

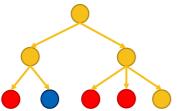
Cheat Sheet: Machine Learning mit KNIME Analytics Platform

Betreutes Lernen: Eine Reihe von maschinellen Lernalgorithmen, um den Wert einer Zielklasse oder Variablen vorherzusagen. Sie erzeugen eine Mapping-Funktion (Modell) von den Eingabefunktionen zur Zielklasse/variable. Zur Schätzung der Modellparameter während der Trainingsphase werden markierte Beispieldaten im Trainingsset benötigt. Die Verallgemeinerung auf ungesiehene Daten wird auf den Testsetsdaten über Scoring Metriken ausgewertet.

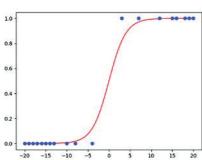
Klassifizierung

Einstufung: Eine Art beaufsichtigtes Lernen, wo das Ziel eine Klasse ist. Das Modell lernt eine Klassennote zu erzeugen und jedem Vektor von Eingabemerkmale der Klasse mit der höchsten zuzuordnen Score. Es können Kosten eingeführt werden, um eine der Klassen während der Klassenzuordnung zu bestrafen.

Entscheidung Tree: Folgen Sie dem C4.5-Entscheidungsbaum Algorithmus. Diese Algorithmen erzeugen eine baumähnliche Struktur, Erstellung von Datensubsets, aka Baumknoten. An jedem Knoten werden die Daten geteilt basierend auf einer der Eingangsmerkmale, Erzeugung zweier oder mehr Zweige als Ausgang. Weitere Spalten werden in folgenden Knoten gemacht, bis ein Knoten generiert, wo alle oder fast alle Daten gehören zur gleichen Klasse.



Logistische Regression: Ein statistischer Algorithmus, die Beziehung zwischen dem Eingang Merkmale und die kategorischen Ausgabeklassen von Maximierung einer Wahrscheinlichkeitsfunktion. Ursprünglich entwickelt für binäre Probleme, es wurde auf Probleme mit mehr als zwei Klassen (multinomische logistische Regression).

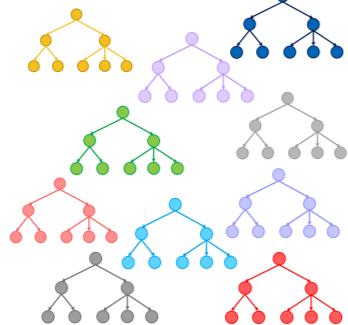


Unternehmen

Ensemble Learning: Eine Kombination mehrerer Modelle von überwachten Lernalgorithmen bis ein stabileres und genaues Gesamtmodell erhalten. Am häufigsten verwendete Ensembletechniken sind Bagging und Boosting.

BAGG

Taschen: Eine Methode zum Training mehrerer Klassifikations-/Regressionsmodelle auf verschiedenen zufällig gezeichnete Teilmengen der Trainingsdaten. Die endgültige Vorhersage basiert auf der Prognosen aller Modelle, so die Chance zu reduzieren, zu übertreffen.

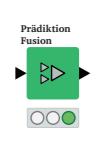


Tree Ensemble der Entscheidung/Regression Bäume: Ensemblemodell mehrerer Entscheidungs-/Regressions Bäume, die auf verschiedenen Teilmengen von Daten trainiert. Datensubsets mit weniger oder gleichen Zeilen und weniger oder gleiche Säulen sind von der Original-Trainingsset. Schlussvorhersage basiert über eine harte Stimme (Majoritätsregel) oder Soft vote (alle Wahrscheinlichkeiten oder numerische Prognosen) auf allen beteiligten Bäumen.

Random Forest of Decision/Regression Trees: Ensemblemodell mehrerer Entscheidungen/Regressions Bäume, die auf verschiedenen Teilmengen von Daten trainiert. Datensubsets mit der gleichen Anzahl von Zeilen werden vom Original-Training abgeschnitten gesetzt. An jedem Knoten wird die Spaltung auf einer subset von sqrt(x) Features aus dem Original x Eingabemerkmale. Die letzte Vorhersage basiert auf einer harten Stimme (Majorität-Regel) oder weichstimmt (alle Wahrscheinlichkeiten oder numerische Prognosen) auf allen beteiligten Bäumen.

Custom Ensemble Modell: Kombination verschiedener überwachter Modelle zu einem benutzerdefinierten Ensemblemodell. Das Finales Vorhersage kann auf Mehrheitsabstimmung sowie mittlere oder andere Funktionen die Ausgabe ergibt.

XGBoost: Eine optimierte verteilte Bibliothek für maschinelle Lernmodelle in der Gradient-Boost-Rahmen, entworfen, um hocheffizient, flexibel und tragbar. Es verfügt über Regelparameeter, um komplexe Modelle zu penalisieren, effektives Handlung von Spardeaten für bessere Leistung, parallele Berechnung, und effizientere Speichernutzung.



Geschäftsbedingungen

Numerische Vordition & Klassifizierung

Künstliche Neuralnetze (ANN, NN): Inspiriert von biologischem Nerven System, Künstliche neuronale Netzwerke basieren auf Architekturen von miteinander verbundene Einheiten genannt künstliche Neuronen. Künstliche Neurone Parameter und Verbindungen sind trainiert über dedizierte Algorithmen, das populärste Wesen der Back-Propagation Algorithmus.

Support Vector Machine (SVM): Ein überwachter Algorithmus-Konstrukt... eine Reihe von Diskriminierung Hyperplane in hochdimensionalen Raum. Zusätzlich zu linear Klassifizierung, SVMs kann nichtlineare Klassifizierung nach implizit ihre Eingänge in hochdimensionale Spielräume, wobei die beiden Klassen linear sind trennbar.

k-Nearest Neighbor (kN): Eine nichtparametrische Methode, die die Klasse der k am meisten zuordnet ähnliche Punkte in den Trainingsdaten, basierend auf einem vordefinierten Abstand Maßnahme. Klassenzuschreibung kann gewichtet durch den Abstand zum k-ten Punkt und/oder durch die Klasse Wahrscheinlichkeit.

Allgemeines Linearmodell (GLM): Eine statistikbasierte Flexibilität Verallgemeinerung gewöhnlicher linearer Regression, gültig auch für nicht normal Verteilungen der Zielgröße. GLM verwendet die lineare Kombination von den Eingabemerkmale zum Modell a beliebige Funktion des Ziels variabel (die Linkfunktion) eher als die Zielgröße selbst.

Numerische Vordition

Numerische Vorhersage: Eine Art überwachtes Lernen für numerische Zielvariablen. Das Modell lernt eine oder mehrere Zahlen mit dem Vektor zu verknüpfen von Eingabemerkmale. Beachten Sie, dass numerische Prädiktionsmodelle auch trainiert werden können, um Klassenspunkte vorherzusagen und daher für die Klassifizierung verwendet werden können.

Linear/Polynom Regression: Lineare Regression ist ein statistischer Algorithmus zur Modellierung einer multivariate lineare Beziehung zwischen numerische Zielgröße und die Eingabemerkmale. Polynom Regression erweitert dieses Konzept um Bestücken einer Polynomfunktion einer vordefinierten Grad.

Deep Learning: Deep Learning die Familie der ANNs mit tiefere Architekturen und zusätzliche Paradigmen, z. Recurrent Neural Networks (RNN). Die Ausbildung solcher Netze hat durch die jüngsten Fortschritte aktiviert auch in Hardwareleistung als parallele Ausführung.

Regression Tree: Erstellt einen Entscheidungsbaum, um vorherzusagen numerische Werte durch einen rekursiven, top-down, gieriger Ansatz bekannt als rekursive binäre Aufteilen. In jedem Schritt teilt der Algorithmus die durch jeden Knoten in zwei oder mehr neue Zweige mit einer gierigen Suche nach dem am besten geteilt. Der Durchschnittswert der Punkte in einem Blatt produziert die numerische Vorhersage.

Allgemeines Linearmodell (GLM): Eine statistikbasierte Flexibilität Verallgemeinerung gewöhnlicher linearer Regression, gültig auch für nicht normal Verteilungen der Zielgröße. GLM verwendet die lineare Kombination von den Eingabemerkmale zum Modell a beliebige Funktion des Ziels variabel (die Linkfunktion) eher als die Zielgröße selbst.

Analyse

Numerische Vordition

ANALYSE

Zeitreihenanalyse: Eine Reihe von numerischen Vorhersagemethoden zur Analyse/Prädikt-Zeitreihen Daten. Zeitreihen sind zeitlich geordnete Sequenzen von numerischen Werten. Insbesondere Zeitreihen Prognose zielt darauf ab, zukünftige Werte basierend auf zuvor beobachteten Werten vorherzusagen.

Saisonale autoregressive integrierte Moving Average (SARIMA): Saisonales (S) Auto-Regressive (AR) Modell wird auf einem bestimmten Anzahl p früherer (saisonaler) Werte; Daten werden nach einem Grad erstellt Differenzierung d zur Korrektur der Nichtstationarität; während einer linearen Kombination - benannt Moving Average (MA) - modellt q Vergangenheit (saisonale) Rest Fehler. Alle SARIMA-Modellparameter werden gleichzeitig von verschiedenen Algorithmen, meist nach dem Box-Jenkins-Ansatz.

ML-basierte TSA: Ein numerisches Vorhersagemodell, das auf Vektoren vergangener Werte können den aktuellen numerischen Wert vorhersagen der Zeitreihe.

Long Short Term Memory (LSTM) Units: LSTM-Einheiten einen versteckten Zustand erzeugen, indem m x n Tensors Eingangswerte, wobei in die Größe des Eingangsvektors an jedem Zeit und n die Anzahl der letzten Vektoren. Der verborgene Zustand kann dann in den aktuellen Vektor der Ziffer transformiert werden Werte. LSTM-Einheiten eignen sich zur Zeitreihenvorhersage als Werte aus vergangenen Vektoren können gemerkt oder vergessen werden durch eine Reihe von Toren.

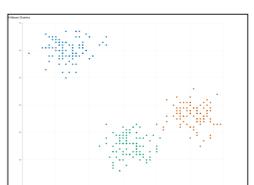
UNSUPERVISED LEARNING

Unsupervised Learning: Eine Reihe von Maschinen Lernalgorithmen, um Muster in der Daten. Ein markierter Datensatz ist nicht erforderlich, da Daten letztendlich organisiert und/oder transformiert auf der Grundlage von Ähnlichkeiten oder statistischen Maßnahmen.

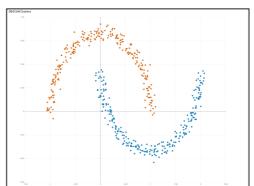
Schlussfolgerung

Clustering: Ein Zweig des ununterbrochenen Lernens Algorithmen, die Daten gemeinsam ordnen über ähnliche Maßnahmen, ohne Hilfe Etiketten, Klassen oder Kategorien.

k-Means: Die n Datenpunkte im Datensatz sind auf Basis der kürzeste Entfernung von den Cluster-Prototypen. Der Cluster-Prototyp wird als Durchschnitt genommen Datenpunkt im Cluster.



DBSCAN: Eine nichtparametrische Dichtebasis Clustering-Algorithmen. Datenpunkte werden klassifiziert als Kern-, Dichte- und Ausreißerpunkte. Kern- und Dichteprüfpunkte in hohen Dichtebereiche werden zusammengefasst, während Punkte ohne nahe Nachbarn in niedriger Dichte Regionen werden als Ausreißer bezeichnet.



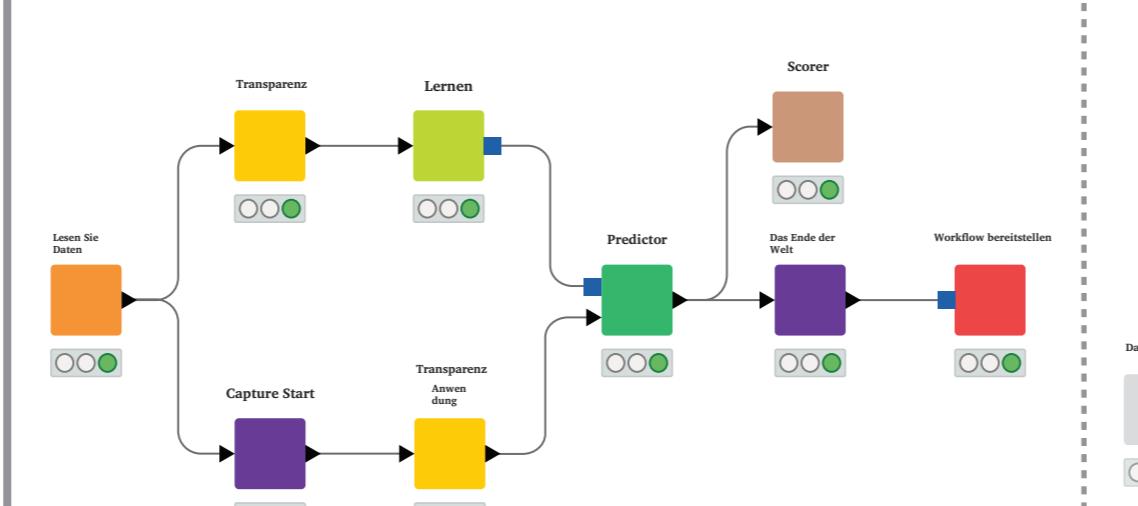
Hierarchisches Clustering: Baut eine Hierarchie von Cluster, indem sie entweder die ähnlichsten sammeln (agglomerativer Ansatz) oder Trennung der am meisten dissimilaren (divisive approach) Daten Punkte und Cluster, entsprechend einer ausgewählten Abstandsmaß. Das Ergebnis ist ein Dendrogramm Zusammenfassen der Daten (agglomerierend) oder Datentrennung in verschiedene Cluster top-down (divisiv).

Selbstorganisierender Baumbasiert Algorithm (SOTA): Ein besonderes Karte zur Selbstorganisation (SOM) neuronales Netz. Seine Zelle Struktur wird mit einer binäre Baumtopologie.

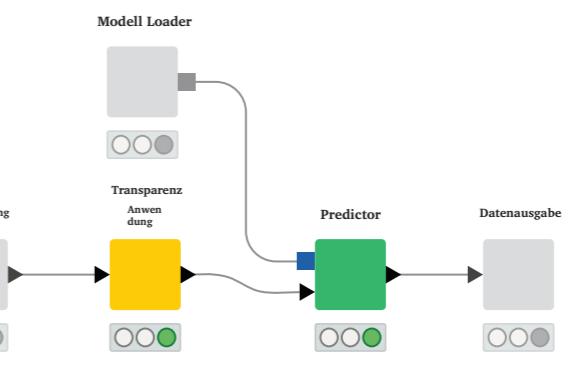
Fuzzy c-Means: Einer der am meisten verbreitet fuzzy Clustering Algorithmen. Es funktioniert ähnlich wie die k-Means Algorithmus, aber es erlaubt Daten Punkte zu mehr als ein Cluster, mit verschiedenen Grad der Mitgliedschaft.

ENTWICKLUNG

Ausbildung



Ausbildung



Ressourcen

E-Books: KNIME Advanced Luck Abdeckungen erweiterte Funktionen und mehr. Praxisdaten Wissenschaft ist eine Sammlung von Datenwissenschaft Fall Studien aus früheren Projekten. Beide erhältlich bei knime.com/knimepress

KNIME Blog: Themen, Herausforderungen, Branchennachrichten, & Wissen nuggets bei knime.com/blog

E-Learning Kurse: Nehmen Sie unsere kostenlose online selbstbefriedigte Kurse, um über die verschiedenen Schritte in einem Data Science-Projekt (mit Übungen & Lösungen zum Testen Ihres Wissens) bei knime.com/knime-self-paced-gänge

KNIME Community Hub: Durchsuchen und teilen Workflows, Knoten und Komponenten. Hinzufügen Bewertungen oder Kommentare zu anderen Workflows bei wohnzimmer.de

KNIME Forum: Schließen Sie sich unserer globalen Community an in Gesprächen bei Forum.de

KNIME Business Hub: Für Teams Zusammenarbeit, Automatisierung, Management, & Einsatz Check-in KNIME Business Hub bei knime.com/knime-business-hub

Empfehlung Motoren: Eine Reihe von Algorithmen, die bekannte Informationen über Benutzerstellungen, um Elemente von Interesse vorherzusagen.

Vereinsregeln: Der Knoten zeigt Regelmäßigkeiten in Co-occurrence Rezensionen von mehreren Produkten in umfangreiche Transaktionsdaten an Verkaufspunkten aufgezeichnet. Basierend auf dem a-priori-Algorithmus, die häufigsten Artikel in der Datensatz verwendet wird, um Empfehlungsbestimmungen.

Collaborative Filtering: Basierend auf die alternierenden Least Squares (ALS) Technik, produziert Empfehlungen (Filtern) über die Interessen eines Nutzers durch Vergleich der aktuellen Vorlieben mit denen von mehrere Benutzer (zusammenarbeiten).

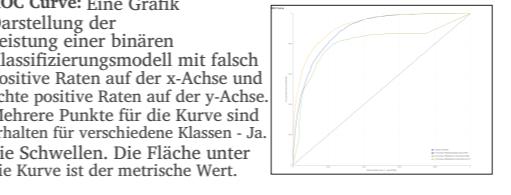
Evaluierung

Bewertung: Verschiedene Scoring Metriken zur Beurteilung der Modellqualität - insbesondere die Vorhersagefähigkeit oder Fehlerwahrscheinlichkeit eines Modells.

Konfusionsmatrix: Darstellung einer Klassifikationsaufgabe Erfolg durch die Anzahl der Spiele und Fehlanpassungen zwischen den tatsächlichen und vorhergesagten Klassen, aka wahre positive, falsche Negative, falsche Positive und wahre Negative. Ein Klasse wird willkürlich als positive Klasse gewählt.



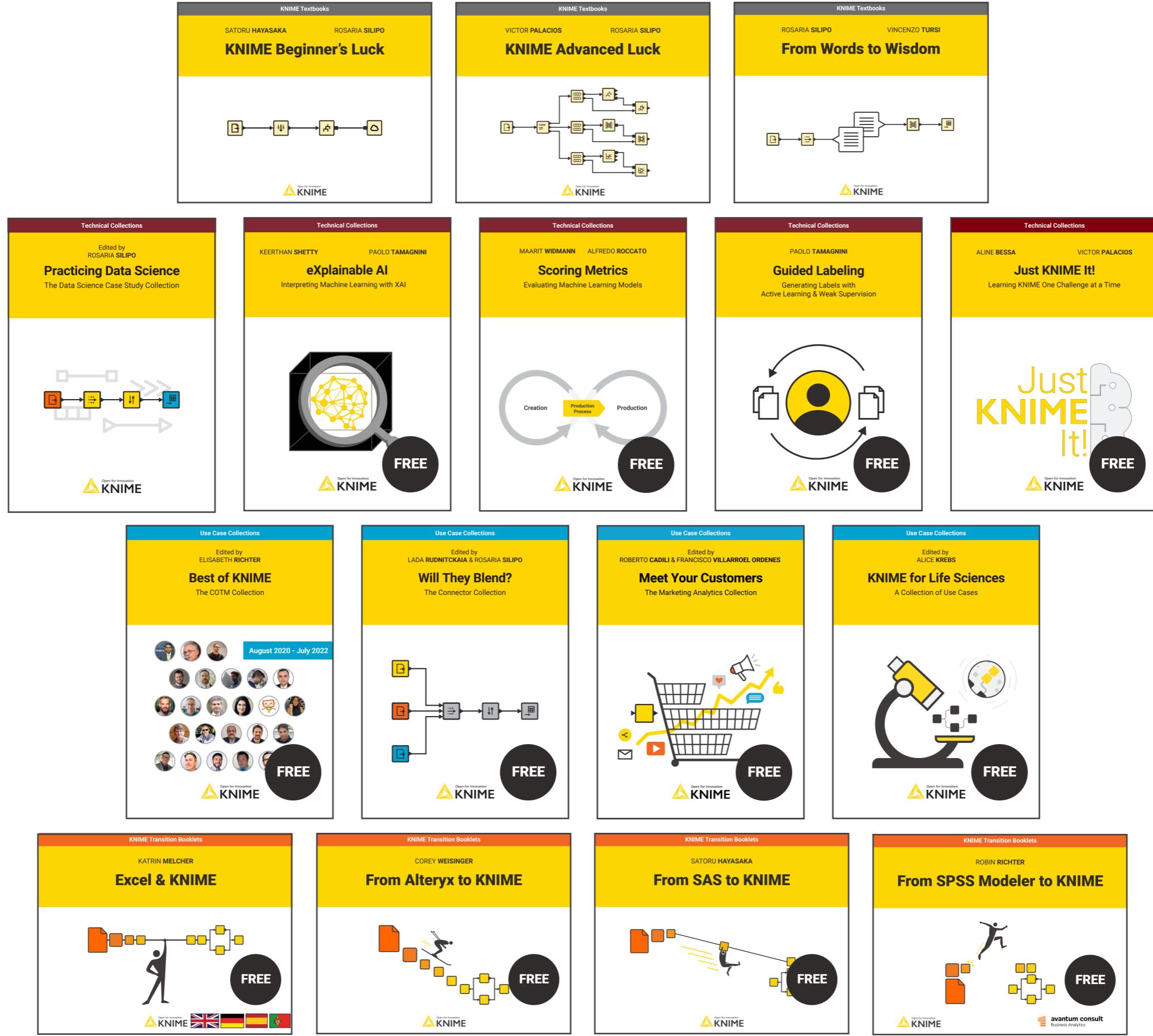
Numerische Fehlermessungen: Bewertungsmetriken für numerische Vorhersagemodelle Fehlergröße und Richtung. Gemeinsame Metriken RMSE, MAE oder R². Die meisten dieser Metriken hängen vom Bereich des Ziels ab variabel.



ROC Curve: Eine Grafik Darstellung der Leistung einer binären Klassifizierungsmodell mit falsch positive Raten auf der x-Achse und echte positive Raten auf der y-Achse. Modell berechnet aus den Werten in der Verwirrungsmatrix, so als Empfindlichkeit und Spezifität, Präzision und Rückruf oder insgesamt Genauigkeit.

Cross-Validation: Eine Modellvalidierungstechnik für Bewertung der Ergebnisse eines maschinellen Lernens Modell wird zu einem unabhängigen Datensatz verallgemeiner. A Modell ist geschult und validiert N mal auf verschiedenen Paaren von Trainingsset und Testset, beide aus den ursprünglichen Datensatz. Einige grundlegende Statistiken über die Ergebnis N Fehler- oder Genauigkeitsmaßnahmen geben Einblicke bei Überbelogung und Verallgemeinerung.

Erweitern Sie Ihr KNIME Wissen mit unserer Sammlung von Büchern von KNIME Press. Für Anfänger und Fortgeschrittene, bis hin zu denen, die an Fachthemen wie Themenerkennung, Datenvermischung und Klassik interessiert sind Lösungen für gängige Anwendungsfälle mit der KNIME Analytics Platform - es gibt für jeden etwas. Verfügbar unter www.knime.com/knimepress.



Brauchen Sie Hilfe?
Kontaktieren Sie uns!

