

SEAMS ガイドライン

Ver2.5.16

SAMPLE

改訂履歴

版	発行日	内容	発行
1.0.0	2019/2/28	新規作成	WITZ
2.0.0	2020/2/29	2019 年度研究知見の反映	WITZ
:	:	各種追記（技術情報、業界動向等）	WITZ
2.5.0	2022/1/17	関連標準の概要 分類整理（付録 D～F） 5.15 ML のデザインパターン 追加 13.1.4. 国内ドローン法改正 追加 8.2.9 欧州 AI レギュレーションの届出手続要件 追記 米国、英国、北京の AI 原則 追記（8.12, 8.15, 8.16） GRVA AI ガイド、PAS 8800 追記（9.6, 9.7）	WITZ
2.5.1	2022/1/19	5.16 Interpretable Machine Learning 追加	WITZ
2.5.2	2022/1/20	5.17 AI 原則整理 追加	WITZ
2.5.3	2022/3/15	9.2 UL 4600 記述更新 9.5 ISO/TR 4804 記述更新 9.8 ISO 22737 追加	WITZ
2.5.4	2022/3/29	5.18 NIST AI RMF 追加 5.19～5.21 日本発行のガイドライン 追加	WITZ
2.5.5	2022/3/30	1.1～1.3 最新内容に合わせて更新 1.4 プロセス構築への活用方法を追加	WITZ
2.5.6	2022/4/1	5.22 信頼できる機械学習（IBM）追加 5.23 About ML Reference Document（PAI）追加 11.12 自動運転の安全性評価メトリクス 追加 13.1.2 ドローンの欧州法規について更新	WITZ
2.5.7	2022/4/11	5.18.3 NIST AI RMF WS 追加	WITZ
2.5.8	2022/4/18	2.3.7 レジリエンスエンジニアリング 追加 5.24 ETSI 追加 12 適用事例 補足説明追記 14 AI セキュリティ 追加 2.3.2 不適切な記述があるため改善	WITZ
2.5.9	2022/5/9	2.3.8 協調安全 追加	WITZ
2.5.10	2022/5/27	8.9 シンガポールの Framework&Toolkit に関して追記 13.4 LiDAR の機能安全対応事例 追加 14 AI セキュリティ に MLSE のガイドを追記	WITZ
2.5.11	2022/5/31	7.2.8 不確実性分析について説明追記	WITZ



2.5.12	2022/6/15	8.19 OECD AI フレームワーク 追加 11.13 AUTOSAR RS Safety 追加	WITZ
2.5.13	2022/7/28	2.1 ユースケースに関して更新	WITZ
2.5.14	2022/8/22	5.18 NIST AI RMF に、Second Draft の件を追記	WITZ
2.5.15	2022/9/12	5.20 フランス自主規制 追加 8.2.10 capAI 追加 13.1.2 ドローン 欧州活動 追記 13.2 IoT 住宅協調安全追記	WITZ
2.5.16	2022/11/11	5.18.4 NIST AI RMF WS (第3回) 追加 5.25 IEEE 発行「信頼できる技術の信頼できる証拠」追加 7.5 医療分野事例 人間中心の説明可能性 追加 8.3 IEEE 70xx シリーズ 内容更新	WITZ

目次

1. 本ガイドラインについて.....	1
1.1. 目的.....	1
1.2. 対象範囲.....	1
1.3. 想定読者.....	1
1.4. AI 開発プロセス構築への活用方法.....	2
1.5. 本ガイドラインによって改善できる AI の課題.....	3
1.6. 本ガイドラインの位置付け	4
1.7. QA4AI との関係.....	5
1.8. 用語集	5
2. 背景.....	10
2.1. 普及する AI 搭載システム	10
2.2. AI の分類整理	11
2.3. AI を搭載した安全関連システムの課題認識.....	14
2.3.1. AI の信頼性.....	14
2.3.2. 安全に影響のある AI の特性	14
2.3.3. AI の構成要素と信頼性	15
2.3.4. AI は機能安全規格では非推奨.....	16
2.3.5. AI の安全立証における課題	18
2.3.6. 説明可能性の高い AI.....	20
2.3.7. レジリエンスエンジニアリングの必要性	20
2.3.8. 協調安全の必要性.....	21
3. AI 搭載システムの安全設計パターン	24
3.1. AI 搭載システムの安全設計における注意点や課題.....	24
3.2. 方針 1 : AI 自体を安全設計する	25
3.2.1. 1a) 機能安全開発パターン	26
3.2.2. 1b) 安全性評価パターン.....	26
3.2.3. 1c) Proven-in-use パターン	26
3.2.4. 1d) 多重化設計パターン.....	28
3.2.5. ソフトウェアの不確かさへの対応	31
3.3. 方針 2 : AI は安全設計せず、外部に安全メカニズムを設ける	32
3.3.1. 2a) 監視機能パターン	32
3.3.2. 2b) 比較パターン	33
3.3.3. 2c) 防御設計パターン.....	34
3.4. AI 搭載システムの安全設計パターンの比較.....	36
3.5. 一時故障の影響と対策.....	36

4. AI 搭載システムを用いたサービスの安全開発プロセス	38
4.1. (1)サービスレベルのシステム定義	40
4.2. (2)ハザード分析及びリスクアセスメント	41
4.2.1. アクシデント・ハザード・安全制約の決定	42
4.2.2. コントロールストラクチャの作成	42
4.2.3. UCA の抽出	42
4.2.4. UCA を含めたユースケースでの非安全シナリオ導出	43
4.2.5. リスクアセスメント	43
4.2.6. 参考：シミュレーション環境による自動化	44
4.3. (3)サービスレベルの安全要求導出	45
4.4. (4)サービスレベルの安全要求割り付け	45
4.5. (5)システム単体のアーキテクチャ設計	45
4.6. (6)システムレベルの安全要求導出	45
4.7. (7)システム安全分析	45
4.8. (8)初期 AI コンポーネント開発	46
4.9. (9)AI 学習フェーズ	46
4.9.1. Automotive SPICE のテーラリング	46
4.9.2. SOTIF における AI 学習ワークフロー	47
4.9.3. 説明可能性の高い AI モデルワークフロー	49
4.10. (10)ハードウェア・ソフトウェア開発	49
4.11. (11)システムレベル統合試験	49
4.12. まとめ	49
5. 付録 A : AI 関連資料の調査	50
5.1. AI 開発ガイドライン素案へのパブコメ調査	50
5.2. 車載ソフトウェアで機械学習を扱うための提言論文	51
5.3. AI の研究開発の原則の策定	53
5.4. QA4AI ガイドラインの概要	54
5.5. WAISE 2020 の注目技術	56
5.5.1. AI の安全上の懸念事項への緩和策	56
5.5.2. ODD を考慮した安全性論証	57
5.5.3. ルールベースの安全エビデンス	58
5.6. 信頼できる AI の評価リスト (ALTAI)	59
5.6.1. 技術的な堅牢性と安全性	60
5.6.2. 透明性	62
5.7. AI の説明可能性へのアプローチ	63
5.7.1. モデリング前の説明可能性	63
5.7.2. 説明可能なモデリング	65

5.7.3.	モデリング後の説明可能性	66
5.8.	透明性モデル	67
5.9.	自律的でインテリジェントなシステムの倫理に関する IEEE グローバルイニシアチブ	68
5.10.	経済産業省「我が国の AI ガバナンスの在り方」	69
5.11.	VDE「AI 倫理を運用化するための学際的なフレームワーク」	69
5.12.	IEEE ユースケース—透明性、説明責任、およびコンタクトトレーシングのプライバシーに おける倫理的課題に対処するための基準—	70
5.13.	GPAI (Global Partnership on AI)	71
5.14.	AI の機能安全対応を検討した論文集	72
5.14.1.	ISO26262 対応への推奨事項	72
5.14.2.	DNN のトレーサビリティ	73
5.14.3.	ISO26262 ソフトウェアプロセス要求への評価と適合	75
5.14.4.	機械学習のライフサイクルの保証	75
5.15.	ML のデザインパターン	75
5.15.1.	Machine Learning Architecture and Design Patterns (早稲田大学 鷲崎先生)	75
5.15.2.	Machine Learning Design Patterns (Nastasia Saby)	80
5.15.3.	機械学習応用システムのアーキテクチャ・デザインパターン	81
5.16.	Interpretable Machine Learning	82
5.17.	国際論文提案されている AI の原則の整理	83
5.18.	NIST AI RMF	83
5.18.1.	概要	83
5.18.2.	リスク管理のフレームワークとしての観点例を記載	85
5.18.3.	NIST AI Risk Management Framework ワークショップ (第 2 回)	89
5.18.4.	NIST AI Risk Management Framework ワークショップ (第 3 回)	91
5.19.	AI 利活用ガイドライン	92
5.20.	AI を用いたクラウドサービスに関するガイドブック	94
5.21.	AI・データの利用に関する契約ガイドライン	95
5.22.	信頼できる機械学習	96
5.22.1.	目次	96
5.22.2.	まとめ	103
5.23.	About ML Reference Document	111
5.23.1.	目次	111
5.23.2.	チェックリスト	113
5.24.	欧州電気通信標準化機構 (ETSI) の AI 標準化	117
5.25.	IEEE 発行「信頼できる技術の信頼できる証拠」	119
6.	付録 B : IEC 62998 の概要と AI 搭載システムへの適合	120
6.1.	IEC 62998 の概要	120

6.2.	IEC 62998 の特徴	120
6.3.	IEC 62998 における開発の流れ	121
6.3.1.	安全関連システム(SRS/SRSS)の定義とハザード分析	121
6.3.2.	SRS/SRSS の目標パフォーマンスクラス決定	122
6.3.3.	SRS/SRSS の設計	122
6.3.4.	信頼性情報の評価（あるいは、SRS のパフォーマンスクラスの評価）	123
6.3.5.	SRSS のパフォーマンスクラスの評価	128
6.3.6.	統合と設置	129
6.3.7.	検証	129
6.4.	IEC 62998 の AI 搭載システムへの応用	130
6.5.	IEC 62998 の課題	130
6.6.	参考情報：IEC 62998 を満たしているセンサの例	131
7.	付録 C：説明可能性の高い AI とは	132
7.1.	説明可能な AI (XAI)	132
7.1.1.	XAI の概要	132
7.1.2.	XAI のための 4 つの基本原則	132
7.2.	説明可能性の高い AI モデル開発ワークフロー	133
7.2.1.	ワークフローの概要	133
7.2.2.	①探索的データ分析 (EDA) & 視覚化	134
7.2.3.	②ベンチマークや再現性の確立	136
7.2.4.	③手動な、個人的な、まばらな、わかりやすい特徴選択	136
7.2.5.	④公平性、プライバシー、セキュリティのための前処理	136
7.2.6.	⑤制約された、公平な、解釈可能な、個人的な、シンプルなモデル	137
7.2.7.	⑥予測の調整	138
7.2.8.	⑦伝統的なモデル評価&診断	139
7.2.9.	⑧説明	140
7.2.10.	⑨モデルデバッグ	141
7.2.11.	⑩社会的偏見の考査と修復	143
7.2.12.	⑪リスクの定量化と計画	143
7.2.13.	⑫人間のレビューと文書化	144
7.2.14.	⑬配備、管理、監視	144
7.2.15.	⑭自動化されたモデル決定のヒューマンアピール	145
7.2.16.	⑮根本原因分析	145
7.2.17.	⑯反復	145
7.2.18.	⑰廃止	146
7.3.	QA4AI ガイドラインの活用	146
7.4.	AutomotiveSPICE の活用	147

7.5. 医療分野事例 人間中心の説明可能性	148
8. 付録 D : 関連標準の概要 (AI 安全、AI 法規)	149
8.1. High-risk AI への規制要件	149
8.2. 欧州 AI レギュレーション (EU AI Act)	149
8.2.1. 目次.....	150
8.2.2. 本規制の目的や位置付け	155
8.2.3. AI システムの定義	156
8.2.4. 禁止される AI の条件.....	156
8.2.5. High-risk AI の条件と対応	157
8.2.6. High-risk AI への主な管理要件の例	160
8.2.7. High-risk AI への主な開発要件の例	160
8.2.8. 参考：要件への準拠費用の概算	162
8.2.9. 届出手続.....	162
8.2.10. EU AI Act 適合評価ツール capAI.....	162
8.3. IEEE 70xx シリーズ	166
8.4. IEC 61508 改訂.....	167
8.5. DIN SPEC 92001	168
8.6. VDE-AR-E 2842-61	169
8.7. BSI PAS 188x シリーズ	169
8.7.1. BSI PAS 1880.....	169
8.7.2. BSI PAS 1881.....	170
8.7.3. BSI PAS 1883.....	170
8.8. GERMAN STANDARDIZATION ROADMAP ON ARTIFICIAL INTELLIGENCE	170
8.9. シンガポールの AI ガバナンスに対するアプローチ	173
8.10. ドバイ AI 倫理自己評価ツール.....	176
8.11. アメリカ大統領令.....	179
8.12. アメリカ AI に関する 10 項目の規制原則	181
8.13. マルタ ETHICAL AI FRAMEWORK.....	183
8.14. カナダ Directive on Automated Decision-Making.....	184
8.15. 北京 AI 原則	186
8.16. 英国 AI 原則	187
8.16.1. AI 原則	187
8.16.2. データ倫理フレームワーク	188
8.17. AI の認証プログラム開発	188
8.18. 厚生労働省による AI 搭載システムの承認審査事例.....	188
8.19. OECD AI フレームワーク	188
8.19.1. 主要素と観点	189

8.19.2.	AI アプリケーション整理	190
8.20.	フランス自主規制	192
9.	付録 E : 関連標準の概要 (自動運転安全)	193
9.1.	ISO/PAS 21448 (SOTIF)	193
9.1.1.	ISO/PAS 21448 のプロセス概要	193
9.1.2.	ISO 26262 との関係	194
9.1.3.	ハザード分析&リスクアセスメントの重要性	194
9.1.4.	自動運転 Level 3 以上への適用	196
9.2.	UL 4600	196
9.2.1.	概要	196
9.2.2.	目次	198
9.2.3.	AI に関する要求項目	200
9.3.	WP29	200
9.4.	道路交通法の改訂	201
9.5.	ISO/TR 4804	201
9.5.1.	概要	201
9.5.2.	目次	202
9.5.3.	自動運転車に必要な機能要求	202
9.5.4.	自動運転車におけるモデルアーキテクチャ	205
9.6.	GRVA AI ガイドライン	206
9.7.	ISO/AWI PAS 8800	207
9.8.	ISO 22737	208
9.8.1.	概要	208
9.8.2.	目次	210
9.8.3.	運行設計領域 (ODD)	211
9.8.4.	最小リスク操作 (MRM)	212
9.8.5.	パフォーマンステスト手順の例	212
10.	付録 F : 関連標準の概要 (その他)	214
10.1.	IEC 62853 (DEOS)	214
10.2.	IEC 60721	214
10.3.	国や地域による安全の考え方の違い	216
11.	付録 G : 自動運転と安全	217
11.1.	SAE J3016 で定義されている自動運転のレベル	217
11.2.	自動運転における「安全状態」の例	218
11.3.	自動運転における「安全時間」の例	220
11.4.	自動運転システムの構成要素	220
11.5.	自動運転における一般的なセンサ・フュージョン	221

11.6.	自動運転システムにおける環境条件の例.....	222
11.7.	自動運転技術のリスクアセスメントの必要性.....	223
11.8.	運行設計領域 (ODD)	224
11.9.	自動運転に関する主なガイドライン.....	224
11.9.1.	国内.....	224
11.9.2.	海外.....	225
11.9.3.	Safety First for Automated Driving (SaFAD)	226
11.10.	自動運転業界動向の参考サイト.....	229
11.11.	自動運転の安全性評価フレームワーク	229
11.12.	自動運転の安全性評価メトリクス	230
11.13.	AUTOSAR RS Safety	232
12.	付録 H : 当社の具体的な適用事例	235
13.	付録 I : その他の分野における動向	236
13.1.	無人航空機システム (UAS)、ドローン	236
13.1.1.	米国活動.....	236
13.1.2.	欧州活動.....	237
13.1.3.	国際標準	238
13.1.4.	国内ドローン法改正 (航空法改正)	238
13.2.	IoT 住宅.....	239
13.3.	農業機械の自動走行.....	239
13.4.	LiDAR の機能安全対応の取り組み事例.....	240
13.4.1.	LeddarTech 社.....	240
13.4.2.	Innoviz Technologies 社.....	241
13.4.3.	Melexis 社、Massachusetts 大学.....	241
14.	付録 J : AI セキュリティ.....	242
14.1.	AI セキュリティとは.....	242
14.2.	参考資料.....	242
14.3.	AI の脆弱性	243
14.3.1.	画像認識	243
14.3.2.	音声認識.....	243
14.4.	AI へのサイバー攻撃の種類.....	244
14.4.1.	MLSE 機械学習システムセキュリティガイドライン.....	244
14.4.2.	神戸大学 小澤先生	244
14.4.3.	日本ネットワークセキュリティ協会.....	245
14.4.4.	日本銀行金融研究所	247
14.5.	AI セキュリティ対策	248
14.5.1.	MLSE 機械学習システムセキュリティガイドライン.....	248



14.5.2.	神戸大学 小澤先生	248
14.5.3.	日本ネットワークセキュリティ協会	249
14.5.4.	日本銀行金融研究所	251
14.6.	標準化動向.....	251
14.7.	第三者認証.....	252

SAMPLE

1d)	多重化設計パターン	アーキテクチャ設計アプローチ
-----	-----------	----------------

3.2.1. 1a) 機能安全開発パターン

AI を機能安全規格に準拠したプロセスで開発することにより、AI が十分に高い信頼性があることを示す方法である。

機能安全規格が要求する開発プロセスにおける検証精度（信頼性実現）のレベル感は、可能な限りのホワイトボックス検証（レビュー、ホワイトボックステストなど）による高信頼性部品化と、高信頼性部品の積み重ねによって実現するものである。つまり、ブラックボックステストのみの実施では達成できないことを意味する。さらに、検証（設計検証、テスト）の網羅性や、可能な限りの実システムに対するテスト実施も要求されている。

このような機能安全プロセス相当の実現が期待できる技術として、「説明可能な AI (eXplainable AI; XAI)」があるが、2.3.6 節で説明したように、現時点では有用な手法は見出されていない。

しかしながら、今後もさまざまな手法が提唱されることが期待されるため、引き続き注視することが必要である。

3.2.2. 1b) 安全性評価パターン

AI について機能安全規格適合相当の安全性評価を行う手法である。ブラックボックス評価手法の位置づけになる。

有力な手法として、センサシステムの機能安全規格 IEC 62998 の評価方法が適用可能だと考える。IEC 62998 の詳細は、6 付録 B を参照されたい。

適用可能と考える理由は、複雑なセンサシステムと AI の特性が似ているためである。例えば、複雑な外部環境の影響を受けセンサの出力値は変化しやすいが、人間の期待値とは異なる情報（非決定的な情報）を出力する可能性がある。この点は、AI の非決定的な特性と類似している。

さらに、IEC 62998 はそもそも AI を含むセンサシステムも対象とした規格である。また、IEC 62998 は屋外環境や一般人との接触環境など、非常に複雑な環境まで対象とした規格である。

一方で課題としては、以下が挙げられる。

- ・ IEC 62998 の知名度（存在認識）が低い。特に自動車業界では知られていない。
- ・ AI に対して IEC 62998 で評価した事例が、現時点では存在していない。
- ・ AI に対して IEC 62998 で評価することを認める国際認証機関が、現時点では存在していない。

これらの課題は、AI の評価方法が国際規格として具体化することで、解消されることが期待される。

3.2.3. 1c) Proven-in-use パターン

AI について、機能安全規格が要求する Proven-in-use（使用実績）基準を示すことによって、十分な安全性があると評価する方法である。ブラックボックス評価手法の位置づけになる。

例えば、IEC 61508-7:2010 C.2.10.1 に記載の基準の一例としては、対象コンポーネントは以下を満

6.3. IEC 62998 における開発の流れ

本節では IEC 62998 を使用したセンサシステム開発の流れを説明する。大枠としては機能安全開発と大きな違いはなく、以下 6.3.1 以降の手順を順に実施する。なお、「SRS/SRSS の実装」については単にセンサの製造部分であるため本書では特に説明を記載しない。

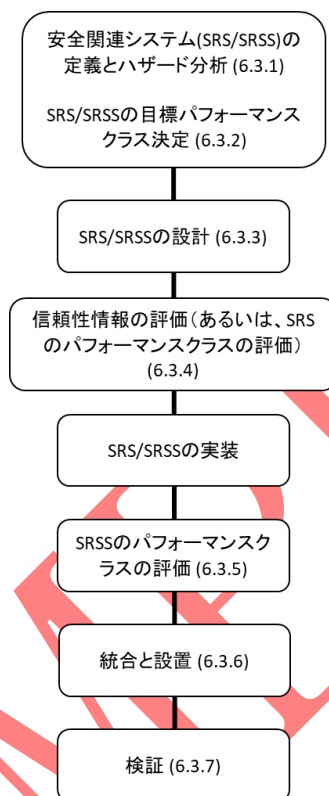


図 17 SRS 開発の流れ

6.3.1. 安全関連システム(SRS/SRSS)の定義とハザード分析

安全関連システム(SRS/SRSS)は、図 18 のように「センシングユニット」「処理ユニット」「入出力ユニット」で構成される。



図 18 SRS アーキテクチャの例

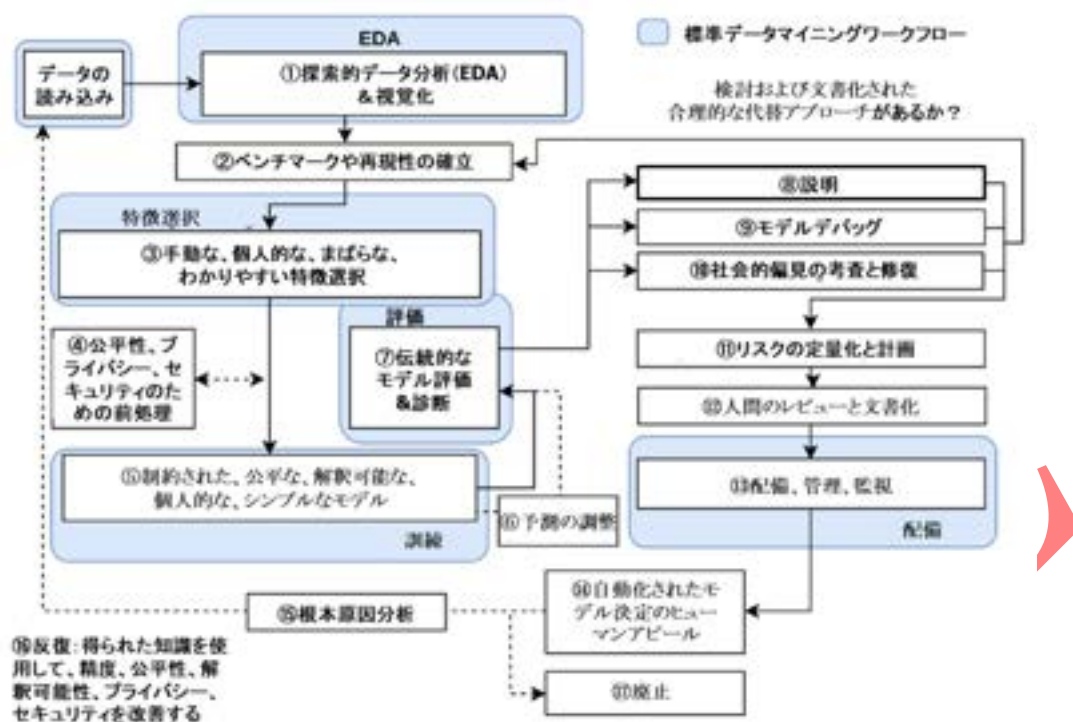


図 22 説明可能性の高い AI モデルワークフロー

7.2.2. ①探索的データ分析（EDA）＆視覚化

<概要>

AI 開発のスタートとして、対象システムの特徴を十分に把握しておくことが重要である。

システム全体を考慮し適切なデータセットの選定を行う。そのために、データを理解しやすいように視覚化しながら、データ分析を繰り返す。

単純な可視化の例としては、クラスごとの学習データの分布を見える化する。

また、ビッグデータの場合は、データ全体の分布などを解析し、分散や偏りなどからデータセットの補正を行う。

最初のシステム設計段階において、何を後処理として調停するかを設計することがある。視覚化を行いながら、システム全体の設計を行うことが望ましい。

公平性のある後処理技術として以下がある。

- ・ Reject option-based classification

<OSS>

- ・ H2O-3 Aggregator

<参考文献>

- ・ 論文 : Visualizing Big Data Outliers through Distributed Aggregation

<参考文献>

- ・論文 : Predicting Good Probabilities with Supervised Learning
<https://www.cs.cornell.edu/~alexn/papers/calibration.icml05.crc.rev3.pdf>

7.2.8. ⑦伝統的なモデル評価&診断

<概要>

予測結果に対して、モデルの評価を行う。

残差分析、QQ プロット、AUC、リフト曲線などにより、モデルが正確であり、仮定基準を満たしているかを評価する。感度分析という手法を用いることもある。

また、予測結果と併せて、不確実性の情報が提供されることもある。これに対して、不確実性分析を実施して、モデルを評価することもある。

Deep Learning では、学習曲線、LOSS 関数によって評価することが一般的である。基準を満たさなかった場合、特徴量選択のフェーズに戻り再調整を行う。評価においては、AI モデル単体としての評価、システムレベルを考慮した評価が必要である。本工程では、工程②で決めたゴール設定に対して評価を行う。

<感度分析>

感度分析は、データを意図的に変更しシミュレートしたときに、モデルの振る舞いと出力が受け入れられるかどうかを調べることである。予測の感度分析は、従来の評価手法以上に機械学習モデルの最も重要な検証およびデバッグ手法である。

従来の線形モデルの場合、入力変数間または入力変数とターゲット変数間の相関関係に起因する回帰パラメータの数値的不安定性に焦点を当てている。

一方、機械学習の場合、モデルパラメーターの数値的不安定性にあまり焦点を当てず、モデル予測の潜在的な不安定性にもっと焦点を合わせるのが賢明である。機械学習アルゴリズムは入力変数値に対して、非常に複雑な非線形、非単調な応答をするので、直接シミュレーションでモデルの予測をテストする方が、隠れた相関関係について静的なトレーニングデータを検索するよりも時間を有効に使用できる。

<不確実性分析>

不確実性とは、予測が確実ではないこと、つまり事前に正しく予測できないことです。例えば、身近なところでは、明日の天気は何か、サイコロの目が何が出るか、ポーカーで相手に何が配られるか、などが該当します。天気予報は当たりもすれば外れもしますので、不確実なものです。また、株価の増減や、この製品が売れるか否かなど、ビジネスの世界においても不確実性なことが多々存在します。

ビジネスにおける不確実性分析では、将来の数値をシミュレーションしたり、将来の状況の仮説を立てたり、将来発生しうるかも知れないリスクを分析したりします。感度分析もその一種で、影響要因の重要度を決めて、変化する影響要因の感度を評価する手法です。