

DATA ANALYTIC

CINNEE_PCY



LEC 1 : Supervised Learning and Bayesian Learning (การเรียนรู้แบบมีผู้สอน)

Machine Learning

↪ Unsupervised (การเรียนรู้แบบไม่มีผู้สอน k-Mean, Clustering)

↪ Reinforcement (การเรียนรู้แบบปรับผิด-ถูก)

↪ Supervised * การเรียนรู้แบบมีผู้สอน

* Naive Bayes

* K-NN (K-Nearest Neighbor)

* Classification กรณีที่ต้องเลือกchoice ex hot or cool, yes or no

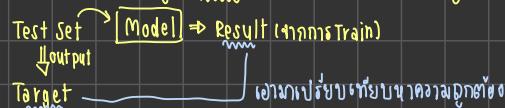
* Linear Regression → กรณีที่ต้อง scale ตามหน่วย ex. อุณหภูมิ, Temperature

Supervised คือ การเรียนรู้จากตัวอย่าง + ประสบการณ์ = งานที่ประดิษฐ์

Task → Experience → Efficiency

จากกิจกรรม Train Set (Data Set) $\Rightarrow "x"$ } หาที่สุด Model

โจทย์เพื่อบอกค่าตอบ Target / Label / GT (Ground Truth) $\Rightarrow "y"$ } Machine learning



งานที่เน้นไป Supervised • งานที่ต้องคำนึงถึงความถูกต้อง

ex Machine Learning • ใช้รูปแบบเดียวกันที่เราเรียนรู้ไปways

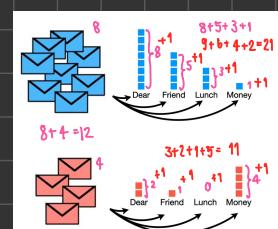
Image จัดงานไปพลาง • ค่าการตัดสินใจที่เราให้ไว้

กรณีเป็นนัก/แมง • ฟังก์ชันและอัลกอริทึม input, output

• input ต้องเข้ารูปแบบ

• Brute-force งานที่ไม่สามารถแก้ไขทุกกรณี

Multinomial (Naive Bayes)



$$\text{normal} \rightarrow \frac{8}{17}, \frac{5}{17}, \frac{3}{17}, \frac{1}{17}$$

$$= 0.47, 0.29, 0.18, 0.06$$

$$\text{spam} \rightarrow \frac{2}{7}, \frac{1}{7}, \frac{0}{7}, \frac{4}{7}$$

$$= 0.29, 0.14, 0.00, 0.57$$

"Dear Friend" friend? or spam?

Prior Probability (Initial Guess) ให้ normal $\frac{8}{12} = 0.67 \times (0.47) = 0.09$

spam $\frac{4}{12} = 0.33 \times (0.29) = 0.01$

$\therefore p(\text{normal}) > p(\text{spam}) ; \text{email-normal}$

"Lunch Money Money Money Money"

$$\text{normal} = (0.67)(\frac{4}{17})(\frac{2}{17})^4 = 0.00001$$

$$\text{spam} = (0.33)(\frac{1}{7})(\frac{5}{21})^4 = 0.00122 \quad \therefore p(\text{spam}) > p(\text{normal}) ; \text{email-spam}$$

\therefore Naive Bayes นิยามว่า order "Bias ต่ำ" > กรณีที่มีปัจจัยหลายต่อ 1 Good

Gaussian Naive Bayes

Bernoulli Naive Bayes

Categorical Naive Bayes

Complement Naive Bayes

Out-of-core Naive Bayes

Naive Bayes for classification

• ต้องมี Data Set, Label (Ground truth) และ Label ที่เรา choice

• Naive Bayes คือการตัดสินใจแบบมีเงื่อนไข (Given)

• ไม่ควรเกิด over fit $\overline{\text{Overfit}}$ จะเกิด Underfit

• อยู่ในช่วงของสมการ (สูตร)

NOTE: แหล่งศึกษา Naive Bayes ได้คือวิธีการเบื้องต้นที่เรียกว่า

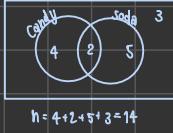
Bayesian Learning → Conditional Prob คือการคำนวณความน่าจะเป็น Given

Bayes' Theorem

Naive Bayes

↳ multinomial, gaussian, Bernoulli, Categorical, Complement, Out-of-core

ex. Class



Contingency Table

	W/Candy	XXxCandy	Total
VSoda	$P = \frac{2}{14}$	$P = \frac{5}{14}$	$\frac{7}{14}$
XXSoda	$P = \frac{4}{14}$	$P = \frac{3}{14}$	$\frac{7}{14}$
Column	$\frac{6}{14}$	$\frac{8}{14}$	
Total	$\frac{6}{14}$	$\frac{8}{14}$	

Like candy and soda from like Candy

$$\frac{2}{14} / \frac{6}{14} = \frac{2}{14} \cdot \frac{14}{6} = \frac{1}{3}$$

Like candy and soda from like Soda

$$\frac{2}{14} / \frac{7}{14} = \frac{2}{7}$$

$$\text{Naive Bayes} \quad P(A|B) = \frac{P(A)}{P(B)} \quad P(A \text{ and } B | B) = \frac{P(A \text{ and } B)}{P(B)}$$

ex. ณ) $P(\text{no c \& s} | \text{no c})$

$$\textcircled{1} \quad \text{if } P(\text{no c \& s} | \text{s}) = 0.71$$

$$P(\text{no c \& s}) = 0.71$$

$$P(\text{no c \& s}) = \frac{0.421}{0.57} = 0.747$$

$$\textcircled{2} \quad P(s) = 0.6$$

$$P(s) = 0.6$$

$$\textcircled{3} \quad P(\text{no c}) = 0.57$$

$$P(\text{no c \& s}) = 0.426$$

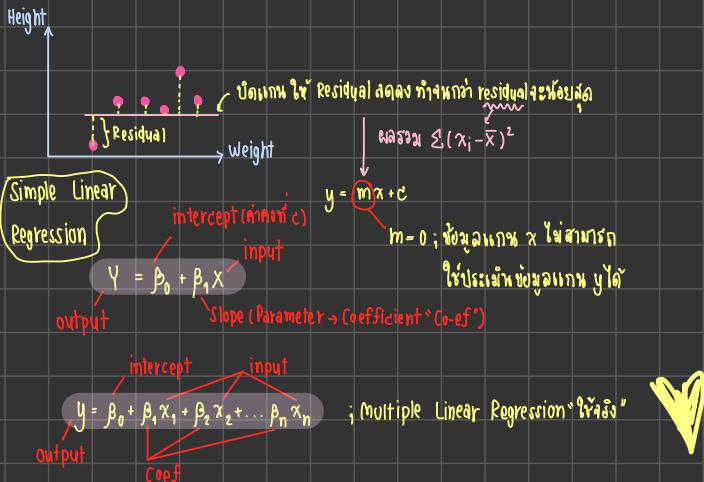
$$\therefore P(\text{no c \& s} | \text{no c}) \approx 0.75$$

LEC 04 Regression (การ预测监督 Supervise)

ການສັກເກຣະຢູ່ປະຈາກສື່ຜົນຮ່ວມງ່າງ Attribute ລາຍການ 1 ທີ່

Linear Regression continuous ຕ້ອງໃຊ້ R

↳ plot graph ແລ້ວຫຼັງໄສການເສັ້ນທຽບ/ໄດ້ເຄີຍ



Linear Regression : R-Square * ນັກສຶກສິ່ນມູນເຕີ 2 ຕອງ Residual = 0 ເສັ້ນ $R^2 = 100\%$ ແລ້ວ

① ພາ mean \rightarrow Least of Squared

② ຮວມຕໍ່ SS(mean) : Sum of Squared Residual // $\sum (x_i - \bar{x})^2$

$$\begin{aligned} SS(\text{mean}) &= (\text{data} - \text{mean})^2 \\ Var(\text{mean}) &= \frac{SS(\text{mean})}{n} \\ &= \frac{1}{n} \sum (x_i - \bar{x})^2 \end{aligned}$$

③ ຮວມຕໍ່ SS(fit) fit ເພື່ອໃຫ້ກຳທານ

$$\begin{aligned} SS(\text{fit}) &= (\text{data} - \text{line})^2 \\ Var(\text{fit}) &= \frac{SS(\text{fit})}{n} \end{aligned}$$

④ ພາ R-Square (S^2)

$$R^2 = \frac{Var(\text{mean}) - Var(\text{fit})}{Var(\text{mean})}$$



ex $Var(\text{mean}) = 11.1$ $Var(\text{fit}) = 4.4$

$$R^2 = \frac{11.1 - 4.4}{11.1}$$

$$R^2 = 0.6 \rightarrow 60\%$$

ຖາກສະກັກ Multiple Linear Regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots$$

ທີ່ co-ef ກ່ອງປົງ 0 ສອກຄົວ / ນີ້ຕ້ອງກ່າວໃຈກ່າວ

ຫຼືຕ້ອງກ່າວຄົວ Attribute ລົງລົງສຸດຂອງ

↳ ລົງ noise, ວິວ, ລົດການໂປ່ກ RAM, ຜົນການ Program ອີງ

Logistic Regression Classification + Linear Regression

↳ ສ່ວນ choice

• Linear Regression ອົດຕະວັດ Height ຢັ້ງ weight ຍັກ

• Logistic Regression ອົດຕະວັດ-ໄວ້ຫຼັງຈາກ weight ຍັກ

↓ Maximum likelihood

ສ້າງ curve line → like? miss?

ເລື່ອງ curve ມີປົວປັງ ກຳຊ້າ

LEC 05 การประยุกต์ประสาทศาสตร์ Model

กระบวนการ Machine Learning

Dataset Preparation: กรณีที่มีข้อมูล → จัดชุด dataset > Model (Kaggle.com)

Design Model

Model ที่ใช้ ex. Neural Network ~ AI, ML

Training The Model, Validation set, Test set

Data Set, Label

Test The Model

Train, Test

Evaluation: วัดประสิทธิภาพของ model

Train, Test

Model = Output from Data set

from Label

ผลลัพธ์ที่บ่งชี้

Validation set หรือ training Train set
วนลูป check "ACC" ต่อเนื่องๆ

วนลูป check "ACC" ต่อเนื่องๆ

k-fold Cross validation กรณีที่มีข้อมูลอย่างจำกัด ให้ unfair ACC หรือผิดๆ (split)

กรณี Train มากกว่า Test ด้วย Acc จะมากกว่าจริง (ผลลัพธ์)] ACC unfair
กรณี Train น้อยกว่า Test ด้วย Acc จะมากกว่าจริง (ผลลัพธ์)]

Accuracy, Confusion Matrix, Precision, Recall, F1-Score, Loss

Accuracy = $\frac{TP + TN}{TP + FP + TN + FN}$

(prediction vs coding) (Actual vs Predict)

	positive	Model	✓	Label	✓	(ถูก)
(TP)	True	Model	✓	Label	✓	(ถูก)
(TN)	True	negative	X	Label	X	(ผิด)
(FP)	False	positive	X	Label	X	(ผิด)
(FN)	False	negative	X	Label	✓	(ถูก)

LOSS function: Model ต่างจาก GT มากแค่ไหน?

* ห้ามต่ำกว่า 0

T (mean square error loss) $MSE = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2$ ยิ่งใหญ่เท่าใด LOSS ก็จะเพิ่ม

RMSE = \sqrt{MSE}

F (mean absolute error loss) $MAE = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i|$

Loss ที่นิยม: Huber Loss (smooth mean absolute error)

→ Log-Cosh Loss

→ Quantile Loss

Confusion → กรณีที่มีผลลัพธ์มากกว่า 2 จำพวก คำนวณจาก confusion matrix Precision / Recall → "True Positive Rate / sensitivity"

Actual			
		(1)	(0)
		+	-
predict	(1)	TP	FP
	(0)	FN	TN

"Positive Predictive Value"

= PPV =

Precision = $\frac{tp}{tp+fp}$ false positive
(Model ✓ Truth ✗)

Detection rate (Medical "ค่าทางการแพทย์")

= TPR =

Recall = $\frac{tp}{tp+fn}$ false negative
(Model ✗ Truth ✓)

$$\text{Precision \& Recall} \rightarrow F1\text{-Score} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = \frac{2}{\frac{\text{Recall} + \text{Precision}}{\text{Precision} \cdot \text{Recall}}} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Recall} + \text{Precision}}$$

Multi-Class

Actually		
+	-	-
+ (2)	0	0
- 2	5	0
- 1	1	2

Precision = $\frac{2}{2+0} = 1$

Recall = $\frac{2}{2+3} = 0.4$

F1-Score = $\frac{(2)(1)(0.4)}{1+0.4} = 0.571$

$2+5+2$

$1+2+2$

Precision = $\frac{9}{9+4} = 0.692$

Recall = $\frac{9}{9+4} = 0.692$

F1-Score = $\frac{9}{9+5+9} = 0.692$

-	+	-
- (2)	0	0
- 2	5	0
- 1	1	2

Precision = $\frac{5}{5+2} = 0.714$

Recall = $\frac{5}{6} = 0.83$

F1-Score = $\frac{(2)(0.714)(0.83)}{0.714+0.83} = 0.768$

$2+5+2$

$3+1+0$

Precision = $\frac{9}{9+4} = 0.692$

Recall = $\frac{9}{9+4} = 0.692$

F1-Score = $\frac{9}{9+5+9} = 0.692$

-	-	+
- (2)	0	0
- 2	5	0
- 1	1	2

Precision = $\frac{2}{2+2} = 0.5$

Recall = $\frac{2}{2+0} = 1$

F1-Score = $\frac{(2)(0.5)(1)}{1.5} = 0.667$

$2+2+2$

$1+5+9$

Precision = $\frac{9}{9+4} = 0.692$

Recall = $\frac{9}{9+4} = 0.692$

F1-Score = $\frac{9}{9+5+9} = 0.692$

Average F1-score = $\frac{(0.571+0.768+0.667)}{3} = 0.7$

3

LEC 07 K-Mean (Unsupervised)

K-mean for Clustering ຈົດກຸ່ມ No Ground Truth!

ຫຼັງ fit data ກຳນົດ No Label!

No Target!

ເຄີຍຄວາມກຸ່ມ \Rightarrow ກຸ່ມທີ່ຕໍ່ກຸ່ມ (ອຸປະນະທາງກາຍການ)

Proximity measurement (Ordinal Attribute; Distance)

Distance \rightarrow Euclidean ; $d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + \dots}$

\rightarrow Manhattan ; $d = |x_2 - x_1| + |y_2 - y_1| + \dots$

\rightarrow Minowski ; $d = ((x_2 - x_1)^p + (y_2 - y_1)^p + \dots)^{1/p}$

K-mean Clustering \rightarrow clustering ແບບ centroid-based

ໃຊ້ຈົດກຸ່ມການ centroid ໃນການຈົດກຸ່ມ ໂຍກົດເຄື່ອງຂ່າຍ/ທ່ານເຕີມການໄຟລູ້ຈົດກຸ່ມ

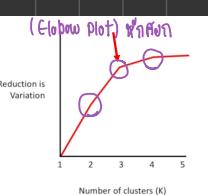
+ ກົດກຸ່ມຕ່າງໆ k ກົດ k ສິນເຫຼົ່າກຸ່ມ

+ ມາງອຸປະນະ initial centroid ຢູ່ກົດກຸ່ມ

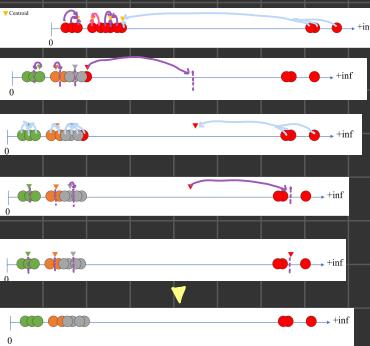
ການຈົດກຸ່ມ \downarrow centroid ອົບເວັດແກ່ກຸ່ມ

+ ເພີ້ມອຸປະນະ centroid ໃປໄສຮອງການຈົດກຸ່ມ

ກຳທັນເຮັດວຽກ centroid ລົງ



ex.



ex.

Centroid	Objects	
	A1	A2
c_1	3.8	9.9
c_2	7.8	12.2
c_3	6.2	18.5

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

New centroid
cluster ① = A_1, A_2
 $c_1 = (4.6, 7.1)$

$$\text{Cluster } ② = \left(\frac{6.8 + 7.8 + 8.4 + 9.6}{4}, \frac{12.6 + 12.2 + 6.9 + 11.1}{4} \right) = (8.2, 10.7)$$

$$\text{Cluster } ③ = (6.6, 18.6)$$

ຄວາມກຸ່ມໃນກົດກຸ່ມ

① k-mean ໄມໄຟທ່ານກະໜົນທີ່ກົດກຸ່ມ clustering (ເພຈະກົດກຸ່ມ 1-3 ພົມ)

x : object

n : number of object

C_i : Cluster ກົດກຸ່ມທີ່

c_i : centroid of C_i

n_i : number of object in the Cluster C_i

C : centroid of all object

k : group of clusters

"SSE sum square error" = $\sum_{i=1}^k \sum_{j=1}^{n_i} (x_j - c_i)^2$ \rightarrow distance

ຢູ່ຈົດກຸ່ມຢູ່ດີ but $\neq 0$

ເສັ້ນ k ແບບ elbow plot (ໜັກສອນ)

k	SSE
1	62.8
2	12.3
3	9.4
4	9.3
5	9.2
6	9.1
7	9.05
8	9.0

k ທີ່ໃຫຍ່ SSE ລວມ

ທ່ານນີ້ແນະນຸຍໃນການເລືອກຕ່າງໆ k

④ ການຈົດກຸ່ມ initial centroid (centroid ຫຼັງຕົ້ນ) ໃຫ້ສະກະສົງ

ວາງ centroid ປັດ ກຸ່ມເປົ້າຫຼື

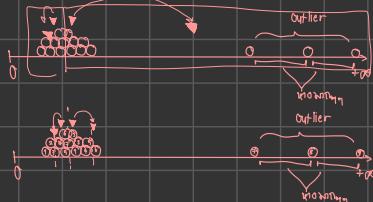
③ ຊົມຈຸດກຸ່ມໃຫຍ່ ກົດກຸ່ມໃຫຍ່ k-mean

Outlier ເຊັ່ນ, ຜົນງານກະຈາຍຕົ້ງ (non-convex), ເຕັກະ class ນາດຕະຫຼາດກົດກຸ່ມ

Variant of K-Means

centroid ກົດກຸ່ມ Med ເຊັ່ນ mean ບໍ່ສົ່ງໃຫຍ່ outlier ເຊັ່ນ:

\Rightarrow K-Medoids



$$(sum absolute error) SAE = \sum_{i=1}^k \sum_{\substack{x \in M_i \\ x \notin C_m \\ x \in C_i}} |x - c_m|$$

ສໍາງກົດກຸ່ມຕ່າງໆ ພັມກົດກຸ່ມ (PAM) (Partitioning around medoid)

ການຕິດກຸດກຸ່ມບໍ່ສົ່ງໃຫຍ່ k-means

(ເກົ່າກວ່າ) robust ກັບ outlier ແລະ outlier ວິກາ k-means

ໄວ່ເວລັກກະບົບຂອງກົດກຸ່ມໃຫຍ່ (time complexity)