីភ្នំ ដោយរំកិលខ្លួនអ្នកបន្តិចម្តងៗ ទៅកាន់ទីដែលទាបជាងកន្លែងដែលអ្នកនៅ។ នៅពេលដែលអ្នករំកិលខ្លួនដល់ទីដែលលែងមានបម្រែបម្រួលនៃរយៈកម្ពស់ អ្នកអាចសន្និដ្ឋានបានថាអ្នកដល់ទីដែលទាបបំផុតហើយ។ ដូចគ្នានេះដែរ នៅក្នុងវិធីសាស្ត្រGradient Descent តាមលក្ខណៈគណិតវិទ្យានៃ gradient (តម្លៃដេរីវេនៃអនុគមន៍ត្រង់ចំនុចណាមួយ) តម្លៃgradientត្រង់ចំណុចណាមួយគឺជាតម្លៃមេគុណ

ប្រាប់ទិសនៃខ្សែកោងត្រង់ចំណុចនោះហើយក៏ជាតម្លៃធំបំផុតនៃបម្រែបម្រួលតម្លៃអនុគមន៍ពេល
អ្នកធ្វើបម្រែបម្រួលលើអថេរមិនអាស្រ័យ។

រូបទី១ គំនិតក្នុង Gradient Descent

រូបទី១បង្ហាញអំពីគំនិតក្នុងវិធីសាស្ត្រធ្វើអប្បបរមាកម្មតាម Gradient Descent។ ដូចដែល
អ្នកអាចធ្វើការកត់សម្គាល់បាន ពេលខ្លះអ្នកអាចនឹងធ្លាក់ចុះទៅក្នុងទីតាំងដែលជាបរមាធៀបតែ
មិនមែនជាកន្លែងអប្បបរមាពិតប្រាកដប្រសិនបើទីតាំងនៃការចាប់ផ្តើមរបស់អ្នក
មិនប្រសើរ។ ប៉ុន្តែក្នុងករណីធ្វើបរមាកម្មតម្លៃផលបូកការេនៃកម្រិតលម្អៀងរបស់យើង ដោយសារ
អនុគមន៍ដែលត្រូវធ្វើបរមាកម្មគឺជាអនុគមន៍ជីក្រេទី២ ហេតុនេះយើងមិនមានការព្រួយបារម្ភក្នុងករណី
នេះឡើយ។

ពេលនេះ យើងពិនិត្យលើការគណនាក្នុងវិធីសាស្ត្រ Gradient Descent។

យើងសិក្សាលើអនុគមន៍ដែលយកតម្លៃស្មាលែ $f(\boldsymbol{x})$ ដែល $\boldsymbol{x} \in \mathbb{R}^d$ ។ សន្មតថាអនុគមន៍
នេះយកតម្លៃអប្បបរមាត្រង់ចំណុច $\boldsymbol{x}^*$ ។ វិធីសាស្ត្រ Gradient Descent អាចឱ្យយើងគណនាតម្លៃ
(ប្រហែល)នៃ
$\boldsymbol{x}^*$ បានដោយចាប់ផ្តើមពីតម្លៃ $\boldsymbol{x}^{(0)}$ ណាមួយ រួចធ្វើការផ្លាស់ប្តូរតម្លៃនេះតាមការគណនាដូចខាងក្រោម
។

$$\boldsymbol{x}^{(t+1)} = \boldsymbol{x}^{(t)} - \eta_t \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}}\bigg|_{\boldsymbol{x}=\boldsymbol{x}^{(t)}}$$

នៅទីនេះ$t = 0,1, \ldots$ គឺជាលេខរៀងនៃការផ្លាស់ប្តូរតម្លៃអថេរ$\boldsymbol{x}$។ $\frac{\partial f(x)}{\partial x}$ គឺជាដេរីវេដោយផ្នែក
នៃអនុគមន៍ f ធៀបនឹងអថេរ $\boldsymbol{x}$ ឬហៅថា gradient ។ η_t គឺជាកម្រិតនៃការផ្លាស់ប្តូរតម្លៃអថេរដោយ
គ្រប់គ្រងលើឥទ្ធិពលនៃតម្លៃgradient។ នៅក្នុង Machine Learning វាត្រូវបានហៅថា
ជា អត្រារៀនឬ learning rate
។ ជាទូទៅតម្លៃនៃη_t ត្រូវបានកំណត់យកចន្លោះ0និង១ដោយតម្លៃយ៉ាងតូច។

យើងអាចកំណត់លក្ខខណ្ឌសម្រាប់បញ្ចប់ការផ្លាស់ប្តូរតម្លៃនៃអថេរបាន ដោយយកពេលដែលតម្លៃដាច់ខាតនៃ gradient យកតម្លៃសូន្យឬក្បែរសូន្យ។

ពិនិត្យលើករណីគម្រូងាយមួយ $f(x) = x^2 - 2x - 3$ ។ ករណីនេះយើងដឹងច្បាស់ថាតម្លៃអប្បបរមានៃអនុគមន៍គឺ -4 នៅពេលដែល $x^* = 1$។ យើងនឹងធ្វើងផ្ទាត់ជាមួយតម្លៃដែលគណនាតាមរយៈGradient Descent។

ដំបូងយើងគណនាអនុគមន៍ដេរីវេ $\frac{df(x)}{dx} = 2x - 2$ និង កំណត់យកអត្រា $\eta = 0.1$ ថេរ។ យើងចាប់ផ្តើមពីចំណុច $x^{(0)} = 0$, $f\left(x^{(0)}\right) = -3$ ។ ដោយផ្លាស់ប្តូរតម្លៃអថេរតាមរយៈGradient Descent ខាងលើយើងបានបម្រែបម្រួលនៃតម្លៃអថេរនិងតម្លៃអនុគមន៍ដូចតារាងខាងក្រោម។

តារាងទី១ បម្រែបម្រួលនៃតម្លៃអថេរនិងអនុគមន៍តាម Gradient Descent

$\boldsymbol{t}$	$\boldsymbol{x^{(t)}}$	$\boldsymbol{\frac{df(x)}{dx}}$	$\boldsymbol{f(x)}$
0	0.00	-2.00	-3.00
1	0.20	-1.60	-3.36
2	0.36	-1.28	-3.59
⋮	⋮	⋮	⋮
44	0.999946	-0.000109	-4.00
45	0.999956	-0.000087	-4.00

យើងត្រលប់ទៅកាន់ម៉ូឌែលតម្រែតម្រង់របស់យើងវិញ។ អនុគមន៍ដែលយើងចង់ធ្វើអប្បបរមាកម្មគឺ $E(\boldsymbol{\beta})$ ដោយយក $\boldsymbol{\beta}$ ជាអថេរ។

$$E(\boldsymbol{\beta}) = \sum_{i=1}^{N} \epsilon_i^2 = (\boldsymbol{y} - X\boldsymbol{\beta})^\top (\boldsymbol{y} - X\boldsymbol{\beta})$$

អនុគមន៍ដេរីវេ(gradient)របស់វាគឺ

$$E(\boldsymbol{\beta}) = (\boldsymbol{y} - X\boldsymbol{\beta})^\top (\boldsymbol{y} - X\boldsymbol{\beta}) = \boldsymbol{y}^\top \boldsymbol{y} - 2\boldsymbol{y}^\top X\boldsymbol{\beta} + \boldsymbol{\beta}^\top X^\top X\boldsymbol{\beta}$$

$$\frac{\partial}{\partial \boldsymbol{\beta}} E(\boldsymbol{\beta}) = -2X^\top \boldsymbol{y} - 2X^\top X\boldsymbol{\beta} = 2X^\top (X\boldsymbol{\beta} - \boldsymbol{y}) = 2X^\top (\hat{\boldsymbol{y}} - \boldsymbol{y})$$

ហេតុនេះ កន្សោមសម្រាប់ការផ្លាស់ប្តូរតម្លៃអថេរគឺ

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \eta_t \left. \frac{\partial E(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right|_{\boldsymbol{\beta} = \boldsymbol{\beta}^{(t)}}$$

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - 2\eta_t X^\top \left(\hat{\boldsymbol{y}}^{(t)} - \boldsymbol{y}\right)$$

ដែល $\hat{\boldsymbol{y}}^{(t)} = X\boldsymbol{\beta}^{(t)}$ ។

យើងសាកល្បងគណនាតម្លៃប្រហែលនៃមេគុណតម្រែតម្រង់ដែលបានសិក្សាក្នុងអត្ថបទមុន ដោយប្រើ gradient descent។ លើកនេះយើងយកតម្លៃកម្ពស់គិតជាម៉ែត្រដើម្បីបង្រួមតម្លៃលេខ។

តារាងទី២ ទិន្នន័យកម្ពស់(m)និងម៉ាស(kg)មនុស្ស៩នាក់

កម្ពស់(m)	1.52	1.57	1.60	1.63	1.50	1.47	1.65	1.68	1.78
ម៉ាស(kg)	54.48	55.84	57.20	58.57	53.12	52.21	59.93	61.29	69.92

ជាមួយPythonអ្នកអាចសរសេរCodeដូចខាងក្រោម។
នៅទីនេះយើងកំណត់យកតម្លៃចាប់ផ្តើមនៃ $\boldsymbol{\beta}^{(0)} = \mathbf{0}$ និង $\eta = 0.001$

```
import numpy as np
X = np.array([1.52,1.57,1.60,1.63,1.50,1.47,1.65,1.68,1.70])
y =
np.array([54.48,55.84,57.20,56.57,53.12,52.21,59.93,61.29,67.92])
XP = np.vstack([np.ones_like(X), X]).T
beta = np.zeros(XP.shape[1])

eta = 1e-3
for t in range(10000):
    y_hat = XP @ beta
    beta -= 2 * eta * XP.T @ (y_hat - y)
```

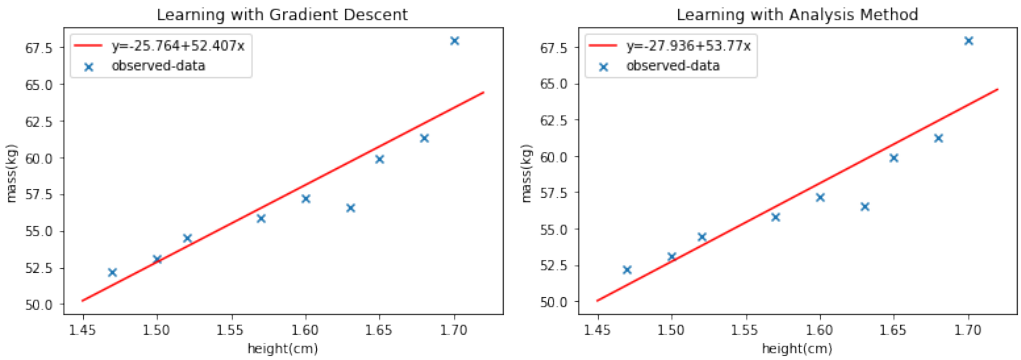
ជាលទ្ធផលយើងបានតម្លៃប្រហែលនៃមេគុណតម្រែតម្រង់គឺ

```
Beta
array([-25.76358113,  52.40677129])
```

អ្នកអាចធ្វើងផ្ទាត់តម្លៃនេះតាមរយៈការគណនាដោយប្រើម៉ាទ្រីសច្រាសដូចក្នុងអត្ថបទមុនបាន។ ក្នុងកណីនេះអ្នកនឹងបានលទ្ធផល

```
Beta
array([-27.93562969,  53.76959967])
```

តាមរយៈលទ្ធផលនេះយើងឃើញថា ការគណនាដោយប្រើ gradient descent អាចជួយយើងឱ្យធ្វើការប៉ាន់ស្មានតម្លៃនៃមេគុណតម្រែតម្រង់បាន។



រូបទី២ លទ្ធផលនៃតម្រែតម្រង់តាម Gradient Descent និង តាមការគណនាដោយគណិតវិទ្យាវិភាគ

2. Stochastic Gradient Descent

ការធ្វើបរមាកម្មលើតម្លៃអនុគមន៍ដោយប្រើ Gradient Descent អាចជួយយើងឱ្យធ្វើការគណនា
បានយ៉ាងមានប្រសិទ្ធភាពទោះបីជាវិមាត្រឬចំនួននៃអថេរពន្យល់ច្រើនក៏ដោយ។ ប៉ុន្តែក្នុងវិធីសាស្ត្រ Gradient Descent ការគណនា gradient ត្រូវបានធ្វើឡើងដោយប្រើប្រាស់ទិន្នន័យទាំងអស់ដែលមានក្នុងដៃ។ ក្នុងករណីដែលចំនួនទិន្នន័យមានច្រើន វិធីនេះត្រូវបានគេដឹងថាមានភាពយឺតយ៉ាវក្នុងការរួមទៅរកតម្លៃបរមារបស់អនុគមន៍។

ដើម្បីដោះស្រាយបញ្ហានេះ Stochastic Gradient Descent (SGD) ត្រូវបានប្រើប្រាស់ជំនួសវិញ។ ក្នុងករណីចំនួនទិន្នន័យដែលមាន
(N) មានបរិមាណច្រើន ក្នុងវិធីSGD ទិន្នន័យម្តងមួយៗ ត្រូវបានជ្រើសយកដោយចៃដន្យ ដើម្បីគណនា gradient នៃអនុគមន៍ រួចធ្វើការផ្លាស់ប្តូរតម្លៃអថេរតែម្តង ដោយមិនចាំបាច់ធ្វើការបូកសរុបគ្រប់ទិន្នន័យដែលមាននោះឡើយ។

ជាទូទៅ ដើម្បីអនុវត្តSGDបាន ចំពោះទិន្នន័យសរុបDដែលមានអនុគមន៍ដែលត្រូវធ្វើបរមាកម្ម ត្រូវតែអាចសរសេរជាផលបូកនៃ
អនុគមន៍ដែលយកករណីទិន្នន័យនិមួយៗជាធាតុចូលដូចខាងក្រោម។

$$E_D(\boldsymbol{\beta}) = \sum_{(\boldsymbol{x},y)\in D} e(\boldsymbol{\beta})$$

ក្នុងករណីយើងកំពុងសិក្សានេះ ដោយសារ$E_D(\boldsymbol{\beta})$ត្រូវបានកំណត់ដោយផលបូកការេនៃកម្រិតលម្អៀងគ្រប់ទិន្នន័យទាំងអស់ $E_D(\boldsymbol{\beta}) = \sum_{i=1}^{N} \epsilon_i^2$ ហេតុនេះ លក្ខខណ្ឌខាងលើត្រូវបានផ្ទៀងផ្ទាត់។

ចំពោះទិន្នន័យនិមួយៗ$(\boldsymbol{x}_i, y_i)$ gradient នៃអនុគមន៍ដែលត្រូវធ្វើបរមាកម្មអាចគណនាបានដូចខាងក្រោម។

$$\frac{\partial e(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{\partial}{\partial \boldsymbol{\beta}}\left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)^2 = -2\left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)\boldsymbol{x}_i^\top = 2(\hat{y}_i - y_i)\boldsymbol{x}_i^\top$$

កន្សោមសម្រាប់ធ្វើការផ្លាស់ប្តូរតម្លៃនៃអថេរតាម SGD គឺអាចបង្ហាញដូចទម្រង់ខាងក្រោម។

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \eta_t \left.\frac{\partial e(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(t)}}$$

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - 2\eta_t\left(\hat{y}_i^{\,(t)} - y\right)\boldsymbol{x}_i^\top$$

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - 2\eta_t\boldsymbol{\delta_i}$$

ដែល $\boldsymbol{\delta_i} = \left(\hat{y}_i^{\,(t)} - y\right)\boldsymbol{x}_i^\top$ ។

ជាមួយPythonអ្នកអាចសរសេរCodeដូចខាងក្រោម។

នៅទីនេះយើងកំណត់យកតម្លៃចាប់ផ្តើមនៃ $\boldsymbol{\beta}^{(0)} = \mathbf{0}$ និង $\eta = 0.001$

```
import random
import numpy as np
```

```
X = np.array([1.52,1.57,1.60,1.63,1.50,1.47,1.65,1.68,1.70])
y =
np.array([54.48,55.84,57.20,56.57,53.12,52.21,59.93,61.29,67.92])
```

```
beta = np.zeros(2)
d_index = list(range(len(X)))

eta = 1e-3
for t in range(100000):
  random.shuffle(d_index)
  for i in d_index :
    XP = np.vstack([np.ones_like(X[i]), X[i]]).T
    y_hat = XP @ beta
    beta -= 2 * eta * XP.T @ (y_hat - y[i])
```

ជាលទ្ធផលយើងបានតម្លៃប្រហែលនៃមេគុណតម្រែតម្រង់គឺ

```
Beta
array([-25.78979689,  52.42501619])
```

បើយើងធ្វើការប្រៀបធៀបរវាង Gradient Descent និង SGD យើងអាចនិយាយបានថា SGD គឺជាវិធីសាស្ត្រដែលសន្មតយកតម្លៃ gradient ចំពោះគ្រប់ទិន្នន័យទាំងអស់ក្នុង Gradient Descent ដោយតម្លៃប្រហែល $\boldsymbol{\delta_i} = \left(\hat{y}_i^{\,(t)} - y\right)\boldsymbol{x}_i^\top$ ពោលគឺ $\frac{\partial E_D(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \approx \frac{\partial e_{x_i,y_i}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \boldsymbol{\delta_i}$ ។

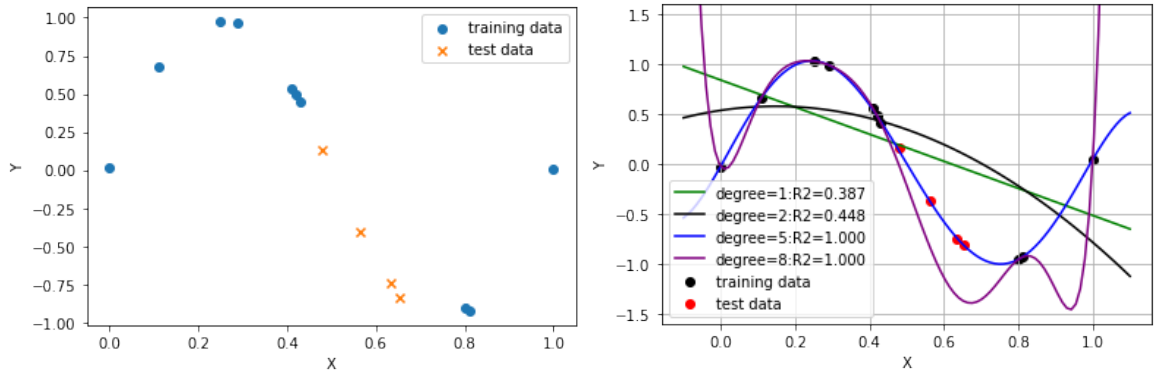
ក្នុងអត្ថបទមុនយើងបានស្វែងយល់អំពីម៉ូឌែលតម្រែតម្រង់លីនេអ៊ែរនិងវិធីសាស្ត្រក្នុងការកំណត់តម្លៃមេគុណតម្រែតម្រង់(ប៉ារ៉ាម៉ែត្រ)
ដោយប្រើប្រាស់គណិតវិទ្យាវិភាគនិងវិធីសាស្ត្រប៉ាន់ស្មានតម្លៃតាមវិធីសាស្ត្រSGD។ ប៉ុន្តែបញ្ហាដែលនៅសល់គឺថាតើយើងគួរជ្រើសរើសយកម៉ូឌែលបែបណាទើបអាចឱ្យវាពណ៌នាទំនាក់ទំនងរវាងទិន្នន័យបានល្អប្រសើរ។ យើងពិនិត្យករណីខាងក្រោមជាឧទាហរណ៍។

សន្មតថាយើងមានទិន្នន័យដូចរូបទី១(a)។ យើងចង់បង្កើតម៉ូឌែលតម្រែតម្រង់ដើម្បីសិក្សាពីទំនាក់ទំនងរវាងអថេរពន្យល់និងអថេរគោលដៅ។

ដើម្បីសិក្សាពីភាពល្អប្រសើរនៃម៉ូឌែល យើងបែងចែកទិន្នន័យជាពីរផ្នែកគឺ training data ដែលប្រើសម្រាប់កំណត់ប៉ារ៉ាម៉ែត្រក្នុងម៉ូឌែលនិង test data សម្រាប់ធ្វើការវាយតម្លៃ។

ក្នុងរូបទី១(b) បង្ហាញពីលទ្ធផលនៅពេលដែលយើងអនុវត្តម៉ូឌែលជាពហុធាដឺក្រេទី១(បន្ទាត់)និងពហុធាដឺក្រេខ្ពស់(ខ្សែកោង)។ យើងអាចពិនិត្យឃើញថា នៅពេលយើងជ្រើសយកម៉ូឌែលសាមញ្ញបំផុតពោលគឺបន្ទាត់ដឺក្រេទី១ នោះកម្រិតនៃការពណ៌នារបស់ម៉ូឌែលទៅលើទិន្នន័យ (Coefficient of determination: R^2)មានតម្លៃទាបដែលបង្ហាញថាមិនអាចពន្យល់បានល្អឡើយចំពោះទិន្នន័យដែលមាន។

ផ្ទុយទៅវិញ នៅពេលដែលយើងតម្លើងដឺក្រេនៃម៉ូឌែលកាន់តែខ្ពស់ យើងពិនិត្យឃើញថាតម្លៃនៃR^2មានការកើនឡើងខ្លះដែលអាចឱ្យយើងនិយាយបានថាវាពន្យល់លើទំនាក់ទំនងរបស់ទិន្នន័យបានប្រសើរ។ ប៉ុន្តែតើការតម្លើងម៉ូឌែលឱ្យកាន់តែស្មុគស្មាញ(តម្លើងដឺក្រេ)វាពិតជាផ្តល់ឱ្យយើងនូវម៉ូឌែលដែលល្អមែនឬ?



(a) (b)

រូបទី១ ទិន្នន័យនិងម៉ូឌែលតម្រែតម្រង់

នៅក្នុងរូបទី១(b)បើយើងពិនិត្យលើទិន្នន័យដែលមិនត្រូវបានប្រើក្នុងការកំណត់តម្លៃមេគុណតម្រែតម្រង់(test data)នោះយើងឃើញថា ម៉ូឌែលដែលមានដឺក្រេលំដាប់ខ្ពស់ឬម៉ូឌែលដែលស្មុគស្មាញខ្លាំងមិនអាចប៉ាន់ស្មាន ឬពន្យល់ទំនាក់ទំនងរវាងអថេរពន្យល់និងអថេរគោលដៅបានល្អឡើយបើប្រៀបធៀបជាមួយម៉ូឌែលដែលមានដឺក្រេទាបជាងវា។

ហេតុនេះ តើយើងគួរធ្វើបែបណាដើម្បីជ្រើសបានម៉ូឌែលដែលអាចពន្យល់បានល្អទាំងចំពោះទិន្នន័យដែលប្រើក្នុងដំណាក់កាលកំណត់ប៉ារ៉ាម៉ែត្រ(learning) trainig data
និងទាំងចំពោះទិន្នន័យដែលមិនត្រូវបានប្រើនៅដំណាក់កាលlearning(test data)?

1. Regularization

ដើម្បីដោះស្រាយបញ្ហានេះ Regularization ត្រូវបានប្រើប្រាស់។ ក្នុងវិធីសាស្ត្រនេះ ភាពស្មុគ
ស្មាញនៃម៉ូឌែលត្រូវបានគិតគូររួមគ្នាជាមួយនិងតម្លៃនៃកម្រិតល្អៀងរបស់ម៉ូឌែល។ ឧទាហរណ៍ក្នុងករណីម៉ូឌែលតម្រែតម្រង់លីនេអ៊ែរដែលយើងបានសិក្សា
កន្លងមកនេះ ការធ្វើបរមាកម្មលើតម្លៃល្អៀងត្រូវបានផ្លាស់ប្តូរទៅជាទម្រង់ $L(\boldsymbol{\beta}, \alpha)$ដូចខាងក្រោម
។ នៅទីនេះ ផ្នែក $R(\boldsymbol{\beta})$ គឺជាផ្នែកដែលបង្ហាញពីកម្រិតនៃភាពស្មុគស្មាញរបស់ម៉ូឌែល ហើយ
α ជាមេគុណដែលប្រើដើម្បីកម្រិតឥទ្ធិពលនៃ $R(\boldsymbol{\beta})$ ពេលធ្វើបរមាកម្ម។

$$\text{ម៉ូឌែល } \boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$
$$L(\boldsymbol{\beta}, \alpha) = E(\boldsymbol{\beta}) + \alpha R(\boldsymbol{\beta})$$

ផ្នែក $R(\boldsymbol{\beta}) : \boldsymbol{\beta} = (\beta_1, \dots, \beta_d)^\top$ ដែលបង្ហាញពីកម្រិតនៃភាពស្មុគស្មាញ
របស់ម៉ូឌែលត្រូវបាន បង្ហាញជាទម្រង់នានាដូចជា

$$Ridge\ penalty(L2\ regularization):\ R(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|^2 = \beta_1^2 + \cdots + \beta_d^2$$

$$L1\ regularization:\ R(\boldsymbol{\beta}) = |\beta_1 + \cdots + \beta_d|$$

ក្នុងអត្ថបទនេះ យើងនឹងណែនាំអំពី $L2\ regularization$
ចំពោះម៉ូឌែលតម្រែតម្រង់ដែលហៅថា Ridge Regression Model។

2. Ridge Regression Model

នៅក្នុង Ridge Regression Model តម្លៃការរនៃណមរបស់វ៉ិចទ័រមេគុណតម្រែតម្រង់ ត្រូវបានប្រើប្រាស់សម្រាប់បង្ហាញពីកម្រិតស្មុគស្មាញរបស់ម៉ូឌែល។ ក្នុងករណីនេះ ដើម្បីកំណត់តម្លៃមេគុណតម្រែតម្រង់ យើងនឹងធ្វើអប្បបរមាកម្មលើ អនុគមន៍ដែលកំណត់ដូចខាងក្រោម។

$$L(\boldsymbol{\beta}, \alpha) = E(\boldsymbol{\beta}) + \alpha R(\boldsymbol{\beta})$$

$$= (\boldsymbol{y} - X\boldsymbol{\beta})^{\top}(\boldsymbol{y} - X\boldsymbol{\beta}) + \alpha\|\boldsymbol{\beta}\|^2$$

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} L(\boldsymbol{\beta}, \alpha)$$

តាមរយៈការធ្វើបរមាកម្មបែបនេះ យើងនឹងអាចកំណត់បាននូវម៉ូឌែលដែលមានកម្រិតលម្អៀងតូចព្រមទាំងទំហំ នៃមេគុណតម្រែតម្រង់ (ដែលយើងសន្មតថាជាកម្រិតភាពស្មុគស្មាញក្នុងករណីនេះ) បានព្រមគ្នា។នៅទីនេះដើម្បីធ្វើតុល្យកម្មរវាងកម្រិតលម្អៀងរបស់ម៉ូឌែលនិងភាពស្មុគស្មាញ(ទំហំនៃមេគុណតម្រែតម្រង់)យើងអាចកែសម្រួលតម្លៃនៃមេគុណ α បាន។

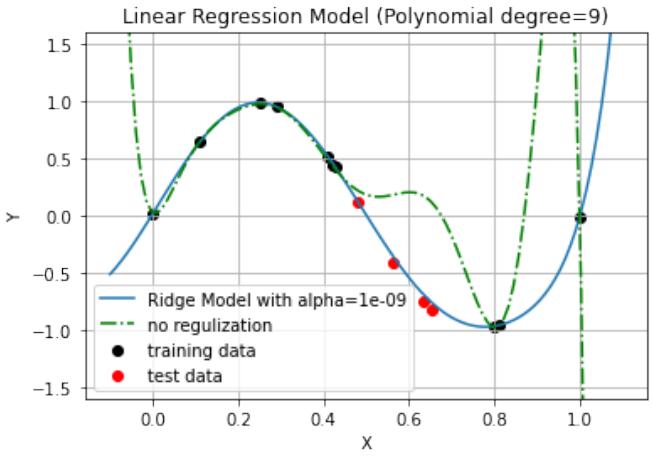
ដើម្បីដោះស្រាយបង្ហាញបរមាកម្មខាងលើ ដូចក្នុងអត្ថបទមុនៗដែរ យើងអាចដោះស្រាយតាមគណិតវិទ្យាវិភាគដោយធ្វើដេរីវេរួចរកតម្លៃនៃអថេរត្រង់ដេរីវេស្មើសូន្យ (ករណីអនុគមន៍ប៉ោង) ឬ ដោះស្រាយដោយប្រើវិធីសាស្ត្រSGD។

ចំពោះចម្លើយតាមគណិតវិទ្យាវិភាគ ទម្រង់នៃមេគុណតម្រែតម្រង់ត្រូវបានគណនាដូចខាងក្រោម(ដំណោះស្រាយទុកជាកិច្ចការផ្ទះជូនមិត្តអ្នកអាន) ដែល I_d ជាម៉ាទ្រីសឯកតា។

$$\widehat{\boldsymbol{\beta}} = (X^{\top}X + \alpha I_d)^{-1}X^{T}\boldsymbol{y}$$

ចំពោះចម្លើយតាមSGD ការផ្លាស់ប្តូរតម្លៃមេគុណតម្រែតម្រង់ត្រូវបានគណនាដូចខាងក្រោម (ដំណោះស្រាយទុកជាកិច្ចការផ្ទះជូនមិត្តអ្នកអាន) ដែលN ជាចំនួនtraining data និងη_t ជា learing rate ។

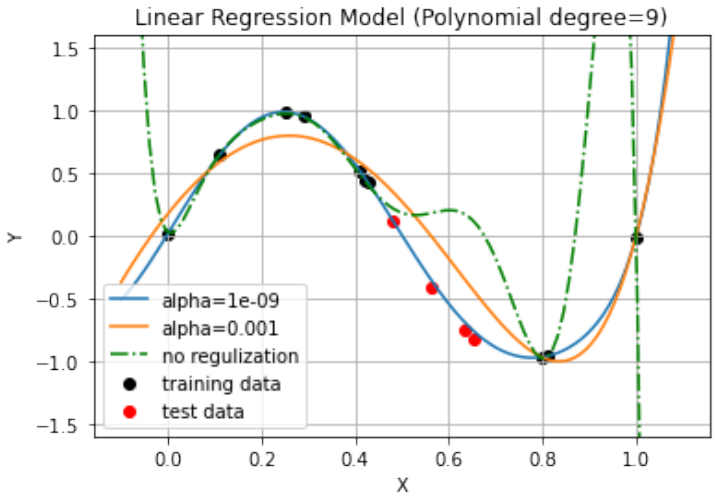
$$\boldsymbol{\beta}^{(t+1)} = \left(1 - \frac{2\alpha\eta_t}{N}\right)\boldsymbol{\beta}^{(t)} - 2\eta_t\left(\hat{y}_i{}^{(t)} - y\right)\boldsymbol{x}_i^{\top}$$



រូបទី២ ការប្រៀបធៀបរវាងម៉ូឌែលដែលប្រើនិងមិនប្រើRegularization

រូបទី២បង្ហាញពីលទ្ធផលនៅពេល L2 regularization ត្រូវបានប្រើលើម៉ូឌែលតម្រែតម្រង់ជាទម្រង់ពហុធាដឺក្រេទី៩។ យើងពិនិត្យឃើញថា នៅពេល regularization ត្រូវបានប្រើ ម៉ូឌែលអាចពន្យល់បានល្អទាំងចំពោះ training data និង test data ព្រមគ្នា ផ្ទុយពីម៉ូឌែលដែលមិនប្រើ regularization។

តាមពិតទៅយើងនៅសល់បញ្ហាមួយទៀតគឺការកំណត់តម្លៃនៃមេគុណ α ។ ដំណោះស្រាយក្នុងបញ្ហានេះអាចធ្វើបានតាមរយៈការសាកល្បងលើតម្លៃជាច្រើននៃ α ចំពោះទិន្នន័យមួយផ្នែកដែលមិនមែនជាtest data , training data ដែលយើងហៅថា validation data ។ យើងអាចកំណត់តម្លៃ α ដោយជ្រើសយកតម្លៃ α ណាដែលធ្វើឲ្យស្ថានភាពនៃម៉ូឌែលល្អបំផុតចំពោះ validation data។ រូបទី៣បង្ហាញពីការប្រៀបធៀបចំពោះតម្លៃមួយចំនួននៃ α ។



រូបទី៣ ការប្រៀបធៀបលើមេគុណ α

ជាមួយPythonអ្នកអាចសរសេរCodeងាយៗដូចខាងក្រោម។

```
import numpy as np
```

ករណីដំណោះស្រាយតាមគណិតវិទ្យាវិភាគ

```
def ridge_fit(x,y,k,alpha):
  X_ = np.zeros((len(x),k+1))
  for i in range(k+1):
    X_[:,i] = x**i
  beta = np.linalg.inv(X_.T@X_+alpha*np.eye(k+1))@X_.T@y
  return beta
```

ករណីដំណោះស្រាយតាមSGD

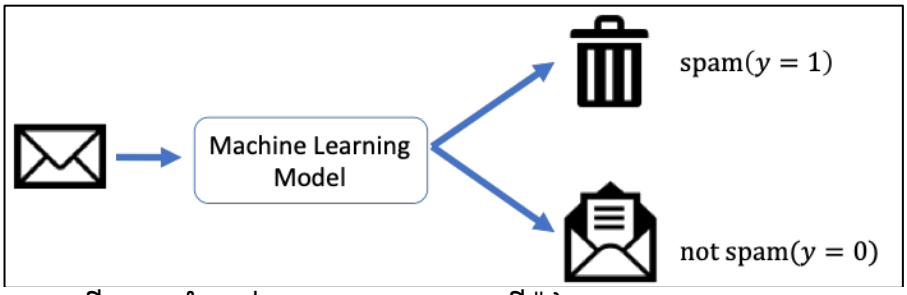
```
# learning with SGD
def ridge_sgd_fit(x,y,k,alpha):
  beta = np.zeros(k+1)
  d_index = list(range(len(x)))

  eta = 1e-4
  for t in range(500000):
    random.shuffle(d_index)
    for i in d_index :
      xi = np.zeros(k+1)
      for j in range(k+1):
        xi[j] = x[i]**j
      y_hat = xi.T @ beta
      beta = (1-2*alpha*eta/len(x))*beta - 2 * eta * (y_hat - y[i]) *
xi
  return beta
```

បញ្ហាចំណាត់ថ្នាក់២ក្រុម(Binary Classification Problem)

ក្រៅពីការសិក្សាពីទំនាក់ទំនងរវាងអថេរពីរឬច្រើនតាមរយៈម៉ូឌែលតម្រែតម្រង់ ការបែងចែកទិន្នន័យដោយផ្អែកលើអថេរពន្យល់របស់វាជាក្រុមហៅថាចំណាត់ថ្នាក់ក្រុម(classification) ។ តាមពិតចំណាត់ថ្នាក់ក្រុម
ក៏អាចបង្ហាញបានជាទម្រង់ម៉ូឌែលតម្រែតម្រង់ផងដែរ ដោយគ្រាន់តែតម្លៃនៃអថេរគោលដៅមិនយកតម្លៃទូលាយលើសំណុំចំនួនពិតឡើយ ផ្ទុយទៅវិញគឺយកតម្លៃដាច់ដូចជា{0,1}ជាដើម ។

ក្នុងអត្ថបទនេះយើងនឹងណែនាំករណីដែលអថេរគោលដៅយកតម្លៃតែពីរប្រភេទដែលយើងកំណត់ហៅថាចំណាត់ថ្នាក់២ក្រុម(binary classification) ។ ឧទាហរណ៍នៃការអនុវត្តចំណាត់ថ្នាក់២ក្រុមក្នុងជីវភាពមានដូចជា៖ ការកំណត់អត្តសញ្ញាណសារអេឡិចត្រូនិចរំខាន(spam or not) ការវិភាគជម្ងឺតាមរយៈរោគសញ្ញា(មានជម្ងឺឬគ្មាន)
ការទស្សន៍ទាយលទ្ធផលបោះឆ្នោត(ជាប់ឬធ្លាក់)ជាដើម ។

ជាឧទាហរណ៍យើងនឹងលើកយកការកំណត់អត្តសញ្ញាណសារអេឡិចត្រូនិចរំខាន(spam or not)មកបង្ហាញដើម្បីស្វែងយល់បន្ថែមពីដំណោះស្រាយក្នុងបញ្ហាចំណាត់ថ្នាក់ក្រុមតាមរយៈ machine learning ។ រូបទី១បង្ហាញពីដំណើរការនៃការកំណត់សារំខាន ។



រូបទី១ ការកំណត់spam mailដោយប្រើម៉ូឌែលmachine learning

1. ចំណាត់ថ្នាក់២ក្រុមដោយម៉ូឌែលលីនេអ៊ែរ

ម៉ូឌែលលីនេអ៊ែរនៃចំណាត់ថ្នាក់២ក្រុមគឺជាម៉ូឌែលដែលទស្សន៍ទាយប្រភេទនៃទិន្នន័យ $\hat{y} \in \{0,1\}$ ដោយកំណត់តាមតម្លៃជ្ជមានឬអវិជ្ជមាននៃផលគុណស្កាលែរវាងធាតុចូល(input)$\boldsymbol{x} \in \mathbb{R}^d$ និងប៉ារ៉ាម៉ែត្រម៉ូឌែល$\boldsymbol{w} \in \mathbb{R}^d$ ។

$$\hat{y} = \begin{cases} 1 & (\boldsymbol{x}^\top \boldsymbol{w} > 0) \\ 0 & (\boldsymbol{x}^\top \boldsymbol{w} \leq 0) \end{cases}$$

ក្នុងករណីការកំណត់spam mail យើងអាចកំណត់ជាទម្រង់ម៉ូឌែលលីនេអ៊ែរខាងលើបាន ដោយកំណត់យក $\hat{y} = 1$ សម្គាល់spam mail និង $\hat{y} = 0$សម្គាល់សារធម្មតា ។

ដើម្បីងាយស្រួលយល់ពីម៉ូឌែលនេះ យើងមកមើលឧទាហរណ៍ងាយមួយក្នុងការកំណត់ spam
mailដូចខាងក្រោម ។ សន្មតថា វ៉ិចទ័រធាតុចូល(អត្ថបទសារ)មានវិមាត្រ10 ។ យើងកំណត់វ៉ិចទ័រនេះ បង្ហាញពីវត្តមាននៃពាក្យគន្លឹះចំនួន១០ ដោយយកតម្លៃ0 ឬ 1 ត្រង់ពាក្យគន្លឹះនិមួយៗប្រសិនបើពាក្យគន្លឹះនោះគ្មានឬមានវត្តមានក្នុងអត្ថបទសារ ។

ឧទាហរណ៍ ពាក្យគន្លឹះដែលរៀបចំទុកមាន “assignment”, “boy”, “file”, “hello”, “love”, “my”, “photo”, “password”, “school” , “text” ។ ក្នុងករណីនេះ អត្ថបទសារ “Hello my boy, I sent you my photo in the attached file” អាចបង្ហាញជាទម្រង់វ៉ិចទ័រ $\boldsymbol{x} \in \mathbb{R}^d$ បានដូចខាងក្រោម

$$\boldsymbol{x} = (0 \quad 1 \quad 1 \quad 1 \quad 0 \quad 1 \quad 1 \quad 0 \quad 0 \quad 0)^\top$$

ចំពោះការប៉ាន់ស្មានដោយម៉ូឌែលលីនេអ៊ែរខាងលើ ចំពោះប៉ារ៉ាម៉ែត្រ $\boldsymbol{w} = (w_1 \quad \cdots \quad w_{10})^\top$ ផលគុណស្កាលែក្នុងករណីអត្ថបទសារខាងលើត្រូវបានគណនាដូចខាងក្រោម ។ ក្នុងករណីនេះប្រសិនបើតម្លៃផលគុណស្កាលែនេះវិជ្ជមាននោះ អត្ថបទនេះត្រូវបានកំណត់ថាជាspam mail ។

$$\boldsymbol{x}^\top \boldsymbol{w} = (0 \quad 1 \quad 1 \quad 1 \quad 0 \quad 1 \quad 1 \quad 0 \quad 0 \quad 0) \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_9 \\ w_{10} \end{pmatrix} = w_2 + w_3 + w_4 + w_6 + w_7$$

ប៉ុន្តែដូចដែលអ្នកកត់សម្គាល់បាន បញ្ហាសម្រាប់យើងគឺថា តើនឹងត្រូវកំណត់តម្លៃនៃប៉ារ៉ាម៉ែត្ររបស់ម៉ូឌែលដោយរបៀបណា ។ មានវិធីសាស្ត្រជាច្រើនត្រូវបានប្រើ តែក្នុងអត្ថបទនេះយើងនឹងណែនាំម៉ូឌែលតម្រែតម្រង់Logistic ។

2. តម្រែតម្រង់Logistic (Logistic Regression)

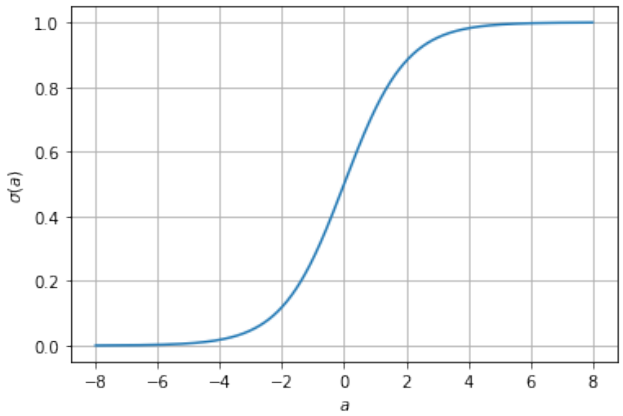
តម្រែតម្រង់Logisticគឺជាម៉ូឌែលលីនេអ៊ែរនៃចំណាត់ថ្នាក់២ក្រុមមួយប្រភេទ ដែលប្រូបាបមាន
លក្ខខណ្ឌ$p(y|\boldsymbol{x})$ពោលគឺប្រូបាបដែលប្រភេទនៃទិន្នន័យyត្រូវបានទស្សន៍ទាយចំពោះអថេរពន្យល់ $\boldsymbol{x}$ ត្រូវបានគណនាដូចខាងក្រោម ។

$$P(\hat{y}=1|\boldsymbol{x}) = \sigma(\boldsymbol{x}^\top \boldsymbol{w}) = \frac{\exp(\boldsymbol{x}^\top \boldsymbol{w})}{1+\exp(\boldsymbol{x}^\top \boldsymbol{w})} = \frac{1}{1+\exp(-\boldsymbol{x}^\top \boldsymbol{w})}$$

$$P(\hat{y}=0|\boldsymbol{x}) = 1 - P(\hat{y}=1|\boldsymbol{x}) = 1 - \sigma(\boldsymbol{x}^\top \boldsymbol{w}) = \sigma(-\boldsymbol{x}^\top \boldsymbol{w})$$

នៅទីនេះ $\sigma(x)$ ជាអនុគមន៍sigmaដែលត្រូវបានគណនាដូចខាងក្រោម។

$$\sigma(x) = \frac{\exp(x)}{1+\exp(x)} = \frac{1}{1+\exp(-x)}$$



រូបទី២ ក្រាបនៃអនុគមន៍Sigmoid

ចំពោះទិន្នន័យ$\boldsymbol{x}$ ក្នុងករណីដែលប្រូបាប$P(\hat{y}=1|\boldsymbol{x}) > 0.5$ នោះទិន្នន័យត្រូវបានទស្សន៍ទាយថាក្នុងចំណាត់ថ្នាក់ក្រុម$\hat{y}=1$ ។ការកំណត់បែបនេះគឺសមមូលគ្នានឹងលក្ខខណ្ឌផលគុណស្កាលែរវិជ្ជមានដែលបានបង្ហាញខាងដើម។

$$P(\hat{y}=1|\boldsymbol{x}) > 0.5 \Leftrightarrow \frac{1}{1+\exp(-\boldsymbol{x}^\top \boldsymbol{w})} > \frac{1}{2} \Leftrightarrow \exp(-\boldsymbol{x}^\top \boldsymbol{w}) < 1 \Leftrightarrow \boldsymbol{x}^\top \boldsymbol{w} > 0$$

3. កម្រិតសាកសមនៃទិន្នន័យ Likelihood

ឧបមាថា ប៉ារ៉ាម៉ែត្រនៃម៉ូឌែលតម្រែតម្រង់Logisticត្រូវបានកំណត់ដូចខាងក្រោម។

$$\boldsymbol{w} = (-2 \quad 1 \quad 1 \quad 0 \quad 1 \quad 0 \quad 1 \quad 1 \quad -1 \quad 0)^\top$$

ក្នុងករណីនេះ តម្លៃផលគុណស្កាលែរវាងវ៉ិចទ័រអថេរពន្យល់នៃទិន្នន័យ $\boldsymbol{x}$ និងប៉ារ៉ាម៉ែត្រអាច

គណនានិងបកស្រាយជាទម្រង់ប្រូបាបដូចខាងក្រោម។

$$\boldsymbol{x}^{\top}\boldsymbol{w} = (0 \quad 1 \quad 1 \quad 1 \quad 0 \quad 1 \quad 1 \quad 0 \quad 0 \quad 0)\begin{pmatrix} -2 \\ 1 \\ \vdots \\ -1 \\ 0 \end{pmatrix} = 3$$

$$P(\hat{y} = 1|\boldsymbol{x}) = \sigma(3) = \frac{1}{1 + \exp(-3)} = 0.95$$

ចំពោះលទ្ធផលនេះយើងអាចបកស្រាយថា ចំពោះអត្ថបទសារដែលបានផ្តល់ ម៉ូឌែលបានប៉ាន់ស្មានថាជាspam
mailដោយតម្លៃប្រូបាប0.95។តម្លៃនេះធំជាង0.5 ហេតុនេះយើងថាម៉ូឌែលទស្សន៍ទាយថាវាជាspam mail។

ក្នុងការពិភាក្សាខាងលើមកដល់ត្រឹមនេះ ចំពោះទិន្នន័យ$\boldsymbol{x}$និងប៉ារ៉ាម៉ែត្រ$\boldsymbol{w}$ ដែលត្រូវបានផ្តល់ឲ្យការប៉ាន់ស្មានរបស់ម៉ូឌែលត្រូវបានគណនាតាមវិធីដែលបានរៀបរាប់ខាងលើ។ ពេលនេះយើងពិនិត្យករណីដែលប្រភេទទិន្នន័យត្រូវបានកំណត់ជាក់លាក់ តែប៉ារ៉ាម៉ែត្រ
អាចត្រូវបានផ្លាស់ប្តូរ។ នៅទីនេះ យើងសិក្សាលើកម្រិតសាកសមនៃទិន្នន័យដែលម៉ូឌែល
ប៉ាន់ស្មានបានដោយកំណត់ជាតម្លៃប្រូបាប$\hat{l}_{(\boldsymbol{x},y)}(\boldsymbol{w})$ដូចខាងក្រោម
និងសន្មតហៅថា កម្រិតសាកសម Likelihood។ពោលគឺកម្រិតដែលម៉ូឌែលអាចប៉ាន់ស្មានបានត្រឹមត្រូវ(គិតជាភាគរយ)លើប្រភេទទិន្នន័យនិមួយៗ។

$$\hat{l}_{(\boldsymbol{x},y)}(\boldsymbol{w}) = P(\hat{y} = y|\boldsymbol{x})$$

ឧទាហរណ៍ក្នុងករណីអត្ថបទសារខាងលើ សន្មតថាប្រភេទទិន្នន័យពិតគឺ $y = 1$ ។ចំពោះប៉ារ៉ាម៉ែត្រ$\boldsymbol{w}$ ដែលត្រូវបានផ្តល់ឲ្យ ការប៉ាន់ស្មានរបស់ម៉ូឌែលខាងលើគឺ $P(\hat{y} = 1|\boldsymbol{x}) = 0.95$
ហេតុនេះកម្រិតសាកសមLikelihood : $\hat{l}_{(\boldsymbol{x},1)}(\boldsymbol{w}) = P(\hat{y} = 1|\boldsymbol{x}) = 0.95$
ដែលមានន័យថា ម៉ូឌែលអាចបែងចែកថាជាspam
mailបានត្រឹមត្រូវក្នុងកម្រិត95%(ប្រូបាប0.95)។ ផ្ទុយទៅវិញ ឧបមាថាយើងមានអត្ថបទសារមិនមែនspam mail មួយផ្សេង “Please submit your assignment file by tomorrow morning” ។
នោះវ៉ិចទ័រ $\boldsymbol{x} = (1 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0)^{\top}$ ។

$$\boldsymbol{x}^{\top}\boldsymbol{w} = (1 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0)\begin{pmatrix} -2 \\ 1 \\ \vdots \\ -1 \\ 0 \end{pmatrix} = -1$$

$$P(\hat{y}=1|\boldsymbol{x})=\sigma(-1)=\frac{1}{1+\exp(1)}=0.27$$

ដោយចម្លើយពិតគឺមិនមែនជាspam mail (y=0) ហេតុនេះកម្រិតសាកសមLikelihood : $\hat{l}_{(x,0)}(\boldsymbol{w})=P(\hat{y}=0|\boldsymbol{x})=1-P(\hat{y}=1|\boldsymbol{x})=1-0.27=0.73$ ដែលមានន័យថាម៉ូឌែលអាចបែងចែកថាមិនមែនspam mailបានត្រឹមត្រូវក្នុងកម្រិត73%(ប្រូបាប0.73) ។

ចំពោះគ្រប់ប្រភេទទិន្នន័យ យើងអាចសរសេរកន្សោមកម្រិតសាកសមបានក្រោមទម្រង់

$$\hat{l}_{(x,y)}(\boldsymbol{w})=P(\hat{y}=y|\boldsymbol{x})=\begin{cases}P(\hat{y}=1|\boldsymbol{x}) & (y=1)\\ P(\hat{y}=0|\boldsymbol{x}) & (y=0)\end{cases}=\pi^{y}(1-\pi)^{1-y}$$

ដែល $\pi=P(\hat{y}=1|\boldsymbol{x})=\sigma(\boldsymbol{x}^{\top}\boldsymbol{w})$ ។

4. ការប៉ាន់ស្មានមេគុណតម្រែតម្រង់តាម Maximum Likelihood

នៅក្នុងម៉ូឌែលតម្រែតម្រង់លីនេអ៊ែរ យើងបានកំណត់តម្លៃមេគុណតម្រែតម្រង់ដោយធ្វើអប្បបរមាកម្មលើតម្លៃកម្រិតលម្អៀងរវាងតម្លៃប៉ាន់ស្មានដោយម៉ូឌែលនិងតម្លៃពិតប្រាកដនៃអថេរគោលដៅ ។ស្រដៀងគ្នានេះ ក្នុងម៉ូឌែលតម្រែតម្រង់Logistic
យើងអាចកំណត់តម្លៃនៃប៉ារ៉ាម៉ែត្ររបស់ម៉ូឌែលដោយធ្វើអតិបរមាកម្មលើកម្រិតសាកសមរបស់ទិន្នន័យដែលម៉ូឌែលអាចប៉ាន់ស្មានបាន ។ វិធីសាស្ត្រនេះត្រូវបានគេហៅថា Maximum Likelihood Estimation(MLE) ។

ចំពោះសំណុំទិន្នន័យទាំងអស់$\mathcal{D}$ដែលមានចំនួនN យើងកំណត់កម្រិតសាកសមនៃទិន្នន័យ(likelihood)ពោលគឺកម្រិតដែលម៉ូឌែលអាចប៉ាន់ស្មានបានត្រឹមត្រូវចំពោះគ្រប់ទិន្នន័យទាំងអស់ដោយកន្សោមខាងក្រោម ។ នៅទីនេះយើងសន្មតថារបាយនៃគ្រប់ទិន្នន័យទាំងអស់គឺឯករាជ្យនិងមានឯកសណ្ឋានភាព(i.i.d : independent and identically dristributed) ។

$$\hat{L}_{\mathcal{D}}(\boldsymbol{w})=\prod_{i=1}^{N}\hat{l}_{(\boldsymbol{x}_i,y_i)}(\boldsymbol{w})$$

ដោយសារ កម្រិតសាកសមនៃទិន្នន័យ(likelihood)គឺជាតម្លៃប្រូបាប ហេតុនេះតម្លៃរបស់វាតូចខ្លាំង
ដែលធ្វើឱ្យតម្លៃផលគុណកាន់តែតូចខ្លាំងពេលចំនួនទិន្នន័យមានច្រើន ។ ដើម្បីបញ្ចៀសនូវបញ្ហាតម្លៃតូចពេកក្នុងការគណនាជាមួយកុំព្យូទ័រ នៅទីនេះយើងសិក្សាបរមាកម្មលើតម្លៃលោការីតរបស់វា ។ ការធ្វើបែបនេះមិនប៉ះពាល់ដល់ការធ្វើបរមាកម្មឡើយ
ព្រោះអនុគមន៍លោការីតជាអនុគមន៍កើនដាច់ខាត ។

$$\log \hat{L}_{\mathcal{D}}(\boldsymbol{w}) = \log \prod_{i=1}^{N} \hat{l}_{(\boldsymbol{x}_i, y_i)}(\boldsymbol{w}) = \sum_{i=1}^{N} \log \hat{l}_{(\boldsymbol{x}_i, y_i)}(\boldsymbol{w})$$

ដើម្បីងាយស្រួលក្នុងការដោះស្រាយបញ្ហាបរមាកម្ម យើងប្តូរពីការធ្វើអតិបរមាកម្មលើLikelihood ទៅជាការធ្វើអប្បបរមាកម្មដោយគុណកន្សោមខាងលើ -1 ។

$$\hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{w}) = -\log \hat{L}_{\mathcal{D}}(\boldsymbol{w}) = -\sum_{i=1}^{N} \log \hat{l}_{(\boldsymbol{x}_i, y_i)}(\boldsymbol{w})$$

ក្នុងករណីយើងចង់សិក្សាបន្ថែមដោយបញ្ចូលផ្នែក Regularization (Ridge) ចូលក្នុងម៉ូឌែល Likehoodដែលត្រូវធ្វើបរមាកម្ម អាចប្តូរទៅសរសេរជាទម្រង់ខាងក្រោម ។

$$\hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{w}) = -\log \hat{L}_{\mathcal{D}}(\boldsymbol{w}) + \alpha \|\boldsymbol{w}\|_2^2 = -\sum_{i=1}^{N} \log \hat{l}_{(\boldsymbol{x}_i, y_i)}(\boldsymbol{w}) + \alpha \|\boldsymbol{w}\|_2^2 \quad (\alpha > 0)$$

5. ការដោះស្រាយតាមរយៈវិធី SGD

ដើម្បីធ្វើអប្បបរមាកម្ម$\hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{w})$ យើងនឹងប្រើប្រាស់វិធីSGDដែលបានសិក្សាក្នុងអត្ថបទមុន ។
ជាដំបូងយើងពិនិត្យលើអនុគមន៍ដេរីវេ $\frac{\partial}{\partial \boldsymbol{w}} \log \hat{l}_{(\boldsymbol{x},y)}(\boldsymbol{w})$ ។

$$\log \hat{l}_{(\boldsymbol{x},y)}(\boldsymbol{w}) = \log(\pi^y (1-\pi)^{1-y}) = y \log \pi + (1-y) \log(1-\pi)$$

$$\frac{\partial}{\partial \boldsymbol{w}} \log \hat{l}_{(\boldsymbol{x},y)}(\boldsymbol{w}) = \frac{y}{\pi} \frac{\partial \pi}{\partial \boldsymbol{w}} + \frac{1-y}{1-\pi} \times \left(-\frac{\partial \pi}{\partial \boldsymbol{w}}\right) = \frac{y-\pi}{\pi(1-\pi)} \frac{\partial \pi}{\partial \boldsymbol{w}}$$

បន្ទាប់ពីនេះ ដើម្បីគណនា $\frac{\partial \pi}{\partial \boldsymbol{w}}$ យើងពិនិត្យលើដេរីវេនៃអនុគមន៍Sigmoid ។

$$\frac{\partial}{\partial a} \sigma(a) = \frac{\partial}{\partial a} \left\{ \frac{1}{1+\exp(-a)} \right\} = -\frac{\frac{\partial}{\partial a} \exp(-a)}{(1+\exp(-a))^2} = \frac{1}{1+\exp(-a)} \times \frac{\exp(-a)}{1+\exp(-a)}$$

$$\frac{\partial}{\partial a} \sigma(a) = \sigma(a)\big(1-\sigma(a)\big)$$

$$\frac{\partial \pi}{\partial \boldsymbol{w}} = \frac{\partial}{\partial \boldsymbol{w}} P(\hat{y}=1|\boldsymbol{x}) = \frac{\partial}{\partial \boldsymbol{w}} \sigma(\boldsymbol{x}^\top \boldsymbol{w}) = \sigma(\boldsymbol{x}^\top \boldsymbol{w})\big(1-\sigma(\boldsymbol{x}^\top \boldsymbol{w})\big) = \pi(1-\pi)$$

នៅទីនេះ $a = \boldsymbol{x}^\top \boldsymbol{w}$, $\frac{\partial a}{\partial \boldsymbol{w}} = \frac{\partial}{\partial \boldsymbol{w}}(\boldsymbol{x}^\top \boldsymbol{w}) = \boldsymbol{x}$ ហេតុនេះ

$$\frac{\partial}{\partial \boldsymbol{w}} \log \hat{l}_{(\boldsymbol{x},y)}(\boldsymbol{w}) = \frac{y-\pi}{\pi(1-\pi)} \frac{\partial \pi}{\partial \boldsymbol{w}} = \frac{y-\pi}{\pi(1-\pi)} \frac{\partial \pi}{\partial a} \frac{\partial a}{\partial \boldsymbol{w}} = \frac{y-\pi}{\pi(1-\pi)} \pi(1-\pi)\boldsymbol{x}$$

$$\frac{\partial}{\partial \boldsymbol{w}} \log \hat{l}_{(\boldsymbol{x},y)}(\boldsymbol{w}) = (y-\pi)\boldsymbol{x}$$

ដូចនេះ តាមរយៈវិធីSGD តម្លៃនៃប៉ារ៉ាម៉ែត្រ $\boldsymbol{w}$ ដែលធ្វើបរមាកម្មលើតម្លៃកម្រិតសាកសមនៃទិន្នន័យត្រូវបានគណនាដោយផ្លាស់ប្តូរតម្លៃដូចខាងក្រោម ។

$$\boldsymbol{w}^{(t+1)} = \boldsymbol{w}^{(t)} - \eta_t \frac{\partial}{\partial \boldsymbol{w}} \{-\log \hat{l}_{(\boldsymbol{x},y)}(\boldsymbol{w})\}\Big|_{\boldsymbol{w}=\boldsymbol{w}^{(t)}}$$

$$\boldsymbol{w}^{(t+1)} = \boldsymbol{w}^{(t)} + \eta_t \frac{\partial}{\partial \boldsymbol{w}} \log \hat{l}_{(\boldsymbol{x},y)}(\boldsymbol{w})\Big|_{\boldsymbol{w}=\boldsymbol{w}^{(t)}}$$

$$\boldsymbol{w}^{(t+1)} = \boldsymbol{w}^{(t)} + \eta_t \big(y - \pi^{(t)}\big)\boldsymbol{x}$$

នៅទីនេះ $\pi^{(t)}$ ជាតម្លៃប្រូបាបដែលគណនាដោយម៉ូឌែលជាមួយតម្លៃប៉ារ៉ាម៉ែត្រនៅដំណាក់កាលt នៃការផ្លាស់ប្តូរតម្លៃប៉ារ៉ាម៉ែត្រ ពោលគឺ $\pi^{(t)} = \sigma\big(\boldsymbol{x}^\top \boldsymbol{w}^{(t)}\big)$ ។ η_t ជាlearning-rate និង y ជាចំណាត់ថ្នាក់ក្រុមពិតប្រាកដនៃទិន្នន័យ $\boldsymbol{x}$ ។

ក្នុងករណីប្រើRidge Regularization កន្សោមខាងលើនឹងប្រែទៅជាទម្រង់ខាងក្រោម ។

$$\boldsymbol{w}^{(t+1)} = \left(1 - \frac{2\alpha\eta_t}{N}\right)\boldsymbol{w}^{(t)} + \eta_t \big(y - \pi^{(t)}\big)\boldsymbol{x}$$

6. ការវាយតម្លៃ

កាលពីសិក្សាម៉ូឌែលតម្រែតម្រង់លីនេអ៊ែរ យើងវាយតម្លៃម៉ូឌែលតាមរយៈតម្លៃកម្រិតលម្អៀង ឬ
មេគុណR^2 ។ នៅករណីម៉ូឌែលLogisticចំពោះបញ្ហាចំណាត់ថ្នាក់ក្រុមនេះ យើងអាចវាយតម្លៃតាមរយៈតម្លៃLikelihood បាន ។ ប៉ុន្តែការសិក្សាលើតម្លៃLikelihood មានការពិបាកក្នុងការបកស្រាយភ្ជាប់នឹងជីវភាពរស់នៅរបស់យើង ។ហេតុនេះក្នុងបញ្ហាចំណាត់ថ្នាក់ក្រុមរង្វាស់សម្រាប់រង្វាយតម្លៃលើម៉ូឌែលត្រូវបានគណនាដូចខាងក្រោម ។

		ចំណាត់ថ្នាក់ក្រុមពិតប្រាកដនៃទិន្នន័យសម្រាប់វាយតម្លៃ		
		$y=1$	$y=0$	សរុប
ចំណាត់ថ្នាក់ក្រុមទស្សន៍ទាយដោយម៉ូឌែល	$y=1$	**TP** (True Positive)	**FP** (False Positive)	ចំនួនករណីដែលត្រូវបានទស្សន៍ទាយថា $y=1$
	$y=0$	**FN** (False Negative)	**TN** (True Negative)	ចំនួនករណីដែលត្រូវបានទស្សន៍ទាយថា $y=0$
	សរុប	ចំនួនទិន្នន័យ $y=1$	ចំនួនទិន្នន័យ $y=0$	ចំនួនទិន្នន័យសរុប

- Accuracy = $\frac{\text{ចំនួនករណីដែលម៉ូឌែលទស្សន៍ទាយបានត្រឹមត្រូវ}}{\text{ចំនួនករណីសរុប}} = \frac{TP+TN}{TP+TN+FP+FN}$
- Precision = $\frac{\text{ចំនួនករណីដែលម៉ូឌែលទាយថា } y=1 \text{ បានត្រឹមត្រូវ}}{\text{ចំនួនករណីដែលម៉ូឌែលទាយថា } y=1} = \frac{TP}{TP+FP}$
- Recall = $\frac{\text{ចំនួនករណីដែលម៉ូឌែលទាយថា } y=1 \text{ បានត្រឹមត្រូវ}}{\text{ចំនួនទិន្នន័យ } y=1} = \frac{TP}{TP+FN}$
- F1-score = $\frac{\text{Precision} \times \text{Recall}}{\frac{1}{2}(\text{Precision}+ \text{Recall})} = \frac{2\ \text{Precision} \times \text{Recall}}{\text{Precision}+ \text{Recall}}$

Accuracyគឺបង្ហាញពីអត្រានៃការទស្សន៍ទាយបានត្រឹមត្រូវរបស់ម៉ូឌែលដោយមិនបែងចែកចំណាត់ថ្នាក់ក្រុមរបស់ទិន្នន័យ ។អត្រានេះដូចគ្នានិងពិន្ទុដែលអ្នកឆ្លើយសំណួរបានត្រឹមត្រូវក្នុងការប្រឡងណាមួយដែរ ។

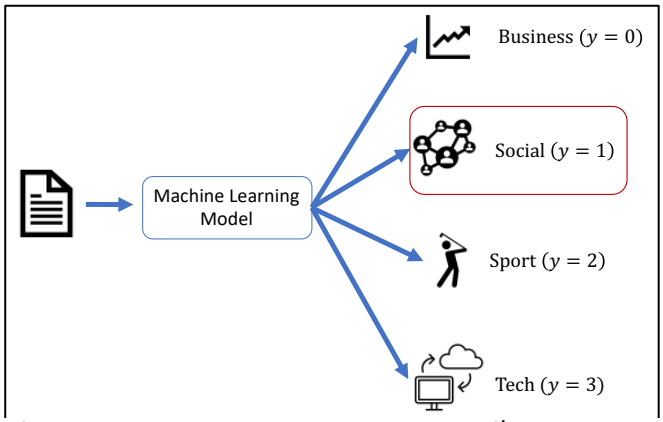
Precisionគឺសំដៅដល់អត្រានៃករណីដែលពិតជានៅក្នុងក្រុម $y=1$មែនក្នុងចំណោមករណីដែលម៉ូឌែលបានទស្សន៍ទាយថាស្ថិតក្នុងក្រុម$y=1$ ។ Recallសំដៅដល់អត្រានៃករណីដែលម៉ូឌែលបានទស្សន៍ទាយថាស្ថិតក្នុងក្រុម$y=1$ ក្នុងចំណោមទិន្នន័យក្នុងក្រុម $y=1$សរុប ។

ជាទូទៅវាជាការលំបាកក្នុងការបង្កើតម៉ូឌែលដែលមានទាំងPrecisionនិងRecallខ្ពស់ដូចគ្នា(trade-off relation) ។ហេតុនេះដើម្បីវាយតម្លៃរួមលើរង្វាស់ទាំងពីរនេះការធ្វើផលធៀបមធ្យមលើតម្លៃទាំងពីរត្រូវបានប្រើពេលគឺតម្លៃ F1-score ។

បញ្ហាចំណាត់ថ្នាក់ច្រើនក្រុម(Multiclass Classification Problem)

នៅក្នុងជីវភាពរស់នៅ ការធ្វើចំណាត់ថ្នាក់ក្រុម មានករណីជាច្រើនដែលចំនួនក្រុមត្រូវកំណត់មានច្រើនលើសពី២។ ឧទាហរណ៍ដូចជា ការបែងចែកអត្ថបទជាក្រុមតាមប្រធានបទ ការធ្វើកំណត់សម្គាល់ប្រភេទសម្ភារៈ ការកំណត់ប្រភេទវត្ថុយានយន្តលើផ្លូវរបស់យានយន្តបើកដោយស្វ័យប្រវត្តិជាដើម។ ដូចទៅនឹងការធ្វើចំណាត់ថ្នាក់២ក្រុមដែរ តម្លៃនៃអថេរគោលដៅក្នុងករណីចំណាត់ថ្នាក់ច្រើនក្រុមយកតម្លៃជាច់ដូចជា{0,1,2,3,.....}។ ករណីដែលអថេរគោលដៅយកតម្លៃច្រើនប្រភេទយើងកំណត់ហៅថាចំណាត់ថ្នាក់ច្រើនក្រុម(multiclass classification)។

ក្នុងអត្ថបទនេះ យើងនឹងលើកយកការកំណត់ប្រភេទអត្ថបទតាមប្រធានបទមកបង្ហាញដើម្បីស្វែងយល់បន្ថែមពីដំណោះស្រាយក្នុងបញ្ហាចំណាត់ថ្នាក់ច្រើនក្រុមតាមរយៈ machine learning ។ រូបទី១បង្ហាញពីដំណើរការនៃការកំណត់សារៈខាន។



រូបទី១ ការកំណត់ប្រភេទអត្ថបទតាមប្រធានបទដោយប្រើម៉ូឌែលMachine Learning

1. ចំណាត់ថ្នាក់ច្រើនក្រុមដោយម៉ូឌែលលីនេអ៊ែរ

ការបង្ហាញបញ្ហាចំណាត់ថ្នាក់ច្រើនក្រុមដោយប្រើម៉ូឌែលលីនេអ៊ែរអាចធ្វើបានដូចករណីនៃចំណាត់ថ្នាក់២ក្រុមដែរ គឺជាម៉ូឌែលដែលទស្សន៍ទាយប្រភេទនៃទិន្នន័យ $\hat{y} \in C = \{0,1,\dots,K\}$ ដោយកំណត់យកក្រុម j ណាដែលមានតម្លៃធំបំផុតនៃផលគុណស្កាលែរវាងធាតុចូល(input)$\boldsymbol{x} \in \mathbb{R}^d$ និងប៉ារ៉ាម៉ែត្រម៉ូឌែល$\boldsymbol{w}_j \in \mathbb{R}^d$របស់ក្រុម j នោះ ។

$$\hat{y} = \arg\max_{j \in C} \ \boldsymbol{x}^{\top}\boldsymbol{w}_j$$

ក្នុងករណីការកំណត់ប្រភេទអត្ថបទតាមប្រធានបទ យើងអាចកំណត់ជាទម្រង់ម៉ូឌែលលីនេអ៊ែរខាងលើបានដោយ ជាឧទាហរណ៍កំណត់យក $\hat{y} = 0$ សម្គាល់ប្រធានបទអំពី business, $\hat{y} = 1$ សម្គាល់ប្រធានបទអំពីSocial, $\hat{y} = 2$ សម្គាល់ប្រធានបទអំពីSport ជាដើម។

ដើម្បីងាយស្រួលក្នុងការបកស្រាយខាងក្រោម យើងសន្មតថា អត្ថបទនិមួយៗត្រូវបានបង្ហាញដោយវ៉ិចទ័រ $\boldsymbol{x} \in \mathbb{R}^d$ ដែលកំប៉ូសង់និមួយៗគឺជាប្រេកង់នៃពាក្យក្នុងវចនានុក្រមដែលមាននៅក្នុងអត្ថបទនោះ។ ដូចដែលអ្នកកត់សម្គាល់បាន បញ្ហាសម្រាប់យើងគឺថា តើនឹងត្រូវកំណត់តម្លៃនៃប៉ារ៉ាម៉ែត្ររបស់ម៉ូឌែលដោយរបៀបណា។ ក្នុងអត្ថបទនេះយើងនឹងណែនាំម៉ូឌែលតម្រែតម្រង់Logisticសម្រាប់ចំណាត់ថ្នាក់ច្រើនក្រុម។

2. តម្រែតម្រង់Logisticសម្រាប់ចំណាត់ថ្នាក់ច្រើនក្រុម (Multiclass Logistic Regression)

តម្រែតម្រង់Logisticសម្រាប់ចំណាត់ថ្នាក់ច្រើនក្រុមគឺជាម៉ូឌែលលីនេអ៊ែរមួយប្រភេទ ដែលប្រូបាបមានលក្ខខណ្ឌ$p(\hat{y} = j|\boldsymbol{x})$ពោលគឺប្រូបាបដែលប្រភេទនៃទិន្នន័យ$\hat{y} = j$ ត្រូវបានទស្សន៍ទាយចំពោះអថេរពន្យល់ $\boldsymbol{x}$ ត្រូវបានគណនាដូចខាងក្រោម។

$$P(\hat{y} = j|\boldsymbol{x}) = \frac{\exp(\boldsymbol{x}^\top \boldsymbol{w}_j)}{\sum_{j=0}^{K} \exp(\boldsymbol{x}^\top \boldsymbol{w}_j)}$$

ដើម្បីសម្រួលក្នុងការសរសេរ នៅទីនេះយើងកំណត់សរសេរ $a_j = \boldsymbol{x}^\top \boldsymbol{w}_j$ ។ ចំពោះក្រុម $j \in C = \{0,1,\dots,K\}$ យើងកំណត់សរសេរវ៉ិចទ័រ $\boldsymbol{a}$ ដែលជាតម្លៃផលគុណស្កាលែចំពោះក្រុមទាំងអស់ដោយទម្រង់ខាងក្រោម។

$$\boldsymbol{a} = (a_0 \quad a_1 \quad \cdots \quad a_K) = (\boldsymbol{x}^\top \boldsymbol{w}_0 \quad \boldsymbol{x}^\top \boldsymbol{w}_1 \quad \cdots \quad \boldsymbol{x}^\top \boldsymbol{w}_K)$$

ប្រូបាបមានលក្ខខណ្ឌដែលអត្ថបទ $\boldsymbol{x}$ ត្រូវបានកំណត់ថាក្នុងក្រុមj : $p(\hat{y} = j|\boldsymbol{x})$ អាចបង្ហាញតាមរយៈអនុគមន៍ Softmax បានដូចខាងក្រោមដែល$\boldsymbol{a}_j$សម្គាល់កំប៉ូសង់ទីjនៃវ៉ិចទ័រ$\boldsymbol{a}$។

$$P(\hat{y} = j|\boldsymbol{x}) = \sigma(\boldsymbol{a})_j = \frac{\exp(\boldsymbol{a}_j)}{\sum_{k=0}^{K} \exp(\boldsymbol{a}_k)}$$

ចំពោះប៉ារ៉ាម៉ែត្រនៃក្រុមទាំងអស់ដែលមាន បើយើងកំណត់សរសេរដោយម៉ាទ្រីស W , ប្រូបាបមានលក្ខខណ្ឌចំពោះការធ្វើចំណាត់ថ្នាក់ក្រុមនិមួយដោយπ_j នោះយើងអាចបង្ហាញការគណនាខាងលើបានដូចខាងក្រោម។

$$\boldsymbol{a} = (\boldsymbol{x}^\top \boldsymbol{w}_0 \quad \boldsymbol{x}^\top \boldsymbol{w}_1 \quad \cdots \quad \boldsymbol{x}^\top \boldsymbol{w}_K) = \boldsymbol{x}^\top W$$

$$(\pi_0 \quad \pi_1 \quad \cdots \quad \pi_K) = (\sigma(\boldsymbol{x}^\top \boldsymbol{w}_0)_0 \quad \sigma(\boldsymbol{x}^\top \boldsymbol{w}_1)_1 \quad \cdots \quad \sigma(\boldsymbol{x}^\top \boldsymbol{w}_K)_K)$$

$$\boldsymbol{\pi} = (\pi_0 \quad \pi_1 \quad \cdots \quad \pi_K) = \sigma(\boldsymbol{x}^\top W)$$

ឧទាហរណ៍ថាចំពោះ ទិន្នន័យ$\boldsymbol{x}$ ប្រូបាបមានលក្ខខណ្ឌចំពោះក្រុមនិមួយៗត្រូវបានគណនា និងបានលទ្ធផល $P(\hat{y} = 0|\boldsymbol{x}) = 0.1$, $P(\hat{y} = 1|\boldsymbol{x}) = 0.4$, $P(\hat{y} = 2|\boldsymbol{x}) = 0.2$, $P(\hat{y} = 3|\boldsymbol{x}) = 0.3$ នោះទិន្នន័យត្រូវបានទស្សន៍ទាយថាក្នុងចំណាត់ថ្នាក់ក្រុម$\hat{y} = 1$ ព្រោះប្រូបាប $P(\hat{y} = 1|\boldsymbol{x})$ មានតម្លៃខ្ពស់ជាងគេ។

ដើម្បីងាយស្រួលក្នុងការសរសេរនិងគណនាកម្រិតនៃភាពសាកសមរបស់ទិន្នន័យ ភាគច្រើនចំណាត់ថ្នាក់ក្រុមរបស់ទិន្នន័យត្រូវបានបង្ហាញជាទម្រង់one-hot vector។ឧបមាថាទិន្នន័យ $\boldsymbol{x}$ ស្ថិតនៅក្នុងក្រុម1 នោះវ៉ិចទ័រតាងចំណាត់ថ្នាក់ក្រុមរបស់វាត្រូវបានសរសេរដោយ

$$\boldsymbol{y} = (0 \quad 1 \quad 0 \quad \cdots \quad 0)^\top$$

ពោលគឺទិន្នន័យស្ថិតនៅក្រុមj ត្រូវបានសរសេរដោយវ៉ិចទ័រ $\boldsymbol{y}$ ដែលកំប៉ូសង់ទីjរបស់វាកំណត់ដោយ1 និង0 ចំពោះកំប៉ូសង់ក្រៅពីនេះ។

$$y_k = \begin{cases} 1 & (k = j) \\ 0 & (k \neq j) \end{cases}$$

3. កម្រិតសាកសមនៃទិន្នន័យ Likelihood

គំនិតនៃការបង្ហាញកម្រិតសាកសមរបស់ទិន្នន័យក្នុងម៉ូឌែលតម្រែតម្រង់Logisticសម្រាប់ចំណាត់ថ្នាក់ច្រើនក្រុម គឺដូចគ្នាទៅនឹងករណីចំណាត់ថ្នាក់២ក្រុមដែរ។ ចំពោះទិន្នន័យ$(\boldsymbol{x}, \boldsymbol{y})$ និងប៉ារ៉ាម៉ែត្រនៃម៉ូឌែលW កម្រិតសាកសមនៃទិន្នន័យ(Likelihood) ត្រូវបានកំណត់ដូចខាងក្រោម។

$$\hat{l}_{(\boldsymbol{x},\boldsymbol{y})}(W) = P(\hat{y} = j|\boldsymbol{x}) = \pi_j$$

ដោយប្រើទម្រង់one-hot vectorនៃចំណាត់ថ្នាក់ក្រុមរបស់ទិន្នន័យ $\boldsymbol{y} \in \mathbb{R}^K$ យើងអាចសរសេរដូចខាងក្រោម ។

$$\hat{l}_{(\boldsymbol{x},\boldsymbol{y})}(W) = \pi_j = \prod_{k=0}^{K} \begin{cases} \pi_k & (y_k = 1) \\ 1 & (y_k = 0) \end{cases} = \prod_{k=0}^{K} \pi_k^{y_k}$$

ចំពោះទិន្នន័យTrainingទាំងអស់ $\mathcal{D}$ ដែលមាន Likelihood អាចសរសេរបានក្រោមទម្រង់

$$\hat{L}_{\mathcal{D}}(W) = \prod_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{D}} \hat{l}_{(\boldsymbol{x},\boldsymbol{y})}(W)$$

ដើម្បីបញ្ចៀសនូវបញ្ហាតម្លៃតូចពេកក្នុងការគណនាជាមួយកុំព្យូទ័រ នៅទីនេះយើងសិក្សាបរមាកម្មលើតម្លៃលោការីតរបស់វា ។ ការធ្វើបែបនេះមិនប៉ះពាល់ដល់ការធ្វើបរមាកម្មឡើយ ព្រោះអនុគមន៍លោការីតជាអនុគមន៍កើនដាច់ខាត ។

$$\hat{\mathcal{L}}_{\mathcal{D}}^{MLE}(W) = -\log \hat{L}_{\mathcal{D}}(W) = -\log \prod_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{D}} \hat{l}_{(\boldsymbol{x},\boldsymbol{y})}(W) = -\sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{D}} \hat{l}_{(\boldsymbol{x},\boldsymbol{y})}(W)$$

4. ការប៉ាន់ស្មានមេគុណតម្រែតម្រង់តាម Maximum Likelihood

ចំពោះសំណុំទិន្នន័យទាំងអស់$\mathcal{D}$ដែលមានចំនួនN យើងកំណត់កម្រិតសាកសមនៃទិន្នន័យ(likelihood)ពោលគឺកម្រិតដែលម៉ូឌែលអាចប៉ាន់ស្មានបានត្រឹមត្រូវចំពោះគ្រប់ទិន្នន័យទាំងអស់ដោយកន្សោមខាងក្រោម ។ នៅទីនេះយើងសន្មតថារបាយនៃគ្រប់ទិន្នន័យទាំងអស់គឺឯករាជ្យនិងមានឯកសណ្ឋានភាព(i.i.d : independent and identically dristributed) ។

$$\hat{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{w}) = -\log \hat{L}_{\mathcal{D}}(\boldsymbol{w}) = -\sum_{i=1}^{N} \log \hat{l}_{(\boldsymbol{x}_i,y_i)}(\boldsymbol{w})$$

5. ការដោះស្រាយតាមរយៈវិធី SGD

ដើម្បីធ្វើអប្បបរមាកម្ម $\hat{\mathcal{L}}_{\mathcal{D}}(W)$ យើងនឹងប្រើប្រាស់វិធីSGDដែលបានសិក្សាក្នុងអត្ថបទមុន ។

$$W^{(t+1)} = W^{(t)} - \eta_t \frac{\partial}{\partial W}\{-\log \hat{l}_{(\boldsymbol{x},y)}(W)\}\Big|_{W=W^{(t)}}$$

$$W^{(t+1)} = W^{(t)} + \eta_t \frac{\partial}{\partial W}\log \hat{l}_{(\boldsymbol{x},y)}(W)\Big|_{W=W^{(t)}}$$

ជាដំបូងយើងពិនិត្យលើអនុគមន៍ដេរីវេ $\frac{\partial}{\partial W}\log \hat{l}_{(\boldsymbol{x},y)}(W)$ ។

$$\log \hat{l}_{(\boldsymbol{x},y)}(W) = \log\left(\prod_{k=0}^{K} \pi_k^{y_k}\right) = \sum_{k=0}^{K} y_k \log \pi_k$$

$$\frac{\partial}{\partial \boldsymbol{w}_j}\log \hat{l}_{(\boldsymbol{x},y)}(W) = \frac{\partial}{\partial \boldsymbol{w}_j}\sum_{k=0}^{K} y_k \log \pi_k = \sum_{k=0}^{K} \frac{y_k}{\pi_k}\frac{\partial \pi_k}{\partial \boldsymbol{w_j}}$$

នៅទីនេះយើងពិនិត្យលើអនុគមន៍បណ្តាក់ $\pi_k = \sigma(\boldsymbol{a})_k$, $a_j = \boldsymbol{x}^{\top}\boldsymbol{w}_j$

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{w}_j}\log \hat{l}_{(\boldsymbol{x},y)}(W) &= \sum_{k=0}^{K} \frac{y_k}{\pi_k}\frac{\partial \pi_k}{\partial \boldsymbol{w}_j} = \sum_{k=0}^{K} \frac{y_k}{\pi_k}\frac{\partial \pi_k}{\partial a_j}\frac{\partial a_j}{\partial \boldsymbol{w}_j} \\
&= \sum_{k=0}^{K} \frac{y_k}{\pi_k}\{\pi_k(\delta_{kj} - \pi_j)\}\boldsymbol{x} \quad , \quad \delta_{kj} = \begin{cases} 1 \ (k = j) \\ 0 \ (k \neq j) \end{cases} \\
&= \boldsymbol{x}\sum_{k=0}^{K} y_k(\delta_{kj} - \pi_j) \\
&= \boldsymbol{x}\left(\sum_{k=0}^{K} y_k\delta_{kj} - \sum_{k=0}^{K} y_k\pi_j\right) = \boldsymbol{x}(y_j - \pi_j)
\end{aligned}$$

$$\frac{\partial}{\partial \boldsymbol{w}_j}\log \hat{l}_{(\boldsymbol{x},y)}(W) = \boldsymbol{x}(y_j - \pi_j)$$

ដូចនេះ តាមរយៈវិធីSGD តម្លៃនៃប៉ារ៉ាម៉ែត្រ $\boldsymbol{w}_j$ នៃ Wដែលធ្វើបរមាកម្មលើតម្លៃកម្រិតសាកសមនៃទិន្នន័យត្រូវបានគណនាដោយផ្លាស់ប្តូរតម្លៃដូចខាងក្រោម។

$$\boldsymbol{w}_j^{(t+1)} = \boldsymbol{w}_j^{(t)} + \eta_t \frac{\partial}{\partial \boldsymbol{w}_j} \log \hat{l}_{(\boldsymbol{x},y)}(W)\Big|_{\boldsymbol{w}_j=\boldsymbol{w}_j^{(t)}}$$

$$\boldsymbol{w}_j^{(t+1)} = \boldsymbol{w}_j^{(t)} + \eta_t\left(y_j - \pi_j^{(t)}\right)\boldsymbol{x}$$

នៅទីនេះ $\pi^{(t)}$ ជាតម្លៃប្រូបាបដែលគណនាដោយម៉ូឌែលជាមួយតម្លៃប៉ារ៉ាម៉ែត្រនៅដំណាក់កាលt នៃការផ្លាស់ប្តូរតម្លៃប៉ារ៉ាម៉ែត្រ ពោលគឺ $\pi^{(t)} = \sigma\left(\boldsymbol{x}^\top \boldsymbol{w}^{(t)}\right)$។ η_t ជាlearning-rate និង y ជាចំណាត់ថ្នាក់ក្រុមពិតប្រាកដនៃទិន្នន័យ $\boldsymbol{x}$។

ក្នុងករណីប្រើRidge Regularization កន្សោមខាងលើនឹងប្រែទៅជាទម្រង់ខាងក្រោម។

$$\boldsymbol{w}^{(t+1)} = \left(1 - \frac{2\alpha\eta_t}{N}\right)\boldsymbol{w}^{(t)} + \eta_t\left(y - \pi^{(t)}\right)\boldsymbol{x}$$

6. ការវាយតម្លៃ

កាលពីសិក្សាបញ្ហាចំណាត់ថ្នាក់២ក្រុម រង្វាស់សម្រាប់វាយតម្លៃលើម៉ូឌែលត្រូវបានគណនាដោយប្រើ Accuracy, Precision, Recall, F1-score។ក្នុងករណីចំណាត់ថ្នាក់ច្រើនក្រុម តម្លៃទាំងនេះក៏ត្រូវបានប្រើដើម្បីវាយតម្លៃម៉ូឌែលដូចគ្នា។ ចំពោះ Accuracy អាចគណនាបានដោយធ្វើផលធៀបចំនួនករណីដែលម៉ូឌែលធ្វើការប៉ាន់ស្មានបានត្រឹមត្រូវ ធៀបនឹងចំនួនទិន្នន័យសរុប។ ចំពោះ Precision, Recall, F1-score នៃក្រុមនិមួយៗត្រូវបានគណនាដោយឡែកៗពីគ្នា រួចធ្វើតម្លៃមធ្យមMacro /Micro ដូចខាងក្រោម។ ជាមួយគ្នានេះការសិក្សាលើវាយតម្លៃដោយប្រើ Confusion Matrix ដូចក្នុងរូបទី២ ក៏ត្រូវបានប្រើផងដែរ។

$$\text{Macro Precision} = \frac{\text{Precision}_0 + \text{Precision}_1 + \cdots + \text{Precision}_\text{K}}{K}$$

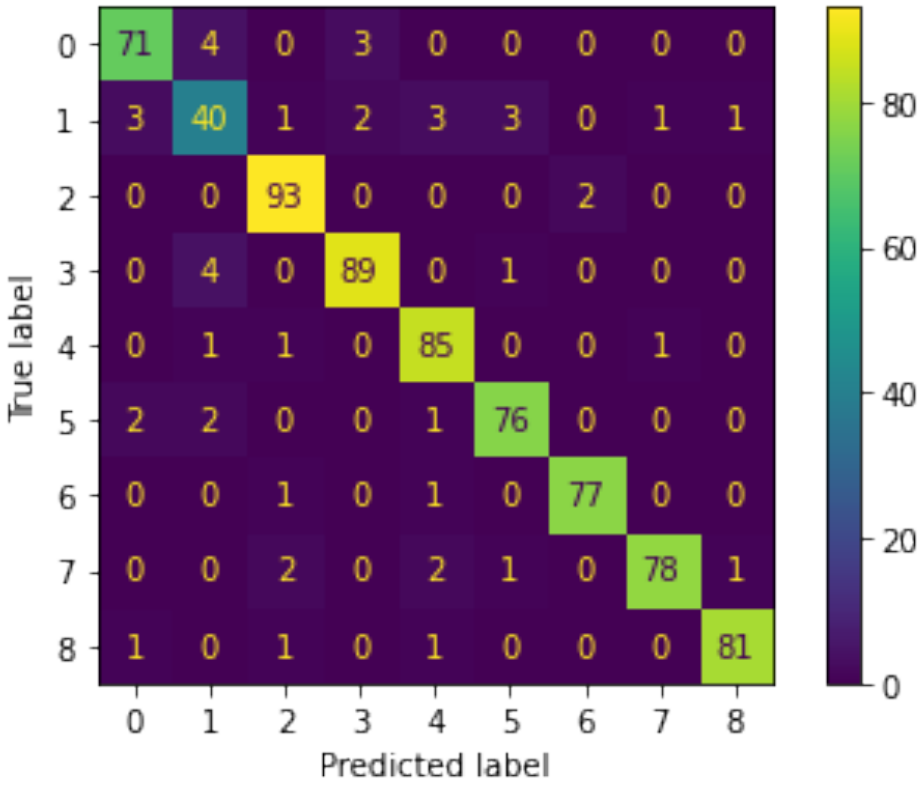
$$\text{Macro Recall} = \frac{\text{Recall}_0 + \text{Recall}_1 + \cdots + \text{Recall}_\text{K}}{K}$$

$$\text{Macro F1} - \text{score} = \frac{\text{F1}_0 + \text{F1}_1 + \cdots + \text{F1}_\text{K}}{K}$$

$$\text{Micro Precision} = \frac{\text{PA}_0 + \text{PA}_1 + \cdots + \text{PA}_\text{K}}{\text{PB}_0 + \text{PB}_1 + \cdots + \text{PB}_\text{K}}, \qquad \text{Precision}_k = \frac{\text{PA}_k}{\text{PB}_k}$$

$$\text{Micro Recall} = \frac{\text{RA}_0 + \text{RA}_1 + \cdots + \text{RA}_\text{K}}{\text{RB}_0 + \text{RB}_1 + \cdots + \text{RB}_\text{K}}, \qquad \text{Recall}_k = \frac{\text{RA}_k}{\text{RB}_k}$$

$$\text{Micro F1} - \text{score} = \frac{\text{FA}_0 + \text{FA}_1 + \cdots + \text{FA}_\text{K}}{\text{FB}_0 + \text{FB}_1 + \cdots + \text{FB}_\text{K}}, \qquad \text{F1} - \text{score}_k = \frac{\text{FA}_k}{\text{FB}_k}$$

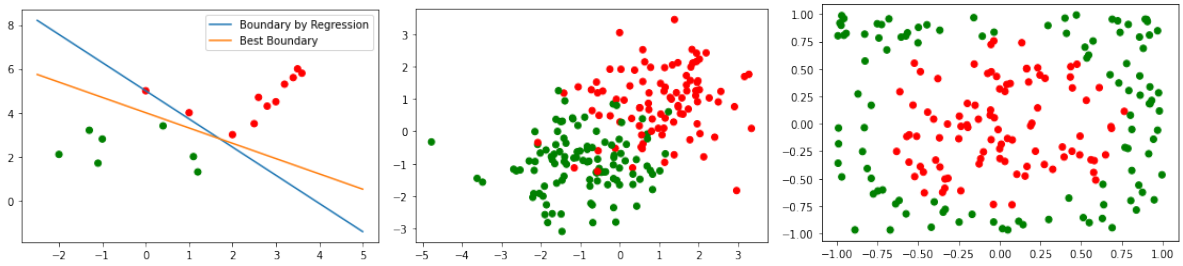


រូបទី២ Confusion Matrix

Support Vector Machine

ក្នុងអត្ថបទមុន យើងបានណែនាំអំពីការធ្វើចំណាត់ថ្នាក់ទិន្នន័យ២ឬច្រើនក្រុមដោយប្រើម៉ូឌែលតម្រែតម្រង់(Linear regression, Logistic regression) ។ ប៉ុន្តែមានចំណុចខ្សោយសំខាន់ពីរកើតមានឡើងក្នុងការប្រើប្រាស់ម៉ូឌែលតម្រែតម្រង់
ក្នុងការធ្វើចំណាត់ថ្នាក់ក្រុមទិន្នន័យ ដូចបង្ហាញក្នុងរូបទី១ខាងក្រោម ។
ក្នុងរូបខាងឆ្វេង ដោយសារមានទិន្នន័យប្រមូលផ្តុំច្រើននៅផ្នែកខាងក្រុមក្រហម ម៉ូឌែលតម្រែតម្រង់លីនេអ៊ែរនឹងផ្តល់ឱ្យនូវបន្ទាត់ព្រំដែនដែលខិតទៅជិតក្រុមក្រហមខ្លាំង (បន្ទាត់ខៀវ) ។ ប៉ុន្តែតាមពិតបន្ទាត់ព្រំដែនទឹកក្រូចអាចមើលឃើញថាល្អប្រសើរជាង ។ ក្នុងករណីរូបកណ្តាលនិងរូបខាងស្តាំ ច្បាស់ណាស់ថា ទិន្នន័យទាំងនេះមិនអាចធ្វើការបែងចែកដោយបន្ទាត់ត្រង់បានឡើយ ពោលគឺជាប្រភេទដែលមិនអាចបែងចែកលីនេអ៊ែរ(linear non-seperable data) ។

ដើម្បីដោះស្រាយបញ្ហាទាំងនេះ ការធ្វើចំណាត់ថ្នាក់ក្រុមទិន្នន័យដោយប្រើSupport Vector Machineត្រូវបានប្រើ ។ ក្នុងអត្ថបទនេះយើងនឹងណែនាំអំពីដំណើរការគណិតវិទ្យាក្នុងការដោះស្រាយបញ្ហាធ្វើចំណាត់ថ្នាក់ក្រុម២ឬច្រើនក្រុមដោយប្រើSupport Vector Machine ។ យើងនឹងពិនិត្យទាំងករណីទិន្នន័យដែលអាចធ្វើបំណែងចែកលីនេអ៊ែរបាន(Linear Seperable) និងមិនបាន(Linear Non- Seperable) ។



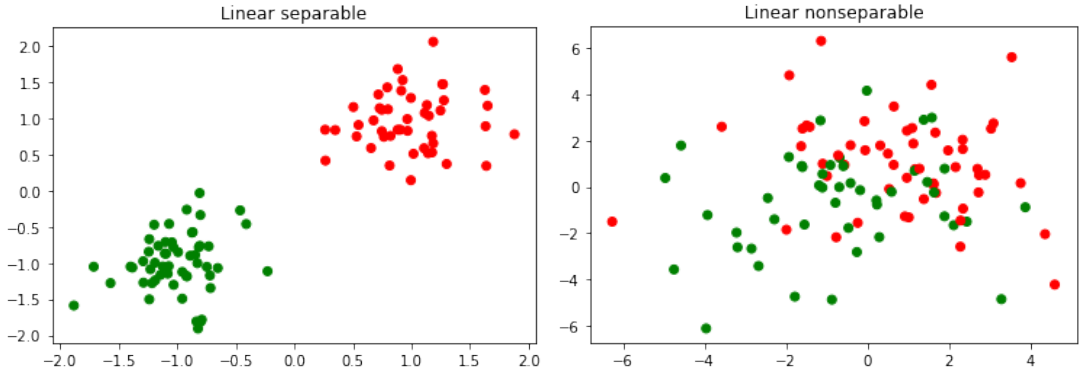
រូបទី១ ករណីធ្វើចំណាត់ថ្នាក់មិនបានល្អជាមួយLinear Regression

1. ចំណាត់ថ្នាក់២ក្រុម

នៅទីនេះយើងសិក្សាលើទិន្នន័យពីរក្រុម($+1, -1$) $\mathcal{D} = \{(\boldsymbol{x_i}, y_i)\}_{i=1}^{N}$ ដែល $\boldsymbol{x_i} \in \mathbb{R}^d$, $y_i \in \{+1, -1\}$ ។ យើងចង់បង្កើតម៉ូឌែលចំណាត់ថ្នាក់ក្រុមមួយដែលធ្វើការបែងចែកតាមម៉ូឌែលលីនេអ៊ែរដែលតាងដោយទម្រង់ខាងក្រោម ។

$$\mathcal{M} = \{\text{sign}(f(\boldsymbol{x})) \mid f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b \,, \boldsymbol{w} \in \mathbb{R}^d, b \in \mathbb{R}\}$$

ក្នុងករណីដែលមាន $m(\boldsymbol{x}) = \text{sign}(\boldsymbol{w}^\top \boldsymbol{x} + b)$ ដែលផ្ទៀងផ្ទាត់ $m(\boldsymbol{x}_i) = y_i \ \forall\, i = 1,2,\dots,N$ នោះ ទិន្នន័យដែលមានហៅថា អាចធ្វើបំណែងចែកលីនេអ៊ែរបាន(Linear Seperable)។ ពោលគឺ យើងអាចកំណត់បន្ទាត់ឬប្លង់ព្រំដែនដើម្បីបែងចែកទិន្នន័យទាំងពីរប្រភេទបានច្បាស់លាស់ដោយគ្មានកំហុសចំពោះគ្រប់ទិន្នន័យ។ ផ្ទុយទៅវិញ ករណីដែលមិនអាចរកបានម៉ូឌែលណាដែលផ្ទៀងផ្ទាត់លក្ខខណ្ឌខាងលើ យើងហៅថា មិនអាចធ្វើបំណែងចែកលីនេអ៊ែរបាន(Linear Non-Seperable)។



រូបទី២ ទិន្នន័យដែលអាចបែងចែកលីនេអ៊ែរបាននិងទិន្នន័យដែលមិនអាចបែងចែកលីនេអ៊ែរបាន

1.1. ករណីទិន្នន័យដែលអាចធ្វើបំណែងចែកលីនេអ៊ែរបាន(Linear Seperable)

ក្នុងករណីទិន្នន័យដែលអាចធ្វើបំណែងចែកលីនេអ៊ែរបាន ចំពោះទិន្នន័យ $\mathcal{D} = \{(\boldsymbol{x_i}, y_i)\}_{i=1}^N$ យើងអាចសរសេរទំនាក់ទំនងខាងក្រោមបាន

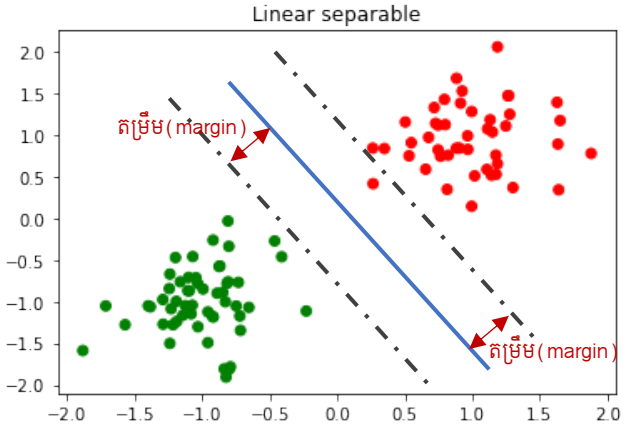
$$y_i = +1 \implies \boldsymbol{w}^\top \boldsymbol{x}_i + b > 0$$
$$y_i = -1 \implies \boldsymbol{w}^\top \boldsymbol{x}_i + b < 0$$

ឬ ជារួម

$$y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) > 0 \ \ (i = 1,2,\dots,N)$$

ជាទូទៅ តម្លៃនៃប៉ារ៉ាម៉ែត្រ $(\boldsymbol{w}, b)$ ដែលផ្ទៀងផ្ទាត់លក្ខខណ្ឌខាងលើមានច្រើនលើសពីមួយ។ ក្នុងចំណោមនោះ ដើម្បីជ្រើសបានប៉ារ៉ាម៉ែត្រដែលប្រសើរ គោលគំនិតក្នុង Support Vector Machine គឺកំណត់យកករណីដែលតម្រឹម(margin)នៃទិន្នន័យនិងព្រំដែនមានទំហំធំបំផុត។

ការកំណត់តម្រឹម(margin)ដែលធំបំផុតនៅទីនេះ គឺសំដៅដល់ការយកប៉ារ៉ាម៉ែត្រណាដែលធ្វើឱ្យចន្លោះរវាងទិន្នន័យទាំងពីរក្រុមនិងបន្ទាត់(ឬប្លង់)ព្រំដែនមានគម្លាតឆ្ងាយពីគ្នាបំផុត(រូបទី៣)។



រូបទី៣ ការកំណត់តម្រឹម(margin)អតិបរមា

យើងនឹងបង្ហាញបញ្ហាខាងលើជាទម្រង់គណិតវិទ្យា។

សន្មតបន្ទាត់ឬប្លង់ព្រំដែនមានទម្រង់ដូចម៉ូឌែលខាងលើ $\boldsymbol{w}^\top \boldsymbol{x} + b = 0$។ ចម្ងាយពីចំណុច $\boldsymbol{x}_0 \in \mathbb{R}^d$ ទៅព្រំដែនអាចកំណត់បានដូចខាងក្រោម។

$$\frac{|\boldsymbol{w}^\top \boldsymbol{x}_0 + b|}{\|\boldsymbol{w}\|}$$

ហេតុនេះ ការកំណត់ប៉ារ៉ាម៉ែត្រ $(\boldsymbol{w}, b)$ ដែលធ្វើឱ្យតម្រឹមអតិបរមាអាចកំណត់បានជាចំណោទបរមាដូចខាងក្រោម។

$$\max_{\boldsymbol{w}\in\mathbb{R}^d, b\in\mathbb{R}} \min_{i=1,2,\dots,N} \frac{|\boldsymbol{w}^\top \boldsymbol{x}_i + b|}{\|\boldsymbol{w}\|}$$

ក្រោមលក្ខខណ្ឌ

$$y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) > 0 \ \ (i = 1,2, \dots, N)$$

ចំណោទបរមាខាងលើអាចសរសេរជាទម្រង់សមមូលបែបងាយដូចខាងក្រោមបានដោយសារតែអនុគមន៍គោលដៅដែលត្រូវធ្វើបរមាកម្ម
មិនប្រែប្រួលតម្លៃឡើយពេល $\boldsymbol{w}, b$ ត្រូវបានគុណនឹងចំនួនថេរក៏ដោយ។

$$\min_{\boldsymbol{w}\in\mathbb{R}^d, b\in\mathbb{R}} \frac{\|\boldsymbol{w}\|^2}{2}$$

ក្រោមលក្ខខណ្ឌ

$$y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) > 0 \ \ (i = 1,2, \dots, N)$$

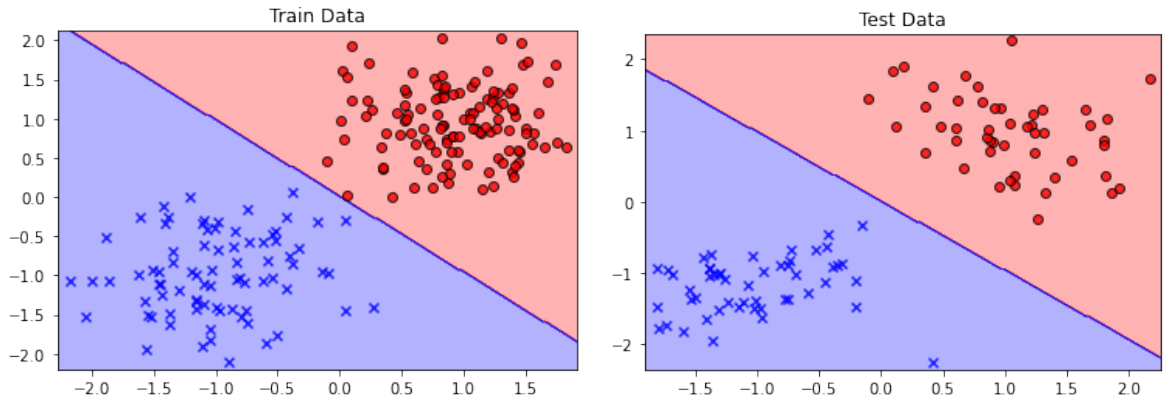
អនុគមន៍គោលដៅខាងលើមានទម្រង់ជាអនុគមន៍ប៉ោងដឺក្រេទី២ ហើយលក្ខខណ្ឌរបស់វាដែលត្រូវផ្ទៀងផ្ទាត់មានទម្រង់

ជាលីនេអ៊ែរ។ ចំណោទបរមាបែបនេះហៅថាចំណោទប្រូក្រាមលំដាប់២(QP:Quadratic Programming)។ ក្នុងការដោះស្រាយចំណោទបែបនេះមានalgorithmដែលមានប្រសិទ្ធិភាពខ្ពស់ជាច្រើនត្រូវបានស្រាវជ្រាវ។នៅទីនេះយើងនឹងមិនធ្វើការបកស្រាយលំអិតឡើយ។ សន្មតថាចម្លើយនៃចំណោទបរមាខាងលើគឺ $\hat{\boldsymbol{w}} \in \mathbb{R}^d, \hat{b} \in \mathbb{R}$ នោះការធ្វើចំណាត់ថ្នាក់ក្រុមទិន្នន័យអាចធ្វើបានដោយគណនាតាមទម្រង់ខាងក្រោម។

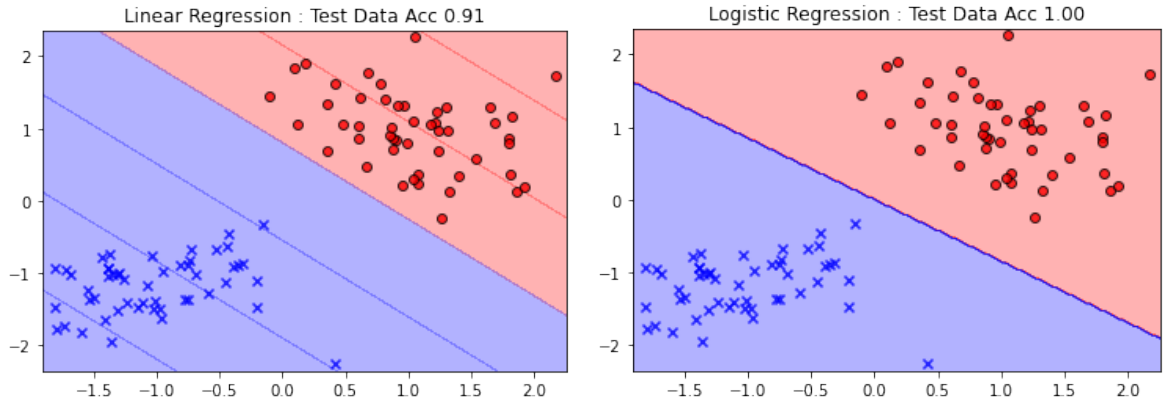
$$\hat{m}(\boldsymbol{x}) = \text{sign}\left(\hat{\boldsymbol{w}}^\top \boldsymbol{x} + \hat{b}\right)$$

ជាមួយPython អ្នកអាចប្រើ sklearn.svm packageបានប្រើSupport Vector Machine Model។

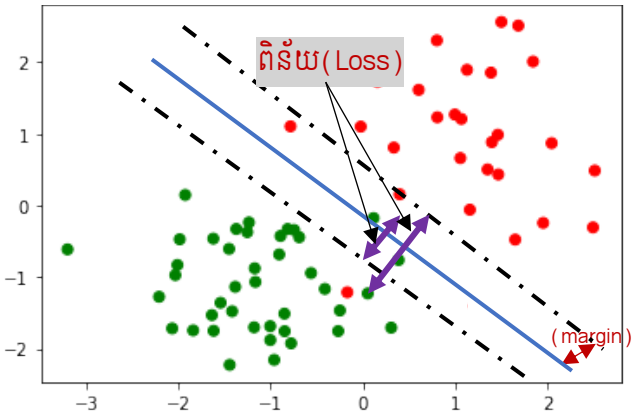
```
from sklearn.svm import SVC,LinearSVC
sv_model = SVC(kernel="linear", C=1.0, random_state=1)
sv_model.fit(Xtrain,ytrain)
```



រូបទី៤ ការធ្វើចំណាត់ថ្នាក់ក្រុម(កំណត់ព្រំដែនក្រុម)ដោយSupport Vector Machineលើ Linear Seperable Data



រូបទី៥ ការធ្វើចំណាត់ថ្នាក់ក្រុម(កំណត់ព្រំដែនក្រុម)ដោយ Regression Model លើ Linear Seperable Data



រូបទី៦ បន្ទាត់ព្រំដែន និង កម្រិតពិន័យ(loss) ក្នុង Soft Support Vector Machine

1.2. ករណីទិន្នន័យដែលអាចធ្វើបំណែងចែកលីនេអ៊ែរបាន(Linear Seperable)

ក្នុងករណីដែលទិន្នន័យមិនអាចធ្វើបំណែងចែកលីនេអ៊ែរបាន ចំណោទបរមាខាងលើគ្មានតម្លៃប៉ារ៉ាម៉ែត្រណាដែលផ្ទៀងផ្ទាត់លក្ខខណ្ឌឡើយ។ ហេតុនេះដើម្បីដោះស្រាយបញ្ហាបំណែងចែកបែបនេះករណីកំហុស(មិនបំពេញលក្ខខណ្ឌ $y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) > 0 \ \ (i = 1,2, \dots, N)$)ក្នុងបំណែងចែកខ្លះត្រូវបានលើកលែង។ដំណោះស្រាយបែបនេះហៅថា Soft Support Vector Machine។

នៅទីនេះដោយប្រើម៉ូឌែលបំណែងចែក $m(\boldsymbol{x}) = \text{sign}(\boldsymbol{w}^\top \boldsymbol{x} + b)$ ករណីទិន្នន័យ$(\boldsymbol{x}_i, y_i)$ តម្លៃនៃការពិន័យ(Loss)លើកំហុសនៃចំណាត់ថ្នាក់ក្រុមត្រូវបានកំណត់ដូចខាងក្រោម។

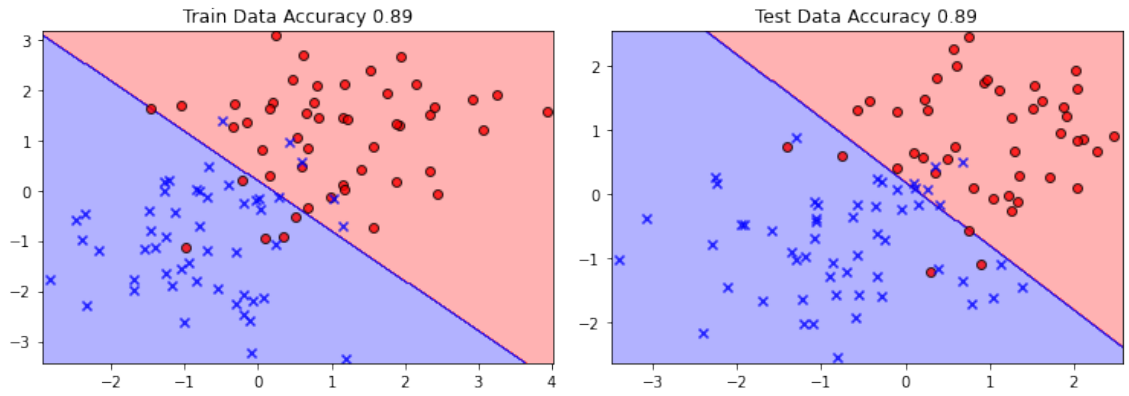
(១) ចំពោះ $y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1$ យើងសន្មតថាម៉ូឌែលអាចបែងចែកបានល្អ ដោយកំណត់តម្លៃពិន័យ 0

(២) ចំពោះ $y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) < 1$ យើងសន្មតថាម៉ូឌែលអាចបែងចែកមិនបានល្អដោយកំណត់តម្លៃនៃពិន័យ $1 - y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) > 0$

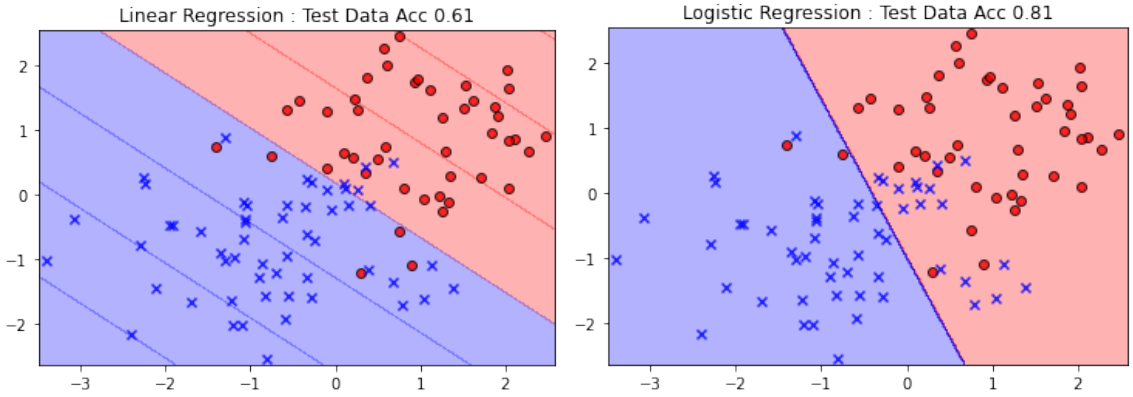
ក្នុងករណីនេះដើម្បីកំណត់ប៉ារ៉ាម៉ែត្រដែលប្រសើរសម្រាប់ធ្វើចំណាត់ថ្នាក់ក្រុមជាមួយSoft Support Vector Machine យើងនឹងកំណត់តម្លៃប៉ារ៉ាម៉ែត្រណាដែលធ្វើឱ្យតម្រឹមធំបំផុត តែកម្រិតពិន័យ (Loss)តូចបំផុត។ យើងអាចបង្ហាញបញ្ហានេះជាទម្រង់ចំណោទបរមាបានដូចខាងក្រោម។

$$\min_{\boldsymbol{w} \in \mathbb{R}^d, b \in \mathbb{R}} \frac{C}{N} \sum_{i=1}^{N} \max\{1 - y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b), 0\} + \frac{1}{2}\|\boldsymbol{w}\|^2$$

C ជាតម្លៃសម្រាប់កំណត់កម្រិតនៃការផ្តោតលើបរមាកម្មរវាងតម្រឹម(margin)និងពិន័យ(loss)។ នៅទីនេះដូចដែលអ្នកអាចចាប់អារម្មណ៍បាន ការកំណត់តម្លៃ C ជារឿងពិបាក។ វិធីដែលច្រើនអនុវត្តគឺការធ្វើ Cross-validation ។ យើងនឹងមិនលំអិតលើវិធីសាស្ត្រនេះឡើយនៅទីនេះ។



រូបទី៧ ការធ្វើចំណាត់ថ្នាក់ក្រុម(កំណត់ព្រំដែនក្រុម)ដោយSupport Vector Machineលើ Linear Non-Seperable Data



រូបទី៨ ការធ្វើចំណាត់ថ្នាក់ក្រុម(កំណត់ព្រំដែនក្រុម)ដោយRegression Model លើ Linear Non-Seperable Data

តាមលទ្ធផលក្នុងរូបខាងលើ យើងអាចផ្ទៀងផ្ទាត់ពីការធ្វើបំណែងចែកក្រុមបានល្អប្រសើរក្នុងករណីប្រើSoft Support Vector Machine ប្រើប្រៀបធៀបទៅនឹងម៉ូឌែលតម្រែតម្រង់ដែលបានសិក្សាកន្លងមក។ប៉ុន្តែទោះជាយ៉ាងណា ករណីទិន្នន័យដែលមិនអាចធ្វើបំណែងចែកលីនេអ៊ែរបាន កម្រិតត្រឹមត្រូវនៃការធ្វើចំណាត់ថ្នាក់ក្រុមនៅមានកម្រិតទោះបីជាប្រើSoft Support Vector Machineក្តី។

2. Kernel Support Vector Machine

ក្នុងករណីដែលទិន្នន័យមិនអាចបែងចែកលីនេអ៊ែរបាន ការប្រើKernelជាជំនួយត្រូវបានអនុវត្តជាទូទៅ។ ជាមួយការប្រើប្រាស់Kernel សមត្ថភាពនៃការបង្ហាញលក្ខណៈពិសេសរបស់ទិន្នន័យនឹងត្រូវបានបង្កើន។ កាលពីសិក្សាពីម៉ូឌែលតម្រែតម្រង់ យើងបានលើកឡើងអំពីការប្រើអនុគមន៍គោលដែលជាអនុគមន៍មិនលីនេអ៊ែរដើម្បីបង្ហាញលក្ខណៈពិសេសរបស់ទិន្នន័យមួយចំនួន។

ស្រដៀងគ្នានេះជាមួយគោលគំនិតក្នុងការប្រើKernel ឧបមាថាយើងមានអនុគមន៍គោល $\varphi_1(\boldsymbol{x}), \varphi_2(\boldsymbol{x}), \ldots, \varphi_D(\boldsymbol{x})$ នោះអនុគមន៍Kernel ត្រូវបានកំណត់ដូចខាងក្រោម ដែលនៅទីនេះ $\Phi(\boldsymbol{x}) = (\varphi_1(\boldsymbol{x}) \quad \cdots \quad \varphi_D(\boldsymbol{x}))^\top$ ។

$$k(\boldsymbol{x}, \boldsymbol{x}') = \sum_{d=1}^{D} \varphi_d(\boldsymbol{x})\varphi_d(\boldsymbol{x}') = \Phi(\boldsymbol{x})^\top \Phi(\boldsymbol{x}')$$

ក្នុងករណីនេះ ទម្រង់នៃ ម៉ូឌែលចំណាត់ថ្នាក់ក្រុមនិងបន្ទាត់ឬប្លង់ព្រំដែនអាចបង្ហាញដូចទម្រង់ខាងក្រោម ដោយសន្មតយក $\boldsymbol{\beta} = (\beta_1 \quad \cdots \quad \beta_n)^\top$ជំនួស $\boldsymbol{w}$ ។

$$f(\boldsymbol{x}) = \boldsymbol{w}^\top \Phi(\boldsymbol{x}) + b$$

$$f(\boldsymbol{x}) = \sum_{i=1}^{N} \beta_i k(\boldsymbol{x}, \boldsymbol{x}_i) + b$$

ដូចគ្នានឹងករណីទូទៅនៃSupport Vector Machine ដែរ ក្នុងករណី Kernel Support Vector Machine បញ្ហាខាងលើអាចបង្ហាញជាទម្រង់ចំណោទបរមាដូចខាងក្រោម។

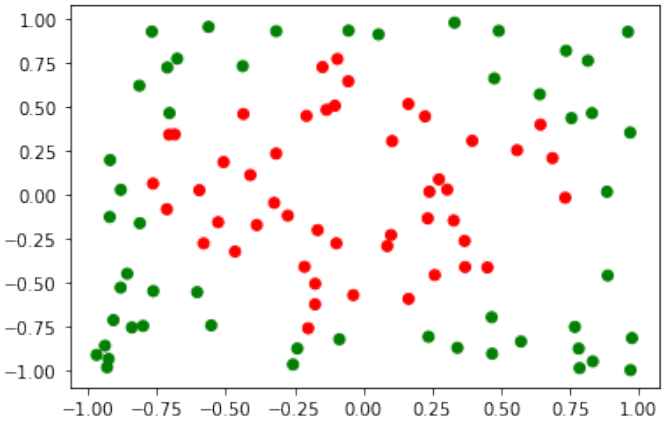
$$\min_{\boldsymbol{\beta}, b} \frac{C}{N} \sum_{i=1}^{N} \max\{1 - y_i f(\boldsymbol{x}_i)\ , 0\} + \frac{1}{2}\boldsymbol{\beta}^\top K \boldsymbol{\beta}$$

ក្រោមលក្ខខណ្ឌ

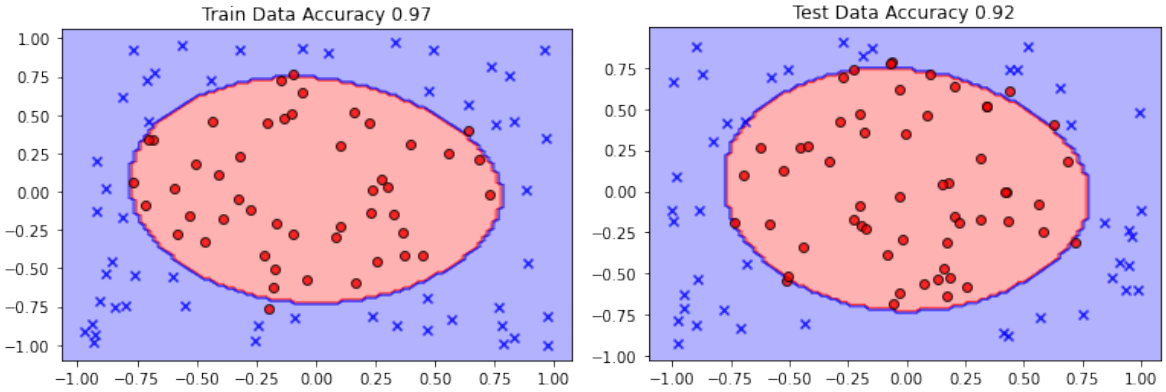
$$f(\boldsymbol{x}_i) = \sum_{j=1}^{N} \beta_i K_{ij} + b \ \ (i = 1,2, \ldots, N)$$

ដែល $K_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ ។

ជាមួយPython អ្នកអាចជ្រើសរើសប្រភេទ Kernel ដែលនិយមប្រើដូចជា: លីនេអ៊ែរ linear , ពហុធា poly , Gaussian Kernel: rbf , Sigmoid Kernel: sigmoid ។



រូបទី៨ ទិន្នន័យដែលមិនអាចបែងចែកលីនេអ៊ែរបាន



រូបទី៩ ការធ្វើចំណាត់ថ្នាក់ក្រុម(កំណត់ព្រំដែនក្រុម)ដោយKernel Support Vector Machine (kernel=rbf) លើ Linear Non-Seperable Data

```
from sklearn.svm import SVC

sv_model = SVC(kernel="rbf", C=1.0, random_state=1)
sv_model.fit(Xtrain,ytrain)
```

```
print("Train score:",sv_model.score(Xtrain,ytrain))
print("Test score:",sv_model.score(Xtest,ytest))
```

```
Train score: 0.97
Test score: 0.92
```

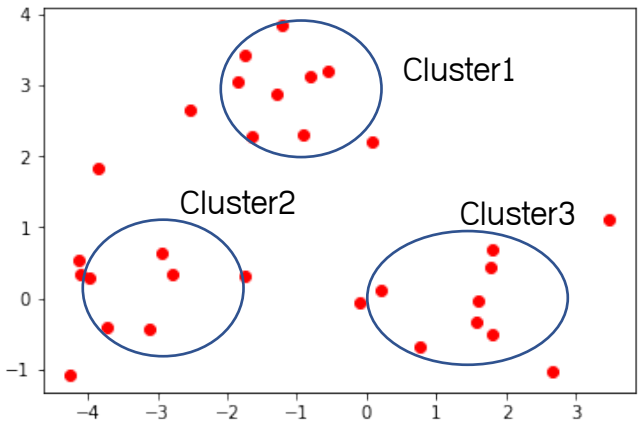
Clustering

ក្នុងបញ្ហានៃការធ្វើចំណាត់ថ្នាក់ក្រុមទិន្នន័យដែលយើងបានណែនាំក្នុងអត្ថបទមុនៗ ទិន្នន័យនិមួយៗជាគូ$(\boldsymbol{x}, y)$ពោលគឺយើងប្រើទិន្នន័យដែលមានចម្លើយជាមុនដើម្បីបង្រៀនដល់ម៉ូឌែលរបស់យើងដែលហៅថា Supervised Learning។ ផ្ទុយពីនេះ ប្រភេទបញ្ហាក្នុង Unsupervised Learning យើងមានតែទិន្នន័យ $\boldsymbol{x}$ តែប៉ុណ្ណោះ។ ក្នុងករណីនេះការប្រមូលផ្តុំទិន្នន័យដែលមានលក្ខណៈដូច ឬស្រដៀងគ្នាជាក្រុមឬជាចង្កោមដោយផ្អែកលើលក្ខណៈបង្ហាញដោយផ្ទាល់ឬប្រយោលរបស់ទិន្នន័យត្រូវបានហៅថាជាClustering។ ក្នុងអត្ថបទនេះយើងនឹងណែនាំអំពីវិធីសាស្ត្រក្នុងការដោះស្រាយបញ្ហាបែបនេះ។

1. ចង្កោម(Cluster)ទិន្នន័យនិងចម្ងាយ

ការធ្វើចំណាត់ក្រុមទិន្នន័យដែលគ្មានសញ្ញាសម្គាល់ប្រភេទជាក្រុមដូចក្នុងរូបទី១គឺជាគោលដៅចំបងនៃការសិក្សាក្នុងUnsupervised Learning។ ក្រុមដែលប្រមូលបានក្នុងទិន្នន័យនោះហៅថាចង្កោមទិន្នន័យ(Cluster) ហើយដំណើរការនៃការប្រមូលជាក្រុមបែបនេះហៅថា Clustering ។

ដើម្បីធ្វើចំណាត់ក្រុមទិន្នន័យបែបនេះ គោលគំនិតសំខាន់គឺការស្វែងរកលក្ខណៈរួមឬប្រហាក់ប្រហែលរវាងទិន្នន័យ។ ពោលគឺ ទិន្នន័យដែលមានលក្ខណៈស្រដៀងគ្នាលើរង្វាស់ណាមួយអាចប្រមូលផ្តុំជាចង្កោមបាន។ ហេតុនេះការកំណត់និយមន័យជាក់លាក់នៃលក្ខណៈរួមឬប្រហាក់ប្រហែលនេះជាដំណើរការសំខាន់ក្នុងការចាប់ផ្តើម។ នៅទីនេះយើងកំណត់យកចម្ងាយជារង្វាស់សម្រាប់បង្ហាញភាពស្រដៀងគ្នានៃទិន្នន័យ។ ដូច្នេះយើងនឹងធ្វើការពិនិត្យលើនិយមន័យនៃចម្ងាយដូចខាងក្រោម។



រូបទី១ ទិន្នន័យនិងចង្កោមដែលអាចកំណត់បាន

1.1. ចម្ងាយ Euclid

សន្មតថា $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ ជាពីរចំណុចក្នុងលំហទិន្នន័យ។

បើ $\boldsymbol{x} = (x_1 \quad \cdots \quad x_d)^\top, \boldsymbol{y} = (y_1 \quad \cdots \quad y_d)^\top$ នោះ ចម្ងាយEuclid រវាងចំណុចទាំងពីរនេះអាចកំណត់បានដូចខាងក្រោម។

$$D_{Euclid}(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_2 = \left(\sum_{i=1}^{d} (x_i - y_i)^2\right)^{\frac{1}{2}}$$

1.2. ចម្ងាយ Manhattan

សន្មតថា $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ ជាពីរចំណុចក្នុងលំហទិន្នន័យ។

បើ $\boldsymbol{x} = (x_1 \quad \cdots \quad x_d)^\top, \boldsymbol{y} = (y_1 \quad \cdots \quad y_d)^\top$ នោះ ចម្ងាយ Manhattan រវាងចំណុចទាំងពីរនេះអាចកំណត់បានដូចខាងក្រោម។

$$D_{Manhattan}(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_1 = \sum_{i=1}^{d} |x_i - y_i|$$

1.3. កាពស្រដៀងគ្នាកូស៊ីនុស

សន្មតថា $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ ជាពីរចំណុចក្នុងលំហទិន្នន័យ។ ដោយប្រើមុំផ្គុំដោយវិចទ័រចំណុចទាំងពីរយើងអាចកំណត់ថាវិចទ័រចំណុចទាំងពីរមានទិសដៅដូចគ្នានៅពេលរង្វាស់មុំនោះកាន់តែតូច។ ហេតុនេះការប្រើតម្លៃកូស៊ីនុសនៃមុំផ្គុំដោយវិចទ័រចំណុចទាំងពីរអាចប្រើជារង្វាស់ប្រៀបធៀបលក្ខណៈស្រដៀងគ្នានៃទិន្នន័យបាន។ បើ $(\boldsymbol{x}, \boldsymbol{y})$ ជាផលគុណស្កាលែរនៃវិចទ័រទាំងពីរ នោះ កាពស្រដៀងគ្នាកូស៊ីនុសរវាងចំណុចទាំងពីរនេះអាចកំណត់បានដូចខាងក្រោម។

$$\cos(\boldsymbol{x}, \boldsymbol{y}) = \frac{(\boldsymbol{x}, \boldsymbol{y})}{\|\boldsymbol{x}\|_2 \cdot \|\boldsymbol{y}\|_2}$$

1.4. មេគុណ Jaccard

ក្នុងករណីដែលទិន្នន័យបង្ហាញជាទម្រង់សំណុំ ពោលគឺ $\boldsymbol{x} = \{x_1, \ldots, x_d\}, \boldsymbol{y} = \{y_1, \ldots, y_d\}$ នោះកម្រិតស្រដៀងគ្នានៃទិន្នន័យទាំងពីរអាចកំណត់បានដោយមេគុណ Jaccard ដូចខាងក្រោម ដែល$|\boldsymbol{a}|$ជាកាឌីណាល់(ចំនួនធាតុ)នៃសំណុំ $\boldsymbol{a}$ ។

$$\text{Jaccard}(\boldsymbol{x}, \boldsymbol{y}) = \frac{|\boldsymbol{x} \cap \boldsymbol{y}|}{|\boldsymbol{x} \cup \boldsymbol{y}|}$$

1.5. KL divergence

ក្នុងករណីដែលទិន្នន័យបង្ហាញជាទម្រង់អនុគមន៍បំណែងចែកប្រូបាប $p(x), p(y)$ នោះកម្រិតស្រដៀងគ្នានៃទិន្នន័យទាំងពីរអាចកំណត់បានដូចខាងក្រោមដែល ។

$$\mathrm{KL}(p||\, q) = \int p(x) \log \frac{p(x)}{q(x)} dx$$

កម្រិតស្រដៀង KL មិនមានលក្ខណៈឆ្លុះឡើយ ពោលគឺ $\mathrm{KL}(p||\, q) \neq \mathrm{KL}(q||\, p)$ ។ ហេតុនេះក្នុងករណីខ្លះJensen-Shannon Divergence ត្រូវបានប្រើជំនួស ។

$$D_{\mathrm{js}} = \frac{1}{2}\big(\mathrm{KL}(p||q) + \mathrm{KL}(q||p)\big)$$

2. វិធីសាស្ត្រ K-means

សន្មតថាយើងមានចំណុចទិន្នន័យ $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N \in \mathbb{R}^d$ ។ យើងចង់ធ្វើចំណាត់ថ្នាក់ក្រុមដោយស្វ័យប្រវត្តិលើចំណុចទិន្នន័យទាំងនេះជា K ក្រុម $C_1, \dots, C_K$ ។ ដូចដែលបង្ហាញខាងលើទិន្នន័យនិមួយៗមិនមានភ្ជាប់ជាមួយនូវកំណត់សម្គាល់(label)អំពីក្រុមដែលខ្លួនស្ថិតនៅឡើយ ។ ហេតុនេះ ការធ្វើចំណាត់ក្រុមត្រូវពិនិត្យលើកម្រិតស្រដៀងគ្នានៃទិន្នន័យដោយផ្អែកលើចម្ងាយរវាងគ្នា ។

ដំបូង យើងកំណត់ហៅចំណុចតំណាងនៃក្រុមដោយ $\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K \in \mathbb{R}^d$ ។ ចំណុចតំណាងទាំងនេះត្រូវបានហៅថា Centroids ។ ចំណុចទិន្នន័យដែលនៅជិតចំណុចតំណាង$\boldsymbol{\mu}_k$ នឹងត្រូវចាត់ចូលជាសមាជិកនៃក្រុមC_k ។

បើចម្ងាយរវាងពីរចំណុចទិន្នន័យកំណត់ដោយ $d(\boldsymbol{x}_1, \boldsymbol{x}_2)$ នោះកម្រិតគម្លាតសរុបរវាងក្រុមនិមួយៗត្រូវបានកំណត់ដូចខាងក្រោម ។ ការធ្វើចំណាត់ថ្នាក់ក្រុមដែលល្អត្រូវមានតម្លៃនៃកម្រិតគម្លាតសរុបរវាងក្រុមតូចបំផុត ។ ការធ្វើចំណាត់ថ្នាក់ក្រុមបែបនេះហៅថា វិធីសាស្ត្រ K-means ។

$$\sum_{k=1}^{K} \sum_{\boldsymbol{x} \in C_k} d(\boldsymbol{x}, \boldsymbol{\mu}_k)^2$$

ក្នុងការបកស្រាយខាងក្រោម យើងកំណត់ប្រើចម្ងាយ Euclid ។ តាង $|C_k|$ ជាចំនួនចំនុចទិន្នន័យក្នុងក្រុម C_k នោះវិចទ័រមធ្យមនៃចំណុចទិន្នន័យក្នុងក្រុមនេះកំណត់ដោយ $\bar{\boldsymbol{x}}_k$ ។

$$\bar{\boldsymbol{x}}_k = \frac{1}{|C_k|} \sum_{\boldsymbol{x} \in C_k} \boldsymbol{x}$$

ចំពោះក្រុម $C_1, \dots, C_K$ យើងបានទំនាក់ទំនងខាងក្រោម ។ ហេតុនេះ បើយើងកំណត់យកចំណុចតំណាងនៃក្រុមនិមួយ $\boldsymbol{\mu}_k$ ដោយវិចទ័រមធ្យមនៃចំណុចទិន្នន័យក្នុងក្រុម នោះយើងនឹងបានតម្លៃនៃកម្រិតគម្លាតសរុបរវាងក្រុមតូចបំផុត ។

$$\sum_{x\in C_k} \|\boldsymbol{x}-\boldsymbol{\mu}_k\|_2{}^2 \geq \sum_{x\in C_k} \|\boldsymbol{x}-\bar{\boldsymbol{x}}_k\|_2{}^2 \quad (k=1,\ldots,K)$$

ដោយផ្អែកលើទំនាក់ទំនងនេះ ដំណើរការនៃវិធីសាស្ត្រK-meansចំពោះចម្ងាយEuclidអាចសរុបដូចខាងក្រោម ។

Input : ចំណុចទិន្នន័យ $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N \in \mathbb{R}^d$, ចំនួនក្រុម K

Initialization : កំណត់តម្លៃចាប់ផ្តើមនៃចំណុចតំណាង $\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_K \in \mathbb{R}^d$

Step-1 : អនុវត្តជំហាន(1),(2),(3)ខាងក្រោមដដែលៗ

(1) ផ្លាស់ប្តូរសមាជិកក្រុម $C_1,\ldots,C_K$

(សន្មតថាទិន្នន័យនិមួយៗត្រូវស្ថិតនៅក្នុងក្រុមណាមួយក្នុងចំណោមនេះ)

$$C_k = \{\boldsymbol{x}_n |\ \|\boldsymbol{x}_n-\boldsymbol{\mu}_k\|_2 \leq \|\boldsymbol{x}_n-\boldsymbol{\mu}_{k'}\|_2\ , k'\neq k\}$$

(2) ផ្លាស់ប្តូរតម្លៃនៃចំណុចតំណាង $\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_K$

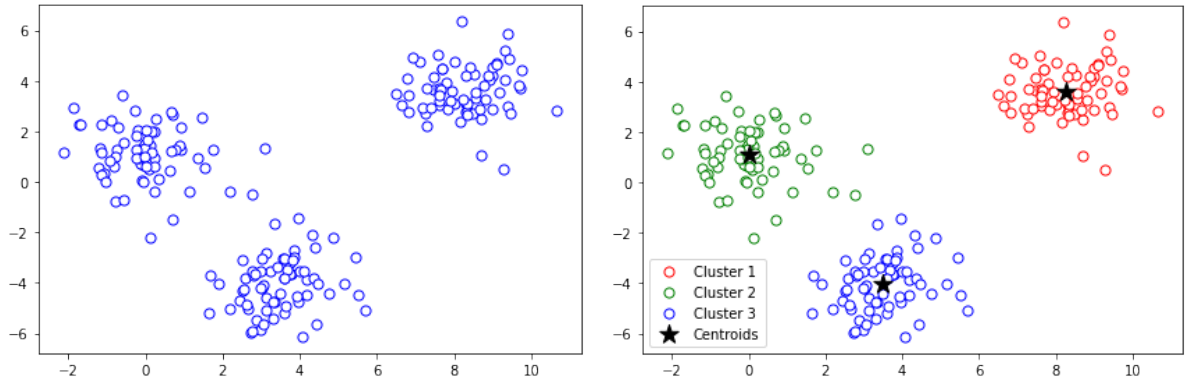
$$\boldsymbol{\mu}_k = \frac{1}{|C_k|}\sum_{x\in C_k} \boldsymbol{x} \quad (k=1,\ldots,K)$$

(3) បន្តទៅStep-2នៅពេលដែលតម្លៃនៃកម្រិតគម្លាតសរុបរវាងក្រុមរួម(ស្ទើឬក្បែរសូន្យ)

Step-2 : យកការធ្វើចំណាត់ថ្នាក់ក្រុម $C_1,\ldots,C_K$ ជាចម្លើយ

ជាមួយPython អ្នកអាចប្រើ sklearn.cluster បាន ។

```
from sklearn.cluster import KMeans
KM = KMeans(n_clusters=3, init='random', n_init=10, max_iter=500, tol=1e-6)
y_KM = KM.fit_predict(X)
```



រូបទី២ ចំណុចទិន្នន័យមុនធ្វើចំណាត់ថ្នាក់ក្រុម និងក្រោយធ្វើចំណាត់ក្រុមដោយK-means

ក្នុងការអនុវត្តវិធីសាស្ត្រK-means ការកំណត់តម្លៃដំបូងនៃចំណុចតំណាងមានឥទ្ធិពលខ្លាំងលើលទ្ធផលនៃការធ្វើចំណាត់ក្រុម។ជាទូទៅការកំណត់តម្លៃដំបូងនេះធ្វើឡើយដោយតម្លៃចៃដន្យ។

ក្នុងករណីsklearn.cluster.KMeans យើងអាចកំណត់ដោយ `init='random'` ។ ប៉ុន្តែករណីខ្លះការកំនត់ដោយចៃដន្យនេះអាចនឹងបានលទ្ធផលចំណាត់ក្រុមដែលមិនល្អ។

ដើម្បីដោះស្រាយបញ្ហានេះ ការប្រើ K-means++ អាចសម្រួលបាន។ ក្នុងK-means++ វិធីសាស្ត្រK-meansត្រូវបានអនុវត្តដោយចៃដន្យជាច្រើនលើកទៅលើទិន្នន័យដែលមាន រួចតម្លៃមធ្យមនៃកម្រិតគម្លាតសរុបនឹងត្រូវធ្វើអប្បបរមាកម្ម។

ករណីsklearn.cluster.KMeans យើងអាចកំណត់ដោយ `init='k-means++'` ។

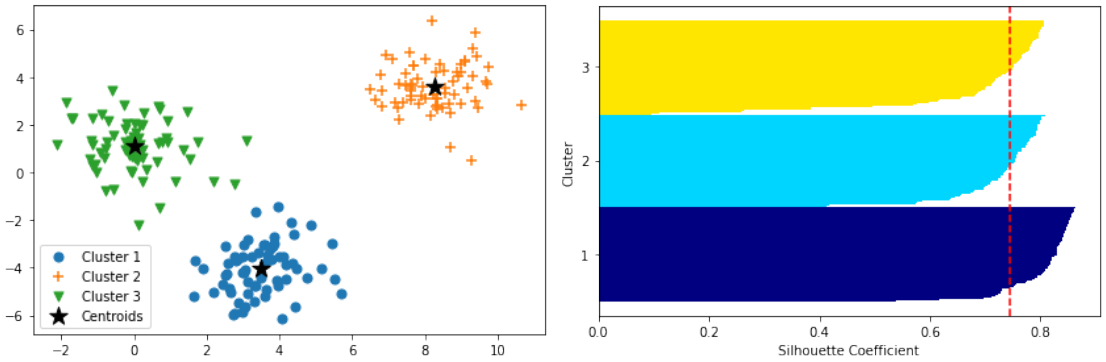
3. ការវាយតម្លៃClusteringដោយប្រើមេគុណSilhouette

ក្រៅពីការសិក្សាលើតម្លៃនៃកម្រិតគម្លាតសរុបរវាងក្រុមនិមួយៗ ការវិភាគលើកម្រិតកំហាប់នៃទិន្នន័យក្នុងក្រុម(ភាពជិតស្និតក្នុងក្រុម)ដូចជា Silhouetter Analysis ក៏ត្រូវបានប្រើប្រាស់សម្រាប់វាយតម្លៃលើ Clustering ផងដែរ។ ក្នុង Silhouetter Analysis កម្រិតកំហាប់នៃការប្រមូលផ្តុំរបស់ទិន្នន័យក្នុងក្រុមនិមួយៗត្រូវបានគណនាដោយមេគុណsilhouetterនិងបង្ហាញជាក្រាប។ មេគុណ silhouetter នៃទិន្នន័យ x_i : $s^{(i)}$អាចគណនាបានតាម៣ជំហានខាងក្រោម។

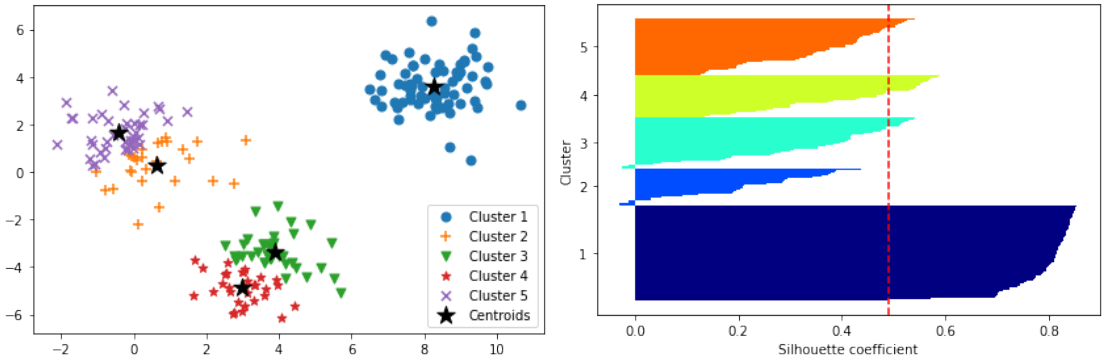
(1) កំណត់កម្រិតកំហាប់ប្រមូលផ្តុំនៃក្រុម $a^{(i)}$ ដោយតម្លៃមធ្យមនៃចម្ងាយរវាងចំណុចទិន្នន័យ x_i និងចំណុចទិន្នន័យដទៃទៀតក្នុងក្រុមជាមួយគ្នា។

(2) កំណត់កម្រិតគម្លាតរវាងក្រុមជិតបំផុត $b^{(i)}$ ដោយតម្លៃមធ្យមនៃចម្ងាយរវាងចំណុចទិន្នន័យ x_i
និងចំណុចទិន្នន័យទាំងអស់នៅក្នុងក្រុមដែលជិតនឹងក្រុមរបស់ខ្លួនបំផុត។

(3) កំណត់តម្លៃមេគុណ silhouetterដោយ $s^{(i)} = \left(b^{(i)} - a^{(i)}\right) / \max\{a^{(i)}, b^{(i)}\}$



រូបទី៣ មេគុណsilhouetterតាមក្រុមនិមួយៗ(ចំណាត់ថ្នាក់ជា៣ក្រុម)



រូបទី៤ មេគុណsilhouetterតាមក្រុមនិមួយៗ(ចំណាត់ថ្នាក់ជា៣ក្រុម)

ក្នុងរូបទី៣និងទី៤ ខាងលើបន្ទាត់ក្រហមបង្ហាញតម្លៃមធ្យមនៃមេគុណsilhouetter លើក្រុមនិមួយៗក្នុងករណីចំនួនក្រុមត្រូវបានកំណត់ជា៣និង៥។ Clustering ដែលល្អនឹងមានតម្លៃនៃមេគុណsilhouetterខិតទៅជិត1។ ក្នុងរូបខាងលើ យើងអាចយើញថាករណីClustering ដោយ ៣ក្រុមមានមេគុណsilhouetterប្រសើរជាងបើប្រៀបធៀបនឹងករណី៥ក្រុម។

4. Gaussian Mixture Models Clustering

ក្នុងវិធីសាស្ត្រ K-means យើងមិនបានធ្វើកំណត់លក្ខខណ្ឌសន្មតណាមួយលើរបាយបំណែងចែកនៃទិន្នន័យឡើយ។ ពេលនេះយើងសន្មតថាទិន្នន័យទាំងអស់គឺស្ថិតក្នុងរបាយបំណែងចែកប្រូបាបមួយ។ ក្នុងករណីនេះយើងកំណត់យកបំណែងចែកនរ៉ាល់ និងម៉ូឌែលបន្សំនៃនរ៉ាល់ ពោលគឺ Gaussian Mixture Models ដើម្បីដោះស្រាយបញ្ហាClustering។

សន្មតថាក្រុមនិមួយត្រូវបានកំណត់ដោយបំណែងចែកប្រូបាបQហើយទិន្នន័យក្នុងក្រុមនិមួយៗត្រូវបានកំណត់ដូចខាងក្រោម។

$$C_k \sim Q \ , \ \boldsymbol{x} \sim p_k(\boldsymbol{x})$$

ពីទិន្នន័យយើងនឹងធ្វើការកំណត់បំណែងចែក Q, p_k ដែលនឹងនាំឱ្យយើងអាចកំណត់ចំណាត់ថ្នាក់ក្រុមសម្រាប់ទិន្នន័យនិមួយៗបាន។ ក្នុងម៉ូឌែលបន្សំ Mixture Models, ចំពោះបំណែងចែកពហុធាQ និង បំណែងចែកនរម៉ាល់ $p_k(\boldsymbol{x}): N_d(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ បើយកប្រូបាបដែលទិន្នន័យជាសមាជិកក្រុមC_kដោយq_kនោះ ម៉ូឌែលស្ថិតិនៃទិន្នន័យ$\boldsymbol{x}$ អាចកំណត់ដោយទម្រង់ខាងក្រោម។

$$\sum_{k=1}^{K} q_k p_k(\boldsymbol{x})$$

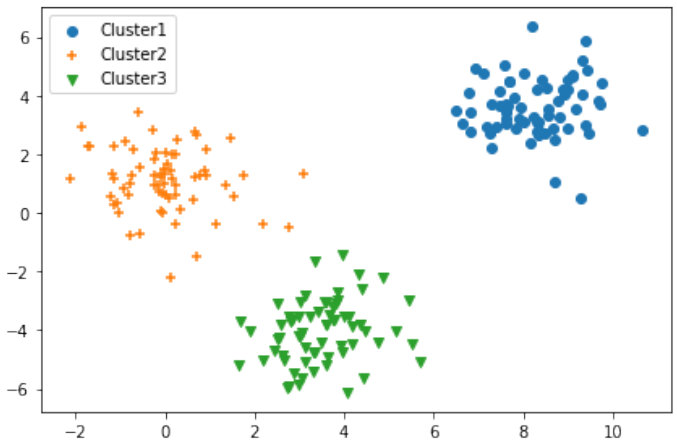
ដើម្បីកំណត់ប៉ារ៉ាម៉ែត្រនៃម៉ូឌែលនេះយើងអាចសិក្សាលើអនុគមន៍លោការីតនៃកម្រិតសាកសមនៃទិន្នន័យ(log-likelihood function)បាន ពោលគឺ

$$\sum_{n=1}^{N} \log\left(\sum_{k=1}^{K} q_k p_k(\boldsymbol{x}_n)\right)$$

ក្នុងការដោះស្រាយបញ្ហាអតិបរមាកម្មនៃអនុគមន៍ខាងលើនេះ EM algorithm ត្រូវបានប្រើ។ នៅទីនេះយើងនឹងមិនលំអិតអំពីEM algorithmឡើយ ប៉ុន្តែយើងណែនាំអំពីការប្រើប្រាស់វាក្នុងការធ្វើ Clustering។

ជាមួយPython អ្នកអាចប្រើ sklearn.mixture.GaussianMixture បាន។

```
from sklearn.mixture import GaussianMixture
model = GaussianMixture(3).fit(X)
classes = model.predict(X)
```



រូបទី៥ លទ្ធផលធ្វើClusteringដោយGaussian Mixture Models

ឯកសារពិគ្រោះ

Takafumi Kanamori, Python で学ぶ統計的機械学習, 2018

Sadanori Konishi, 多変量解析入門—線形から非線形へ, 2010

Hiroshi Nakagawa, Machine Learning, 2015

Sebastian Raschka, Vahid Mirjalili, Python Machine Learning, 2018