Đánh giá mô hình phân lớp

(Classification Model Assessment)

Phan Xuân Hiếu

Khoa Công nghệ Thông tin
Trường ĐH Công nghệ (UET), ĐHQG Hà Nội (VNU)
hieupx@vnu.edu.vn

(last updated: 25–02–2016)

Nội dung

1. Đo đạc hiệu quả phân lớp (classification performance measures
   - Tỉ lệ lỗi (error rate) và độ chính xác (accuracy)
   - Ma trận lỗi/sai số (confusion matrix)
   - Precision, recall, f–measure
   - Receiver operating characteristics (ROC) và AUC (Area Under ROC)
2. Đánh giá mô hình phân lớp (classification evaluation)
   - Kiểm tra chéo (k–fold cross validation)
   - Bootstrap resampling
   - Confidence intervals
   - So sánh hai mô hình phân lớp với t–test
3. Các dạng hàm lỗi (loss functions)
4. Một số vấn đề khác
   - Quá khớp (overfitting)
   - Lựa chọn phương pháp/mô hình phân lớp
   - Triển khai mô hình phân lớp

Nội dung

Mô hình phân lớp

- Một mô hình phân lớp là một hàm (ánh xạ) có dạng $f : \mathbf{X} \rightarrow \mathbf{C}$. Cụ thể hơn:

$$\hat{y} = f(\mathbf{x}) \tag{1}$$

  Trong đó:

  - $\mathbf{x} \in \mathbf{X}$ là một đối tượng dữ liệu cần phân lớp.
  - $\hat{y} \in \mathbf{C} = \{c_1, c_2, \ldots, c_k\}$ là nhãn lớp được đoán nhận bởi mô hình f cho đối tượng $\mathbf{x}$.

- Để xây dựng mô hình f, chúng ta cần:
  - Một phương pháp phân lớp như naive Bayes, cây quyết định, SVMs, ...
  - Huấn luyện mô hình f dựa trên một tập dữ liệu có gắn nhãn (labeled training dataset) $\mathbf{D}_T = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_m, y_m)\}$, trong đó $\mathbf{x}_i \in \mathbf{X}$ và $y_i \in \mathbf{C}$, $i = 1..m$.

- Sau khi huấn luyện, cần đánh giá hiệu quả của mô hình phân lớp f trên một tập dữ liệu kiểm thử có gắn nhãn riêng biệt (labeled test dataset): $\mathbf{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}$.

Nội dung

1. Đo đạc hiệu quả phân lớp (classification performance measures
   - Tỉ lệ lỗi (error rate) và độ chính xác (accuracy)
   - Ma trận lỗi/sai số (confusion matrix)
   - Precision, recall, f–measure
   - Receiver operating characteristics (ROC) và AUC (Area Under ROC)
2. Đánh giá mô hình phân lớp (classification evaluation)
   - Kiểm tra chéo (k–fold cross validation)
   - Bootstrap resampling
   - Confidence intervals
   - So sánh hai mô hình phân lớp với t–test
3. Các dạng hàm lỗi (loss functions)
4. Một số vấn đề khác
   - Quá khớp (overfitting)
   - Lựa chọn phương pháp/mô hình phân lớp
   - Triển khai mô hình phân lớp

Tỉ lệ lỗi (error rate) và độ chính xác (accuracy)

- Với mỗi đối tượng $\mathbf{x}_i$ trong tập dữ liệu kiểm thử $\mathbf{D}$, y_i là nhãn lớp thật sự (true class label) và $\hat{y}_i$ là nhãn lớp đoán nhận bởi mô hình phân lớp f của $\mathbf{x}_i$, tức $\hat{y}_i = f(\mathbf{x}_i)$.
- Tỉ lệ lỗi của mô hình f trên tập $\mathbf{D}$ là số phần trăm đối tượng bị đoán nhận sai nhãn lớp bởi mô hình f:

$$ErrorRate = \frac{1}{n}\sum_{i=1}^{n} \mathbf{I}(\hat{y}_i \neq y_i) \tag{2}$$

  Trong đó: $\mathbf{I}$ là hàm indicator cho giá trị 1 nếu $\hat{y}_i = y_i$ và bằng 0 nếu ngược lại.
- Độ chính xác là tỉ lệ đối tượng trong $\mathbf{D}$ được đoán nhận nhãn lớp chính xác:

$$Accuracy = \frac{1}{n}\sum_{i=1}^{n} \mathbf{I}(\hat{y}_i = y_i) = 1 - ErrorRate \tag{3}$$

Nội dung

Ma trận lỗi/sai số (confusion matrix) tổng quát

- Gọi $\mathbf{G} = \{\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_k\}$ là phân hoạch của các đối tượng trên tập dữ liệu kiểm thử $\mathbf{D}$ dựa trên nhãn lớp thật sự của chúng. Tức với $j = 1..k$, $\mathbf{D}_j = \{\mathbf{x}_i \in \mathbf{D} | y_i = c_j\}$.
- Ký hiệu $n_j = |\mathbf{D}_j|$ là số lượng đối tượng thuộc lớp c_j $(j = 1..k)$.
- Ký hiệu $\mathbf{R} = \{\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_k\}$ là một phân hoạch của các đối tượng trên $\mathbf{D}$ dựa trên nhãn lớp được đoán nhận (predicted class labels). Tức với $j = 1..k$, $\mathbf{R}_j = \{\mathbf{x}_i \in \mathbf{D} | \hat{y}_i = c_j\}$.
- Ký hiệu $m_j = |\mathbf{R}_j|$ là số lượng đối tượng có nhãn lớp được đoán nhận là c_j $(j = 1..k)$.
- Khi đó, từ hai phân hoạch $\mathbf{G}$ và $\mathbf{R}$, chúng ta có thể tạo lập ma trận lỗi/sai số (confusion matrix) có kích thước $k \times k$, trong đó, giá trị tại hàng i, cột j được định nghĩa như sau:

$$\mathbf{N}(i,j) = n_{ij} = |\mathbf{R}_i \cap \mathbf{D}_j| = |\{\mathbf{x}_a \in \mathbf{D} | \hat{y}_a = c_i \text{ và } y_a = c_j\}| \quad (4)$$

Ma trận lỗi/sai số (confusion matrix) tổng quát (2)

	True class labels				
Predicted class labels	c_1	c_2	$\dots$	c_k	
c_1	n_{11}	n_{12}	$\dots$	n_{1k}	m_1
c_2	n_{21}	n_{22}	$\dots$	n_{2k}	m_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
c_k	n_{k1}	n_{k2}	$\dots$	n_{kk}	m_k
	n_1	n_2	$\dots$	n_k	$n = \|\mathbf{D}\|$

- n_{ii} là số đối tượng có nhãn lớp đoán nhận (predicted class labels) và nhãn lớp thực sự (true class labels) giống nhau.
- n_{ij} là số đối tượng được đoán nhận thuộc lớp c_i trong khi thực sự thuộc lớp c_j.

Nội dung

Độ chính xác, độ hồi tưởng, và f–measure cho từng lớp

- Độ chính xác từng lớp (per-class accuracy/precision): đối với mỗi lớp c_i, độ chính xác phân lớp được tính:

$$accuracy_i = precision_i = \frac{n_{ii}}{m_i}$$

  Với m_i số đối tượng được mô hình f đoán nhận thuộc lớp c_i.

- Độ phủ hay độ hồi tưởng từng lớp (per-class coverage/recall): đối với mỗi lớp c_i, độ hồi tưởng được tính:

$$coverage_i = recall_i = \frac{n_{ii}}{n_i}$$

  Với n_i số đối tượng thực sự thuộc lớp c_i.

- Độ đo hài hoà, F_1–measure, giữa độ chính xác và độ hồi tưởng được tính như sau:

$$F_1\text{–measure}_i = \frac{2 \times precision_i \times recall_i}{precision_i + recall_i} = \frac{2 \times n_{ii}}{n_i + m_i}$$

Ví dụ về ma trận lỗi/sai số

	True class labels			
Predicted class labels	Iris–setosa (c_1)	Iris–versicolor (c_2)	Iris–virginica (c_3)	
Iris–setosa (c_1)	10	0	0	$m_1 = 10$
Iris–versicolor (c_2)	0	7	5	$m_2 = 12$
Iris–virginica (c_3)	0	3	5	$m_3 = 8$
	$n_1 = 10$	$n_2 = 10$	$n_3 = 10$	$n = 30$

- $\text{Precision}_{c_1} = \frac{n_{11}}{m_1} = \frac{10}{10} = \mathbf{1.0}$; $\text{Recall}_{c_1} = \frac{n_{11}}{n_1} = \frac{10}{10} = \mathbf{1.0}$; $\text{F–measure}_{c_1} = \frac{2 \times n_{11}}{n_1 + m_1} = \frac{20}{20} = \mathbf{1.0}$
- $\text{Precision}_{c_2} = \frac{n_{22}}{m_2} = \frac{7}{12} = \mathbf{0.583}$; $\text{Recall}_{c_2} = \frac{n_{22}}{n_2} = \frac{7}{10} = \mathbf{0.7}$; $\text{F–measure}_{c_2} = \frac{2 \times n_{22}}{n_2 + m_2} = \frac{14}{22} = \mathbf{0.636}$
- $\text{Precision}_{c_3} = \frac{n_{33}}{m_3} = \frac{5}{8} = \mathbf{0.625}$; $\text{Recall}_{c_3} = \frac{n_{33}}{n_3} = \frac{5}{10} = \mathbf{0.5}$; $\text{F–measure}_{c_3} = \frac{2 \times n_{33}}{n_3 + m_3} = \frac{10}{18} = \mathbf{0.556}$

- Average F_1–measure $= \frac{1.0 + 0.636 + 0.556}{3} = \mathbf{0.731}$
- Accuracy $= \frac{n_{11} + n_{22} + n_{33}}{n} = \frac{10 + 7 + 5}{30} = \frac{22}{30} = \mathbf{0.733}$
- Macro average: Precision $= \frac{1.0 + 0.583 + 0.625}{3} = \mathbf{0.736}$; Recall $= \frac{1.0 + 0.7 + 0.5}{3} = \mathbf{0.733}$
- Macro average: F_1–measure $= \frac{2 \times 0.736 \times 0.733}{0.736 + 0.733} = \mathbf{0.734}$

Ma trận lỗi/sai số trong phân lớp nhị phân

- Với $k = 2$, tức $\mathbf{C} = \{c_1, c_2\}$, khi đó quy ước c_1 là lớp dương (positive class), c_2 là lớp âm (negative class).
- Ma trận lỗi/sai số (confusion matrix) có dạng:

	True class labels	
Predicted class labels	Positive (c_1)	Negative (c_2)
Positive (c_1)	True Positive (TP)	False Positive (FP)
Negative (c_2)	False Negative (FN)	True Negative (TN)

- ErrorRate $= \frac{FP+FN}{n}$; Accuracy $= \frac{TP+TN}{n}$
- $\text{Precision}_{c_1} = \frac{TP}{TP+FP}$; $\text{Precision}_{c_2} = \frac{TN}{TN+FN}$
- Sensitivity (True Positive Rate): TPR $= \text{Recall}_{c_1} = \frac{TP}{TP+FN}$
- Specificity (True Negative Rate): TNR $=$ Specificity $= \text{Recall}_{c_2} = \frac{TN}{TN+FP}$
- False Negative Rate: FNR $= \frac{FN}{TP+FN} = 1 - \textit{Sensitivity}$
- False Positive Rate: FPR $= \frac{FP}{FP+TN} = 1 - \textit{Specificity}$

Nội dung

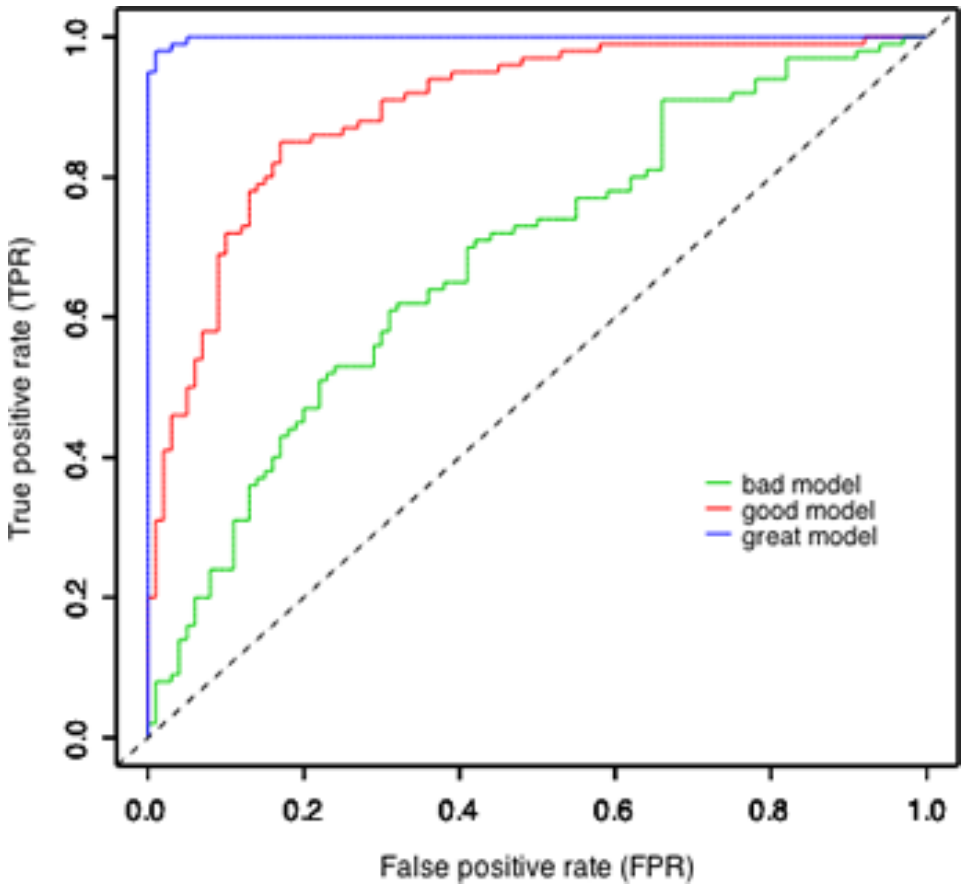
Receiver operating characteristics (ROC)

- Là một cách thức để đánh giá hiệu quả của bộ phân lớp (classifiers) khi có hai lớp đầu ra (nhị phân).
- Cần biết giá trị điểm cho lớp dương (output score value for positive class) tương ứng với mỗi đối tượng $\mathbf{x}_i$ trong tập dữ liệu kiểm thử (test dataset) $\mathbf{D}$. Ký hiệu giá trị điểm cho lớp dương là $S(\mathbf{x}_i)$. Ví dụ:
  - Với phân lớp naive Bayes hoặc MaxEnt: có thể sử dụng giá trị xác suất $P(c_1|\mathbf{x}_i)$ (giả sử c_1 là positive class).
  - Với phân lớp SVM: có thể sử dụng khoảng cách từ đối tượng đến siêu phẳng (hyperplane) chia tách.
- Cho một bộ phân lớp nhị phân (binary classifier), cần tìm một ngưỡng điểm α cho lớp dương. Cụ thể:
  - Lớp dương sẽ bao gồm tất cả các đối tượng $\mathbf{x}_i$ thoả mãn $S(\mathbf{x}_i) \geq \alpha$.
  - Lớp âm sẽ là tất cả $\mathbf{x}_i$ với $S(\mathbf{x}_i) < \alpha$.

Receiver operating characteristics (ROC) - tiếp

- Gọi $\alpha^{min} = \min_i S(\mathbf{x}_i)$ và $\alpha^{max} = \max_i S(\mathbf{x}_i)$ tương ứng là giá trị điểm lớp dương cực tiểu và cực đại trên tập kiểm thử **D**.
- Phân tích ROC (ROC analysis) chính là xem xét hiệu quả của bộ phân lớp f khi α biến đổi trong khoảng $[\alpha^{min}, \alpha^{max}]$.
- Cụ thể, chúng ta vẽ đồ thị đường cong ROC trong không gian hai chiều với trục tung là *TPR* (True Positive Rate hay Sensitivity) và trục hoành là *FPR* (False Positive Rate hay 1 - Specificity) của mô hình f khi α biến đổi trong $[\alpha^{min}, \alpha^{max}]$.
- Có 3 trường hợp cụ thể:
  - Khi $\alpha > \alpha^{max}$: tất cả các đối tượng sẽ thuộc về lớp âm (negative class), khi đó $TPR = \frac{TP}{TP+FN} = FPR = \frac{FP}{FP+TN} = 0.0$. Khi đó, toạ độ của (FPR, TPR) sẽ là $(0, 0)$.
  - Khi $\alpha = \alpha^{min}$: tất cả các đối tượng sẽ thuộc về lớp dương (positive class), khi đó $TPR = \frac{TP}{TP+FN} = FPR = \frac{FP}{FP+TN} = 1.0$. Khi đó, toạ độ của (FPR, TPR) sẽ là $(1, 1)$.
  - Khi $\alpha^{min} < \alpha \leq \alpha^{max}$: khi đó, toạ độ $(FPR, TPR) = (\frac{FP}{FP+TN}, \frac{TP}{TP+FN})$.

Ý nghĩa của đường cong ROC



Nội dung

1. Đo đạc hiệu quả phân lớp (classification performance measures
   - Tỉ lệ lỗi (error rate) và độ chính xác (accuracy)
   - Ma trận lỗi/sai số (confusion matrix)
   - Precision, recall, f–measure
   - Receiver operating characteristics (ROC) và AUC (Area Under ROC)
2. Đánh giá mô hình phân lớp (classification evaluation)
   - Kiểm tra chéo (k–fold cross validation)
   - Bootstrap resampling
   - Confidence intervals
   - So sánh hai mô hình phân lớp với t–test
3. Các dạng hàm lỗi (loss functions)
4. Một số vấn đề khác
   - Quá khớp (overfitting)
   - Lựa chọn phương pháp/mô hình phân lớp
   - Triển khai mô hình phân lớp

Đánh giá hiệu quả mô hình phân lớp

- Đánh giá hiệu quả của mô hình f trên tập dữ liệu kiểm thử $\mathbf{D}$ theo một phương pháp/kết quả đo đạc θ nào đó.
- Kết quả đánh giá có thể phụ thuộc vào nhiều yếu tố:
  - Phụ thuộc vào cách chia tập dữ liệu huấn luyện và tập kiểm thử.
  - Phụ thuộc vào việc tập dữ liệu kiểm thử chứa những đối tượng quá dễ hoặc quá khó phân lớp.
- Chúng ta muốn biết kỳ vọng $E[\theta]$ của kết quả đo đạc θ bằng cách thực hiện đánh giá lặp đi lặp lại trên các tập dữ liệu kiểm thử khác nhau.
- Do chúng ta không biết phân bố thật của các đối tượng dữ liệu, chúng ta có thể ước lượng $E[\theta]$ theo một số cách:
  - Kiểm tra chéo (cross–validation).
  - Lấy mẫu lại (resampling).

Nội dung

Kiểm tra chéo (k–fold cross validation)

- Chia tập dữ liệu kiểm thử $\mathbf{D}$ thành k phần bằng nhau (gọi là folds): $\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_k$.
- Với mỗi $fold_i$ $(i = 1..k)$:
  - Tập huấn luyện: $\mathbf{D} \setminus \mathbf{D}_i = \bigcup_{j \neq i} \mathbf{D}_j$.
  - Tập kiểm thử: $\mathbf{D}_i$.
  - Huấn luyện mô hình phân lớp f_i trên tập huấn luyện.
  - Đánh giá hiệu quả của mô hình f_i trên $\mathbf{D}_i$ và thu được kết quả θ_i.
- Kỳ vọng và phương sai của θ:
  - Kỳ vọng $\hat{\mu}_\theta = E[\theta] = \frac{1}{k} \sum_{i=1}^{k} \theta_k$
  - Phương sai: $\hat{\sigma}^2_\theta = \frac{1}{k} \sum_{i=1}^{k} (\theta_i - \hat{\mu}_\theta)^2$
- Giá trị k thường là 5 hoặc 10. Nếu $k = |\mathbf{D}|$ thì còn được gọi là *leave–one–out cross–validation*.

Nội dung

Bootstrap resampling

- Cho tập dữ liệu gắn nhãn $\mathbf{D}$ gồm n đối tượng dữ liệu.
- Tiến hành xây dựng (lấy mẫu) tập $\mathbf{D}_i$ từ $\mathbf{D}$ bằng cách: lặp lại n lần việc chọn ngẫu nhiên một đối tượng từ $\mathbf{D}$ có hoàn lại (sampling with replacement). Như vậy $|\mathbf{D}_i| = |\mathbf{D}|$.
- Lấy mẫu k lần, chúng ta có k tập mẫu: $\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_i, \ldots, \mathbf{D}_k$.
- Do chọn ngẫu nhiên có hoàn lại, xác suất một đối tượng trong $\mathbf{D}$ được chọn là $\frac{1}{n}$ và xác suất không được chọn là $1 - \frac{1}{n}$.
- Vì một tập $\mathbf{D}_i$ có n phần tử nên xác xuất một đối tượng $\mathbf{x}_j \in \mathbf{D}$ không thuộc $\mathbf{D}_i$ (sau n lần lấy ngẫu nhiên có hoàn lại) là $P(\mathbf{x}_j \notin \mathbf{D}_i) = (1 - \frac{1}{n})^n \simeq e^{-1} = 0.368$.
- Như vậy, xác xuất để một đối tượng $\mathbf{x}_j \in \mathbf{D}_i$ sau k lần chọn ngẫu nhiên là $P(\mathbf{x}_j \in \mathbf{D}_i) = 1 - P(\mathbf{x}_j \notin \mathbf{D}_i) = 1 - 0.368 = 0.632$.

Đánh giá mô hình phân lớp với bootstrap resampling

1. Với $i = 1..k$, thực hiện:
   1. Lấy mẫu tập $\mathbf{D}_i$ từ tập $\mathbf{D}$ (chọn ngẫu nhiên n đối tượng có hoàn lại).
   2. Huấn luyện mô hình phân lớp f_i trên tập $\mathbf{D}_i$.
   3. Đánh giá hiệu quả mô hình f_i trên tập $\mathbf{D}_i$ và thu được θ_i.
2. Tính kỳ vọng: $\hat{\mu}_\theta = E[\theta] = \frac{1}{k}\sum_{i=1}^{k}\theta_k$
3. Tính phương sai: $\hat{\sigma}^2_\theta = \frac{1}{k}\sum_{i=1}^{k}(\theta_i - \hat{\mu}_\theta)^2$

- Về mặt xác suất, mỗi tập $\mathbf{D}_i$ có thể phủ 63.2% số phần tử trong $\mathbf{D}$.
- Do đó, kết quả đánh giá thường "lạc quan" hơn so với k–fold cross–validation.

Nội dung

Nội dung

1. Đo đạc hiệu quả phân lớp (classification performance measures
   - Tỉ lệ lỗi (error rate) và độ chính xác (accuracy)
   - Ma trận lỗi/sai số (confusion matrix)
   - Precision, recall, f–measure
   - Receiver operating characteristics (ROC) và AUC (Area Under ROC)
2. Đánh giá mô hình phân lớp (classification evaluation)
   - Kiểm tra chéo (k–fold cross validation)
   - Bootstrap resampling
   - Confidence intervals
   - So sánh hai mô hình phân lớp với t–test
3. Các dạng hàm lỗi (loss functions)
4. Một số vấn đề khác
   - Quá khớp (overfitting)
   - Lựa chọn phương pháp/mô hình phân lớp
   - Triển khai mô hình phân lớp

Nội dung

Nội dung

1. Đo đạc hiệu quả phân lớp (classification performance measures
   - Tỉ lệ lỗi (error rate) và độ chính xác (accuracy)
   - Ma trận lỗi/sai số (confusion matrix)
   - Precision, recall, f–measure
   - Receiver operating characteristics (ROC) và AUC (Area Under ROC)
2. Đánh giá mô hình phân lớp (classification evaluation)
   - Kiểm tra chéo (k–fold cross validation)
   - Bootstrap resampling
   - Confidence intervals
   - So sánh hai mô hình phân lớp với t–test
3. Các dạng hàm lỗi (loss functions)
4. Một số vấn đề khác
   - Quá khớp (overfitting)
   - Lựa chọn phương pháp/mô hình phân lớp
   - Triển khai mô hình phân lớp

Nội dung

Nội dung

Nội dung