Fragen und Antworten zu KI – Forumbeiträge

Einsatz eines Bayes-Netzes in medizinischer Forschung
Hallo,

heute in einem Vortrag darauf gestoßen:
Preoperative risk stratification in endometrial cancer (ENDORISK) by a Bayesian network model: A development and validation study - PubMed (nih.gov)

Viele Grüße,
Mareike

Frage VL Woche 5
Halllo Mareike,

ich habe eine Frage zu deiner VL zu den Markov-Ketten, speziell zu dem Beispiel mit den Impfzuständen. Es ist ja nur die Rede von einem generellen Zustand in der Zukunft. Bei dem Beispiel macht es ja einen Unterschied, ob die Zukunft in einem Monat liegt oder in 5 Jahren. Dementsprechend wären auch die Wahrscheinlichkeiten von der Zeit abhängig (Bsp. Übergang von Zustand geimpft zu ungeimpft). Gibt es da eine Möglichkeit das in dem System zu berücksichtigen?

Und noch ein Frage zu der Darstellungsweise: Ist es bloß der Übersichtlichkeit und dem Beispiel geschuldet, dass es keinen Zustandsübergang von geimpft zu geimpft gibt, bzw. keinen direkten Übergang von ungeimpft zu geboostert, oder sind diese Wege genrell nicht zulässig?

Dankeschön und viele Grüße

Tabea

Hallo Tabea,

Die Zeitspanne zwischen den Zuständen wird in diesem Modell nicht mit einbezogen.
Es ist schon richtig, dass die Übergangswahrscheinlichkeiten hier in diesem Beispiel in der Realität nicht 100% von der Zeit unabhängig wären.
Dementsprechend würden je nach Zeitspanne dann auch andere Übergangswahrscheinlichkeiten vorliegen.
Diese Übergangswahrscheinlichkeiten würden dann auch für mehrere Schritte (mit der gleichen Zeitspanne) als gleich angenommen.
Es geht auch hier um eine Annäherung an die Realität. Die Realität zu 100% in einem Modell abzubilden ist generell schwierig.

Zu Deiner zweiten Frage: Ja. Das Beispiel soll explizit zeigen, dass nicht immer alle Übergänge existieren müssen.

Viele Grüße,
Mareike

Frage zu balancierten und unbalancierten Daten
Hallo,

ich wollte fragen, ob es eine Daumenregel gibt, wie groß das Verhältnis etwa sein muss, um zu entscheiden, ob man
unbalancierten Daten hat oder nicht?

Klar, ist ein Datensatz mit 2 Klassen und einer Prävalenz von 0,5 balanciert und ein Datensatz mit einer Prävalenz von

0.01 wäre extrem unbalanciert, weil gerade 1 % des Datensatzes Actual Positives sind. Aber wie sieht es mit dem Raum dazwischen aus?
Es kann doch nicht sein, dass 0.49 bereits unbalanciert ist und einen Einfluss auf die Ergebnisse der Metriken und ROC-Kurve hat.
Ab wann fängt ein unbalancierter Datensatz also an, herkömmliche Metriken wie Accuracy und die ROC-Kurve zu beeinflussen und wann würde ein F1-Score
oder eine Precision-Recall-Kurve eine bessere Aussagekraft haben?

MfG Sergej Ruff

Hallo Sergej,

formal definierte klare Grenzen gibt es da nicht.
Oft wird eine Imbalance mit 20-40% der Minority Klassen als mild angesehen. Bei der Wahl bzw. Interpretation der Metrik sollte man das hier aber schon beachten.
Die Wahl der Metrik hängt auch immer von der Problemstellung ab. Wenn die korrekte Klassifizierung der einen Klasse wichtiger ist, als die der anderen, dann kann man das durch die Wahl einer geeigneten Metrik abbilden.
Man kann auch die Modellperformance mit einem "Dummy-Classifier" vergleichen, dann sieht man, wie viel besser man ist, als wenn man einfach die Majority Klasse vorhersagt.
Die ROC Kurve und Precision Recall Kurve sind eher geeignet, um einen einzelnen Classifier zu evaluieren. Schwieriger wird es dann, wenn man mehrere Classifier vergleichen möchte. Da könnte man dann ROC AUC oder PR AUC verwenden (oder F1, F2, etc).
Wie man mit Klassenungleichgewichten umgeht (neben der Wahl der Metrik auch Upsampling, Downsampling, Wahl des Algorithmus), hängt aber nicht nur von der relativen und absoluten Verteilung der Klassen ab. Wenn die Minority Klasse auch in absoluten Zahlen ausreichend viele Fälle hat, kann man damit oft noch recht gut lernen. Wenn die Majority Klasse sehr viel Varianz enthält, kann das natürlich durch Donwsampling verloren gehen.
Es gibt wirklich sehr viele verschiedene Metriken die für die verschiedensten Probleme entwickelt wurden. Es wird auch immer noch an neuen geforscht. Da sollte man also die verschiedenen Eigenschaften des Problems (also nicht nur Imbalance, sondern auch Folgen von Fehlklassifikationen etc..) mitbedenken. Auch wenn in dem ersten Link unten eine kleine Grafik zur Auswahl der Metrik mit bei ist, sollte man sich immer gut überlegen, wie die Ergebnisse der gewählten Metrik dann zu interpretieren sind und was sie bedeuten und eben vielleicht auch nicht aussagen können.

https://machinelearningmastery.com/tour-of-evaluation-metrics-for-imbalanced-classification/
https://datascience.stackexchange.com/questions/90562/how-much-imbalance-in-a-training-set-is-a-problem
https://datascience.stackexchange.com/questions/810/should-i-go-for-a-balanced-dataset-or-a-representative-dataset/8628#8628
https://datascience.stackexchange.com/questions/11788/when-should-we-consider-a-dataset-as-imbalanced

Liebe Grüße
Sarah

Frage zu Bayes 2
Hallo,

ich habe mir die Bayes/ Markov Vl nochmal angeschaut und wollte fragen, ob ich es richtig verstanden habe.

Im Internet steht oft folgendes zu Bayes und Markov (von Google entnommen):

"A Bayesian network is a directed graphical model. (A Markov random field is a undirected graphical model.)"

Alternativ wird oft behauptet, dass Bayes Netzwerke nicht zyklisch sind und Markov Graphen sind zyklisch.

Ich wollte fragen, was genau damit gemeint ist.

Was ich mir darunter vorstelle, ist folgendes:

In Bayes kann ein Status von Status A zu B wechseln, aber nicht von B zu A zurück.
In Markov kann es hingegen von A zu B und von B zurück zu A wechseln.

Um ein Beispiel zu geben.

Wenn ich von Symptomen auf eine Krankheit schließen will, dann würde ich Bayes nutzen, da ich eine gerichtete Beziehung
habe: Ich will von Symptomen auf Krankheit schließen und möchte nicht von Krankheit zurück auf Symptome schließen.
ich würde Markov also nicht nutzen, weil es hier auch zurück aufs Symptom schließen kann.

Habe ich das richtig verstanden? Ich versuche nämlich gerade zu verstehen, wann ich Bayes Netze und wann markov in der
Praxis anwenden würde.

MfG Sergej Ruff

Hallo Sergej,

bis auf das Beispiel ist Deine Zusammenfassung soweit richtig.
Dein Beispiel passt eher zu einem Hidden-Markov-Modell, da Krankheit einem (verstecktem) Zustand entspricht und das Symptom die beobachtbare Auswirkung des Zustandes ist.

Generell wird Bayes benutzt, wenn es um azyklisch gerichtete Graphen geht.
Markov wird dagegen bei zyklisch gerichteten Graphen verwendet.

Viele Grüße,
Sarah und Mareike

Frage zu Features im breast cancer data set
Hallo,

wir fragen uns ob es noch Tipps dazu gibt, welche features am relevantesten sind? Wir haben uns die pairplots angeschaut und es scheint ja mehr als zwei features zu geben, die B und M diskriminieren könnten. Wie könnte man dann mehr als zwei features plotten um dann noch die decision boundaries zu zeigen?

Danke!
Mattea

Frage zu Features im breast cancer data set
Hallo Mattea,
super, dass ihr euch die Pairplots angeguckt habt. Ein Problem beim Plotten ist immer, dass wir nur einen 2D-Bildschirm haben. Um 3 Dimensionen sinnvoll darzustellen, sind entweder mehrere Ansichten oder noch besser Animationen notwendig (hoher Aufwand!). Bei 4 oder mehr Dimensionen ist es nicht wirklich möglich. Daher betrachtet man idR die Gesamtheit der Correlationplots und versucht dort Schlüsse zu ziehen. Ihr könntet hier auch Featureengineeren versuchen und Linearkombinationen der interessanten Feature bilden (zB mittels PCA), wenn es um die Abbildbarkeit in 2D geht. Ansonten kann man sich ein sinnvolles Featureset durch Trail-and-Error erarbeiten, indem man verschiedene Kombinationen ausprobiert und die Leistung des stetig Modells evaluiert. Den Prozess könnte man auch analog zum Hyperparametertuning automatisieren.

Liebe Grüße
Dominik

Frage zu GGM

Hallo,

ich habe eine Frage, das nicht unbedingt etwas mit unserem jetzigen Thema zu tun hat.

Ich soll nämlich an der Tiho am 8.11 eine kurze Präsentation zum Thema Gaussian Graph Models halten und zur Covariance Shrinkage.

ich verstehe die Idee hinter beiden, aber ich versuche gerade tiefer in die Mathematik einzutauchen und wollte fragen, ob es gute Quellen gibt,
die die Mathematik leicht erklären kann.

Ich versuche zum Beispiel herauszufinden, wieso man Maximum Likelyhood in GGM nutzt oder wieso eine Covarianzmatrix mit p>n nicht invertierbar ist
und wie Covarianzshrinkage genau bei dem Problem hilft.

MfG Sergej Ruff

Hi Sergej,
leider zu spät, sorry, aber ich finde diese VL gut für das Thema: https://www.youtube.com/watch?v=V6NMDZB6LI4

Liebe Grüße
Dominik

Frage zu Bayes
Hallo,

ich habe noch eine Frage zu Bayes.

From the definition of the conditional distribution, we immediately see that

$$P(\alpha \cap \beta) = P(\alpha)P(\beta \mid \alpha).$$

Ich habe folgendes in einem Buch gefunden, das ich gerade lese.

Verstehe ich es richtig, dass der Zähler in der Bayes Regel damit im Grunde der Intercept zweier Variablen wäre.
Der Zähler sagt damit also, wie wahrscheinlich es ist, dass beide Events simultan passieren.
Für P(A/B) würde man dann den Intercept durch die Wahrscheinlichkeit für B (P(B)) teilen. Würde ein P(A/B)= 0 auch auf Conditional Independence deuten?
Wenn P(B) groß und der Intercept klein ist, dann sollte doch der Einfluss der Wahrscheinlichkeit für P(B) (die Variable allein) zunehmen und wenn der Zähler = 0 ist, dann würde es auch
auf eine Indepence hindeuten oder kann man diese Annahme aus irgendeinem Grudn nicht machen?

Weitere Frage: Einfach nur um sicher zugehen, eine Sache, die ich fragen wollte, einfach nur um sicher zu sein, ist wann ich Bayes und wann ich Markov für meine Modelle nutze?
Ich wollte wissen, wenn ich ein Szenario vor mir habe und ich muss entscheiden, was von beiden für eine Modellierung besser geeignet wäre, wie ich es besser auseinander halten würde.
Ich schätze mal, Markov ist gut geeignet, wenn die vergangenen Ereignisse egal sind, aber wann würde ich Bayes nutzen?
Ich wollte es nochmal nachfragen, bevor ich mir etwas falschen zusammenreime.

MfG Sergej Ruff

Keine Antwort erhalten, aber später in einem Buch herausgefunden, dass ich es richtig verstanden habe.

Frage zu Indikatorvariablen
Hallo,

in der Vorlesung haben wir die Vor- und Nachteile von Indikatorvariablen behandelt.

Zu dem Nachteilen habe ich folgendes aufgeschrieben:

- Selbst bei MCAR hat man ein Biasrisiko. In der klassischen statistischen Datenmodellierung, wo man häufig den Zusammenhang zwischen Variable X und dem Outcome darstellen will, sollte man Indikatorvariablen nicht nutzen. Hier wären Indikatoren schlechter als listwises löschen.

Ich habe aber gerade Probleme zu verstehen, wieso Listwises Löschen besser ist und wieso Indikatorvariablen für klassische Modellierung ungeeignet ist.

MfG Sergej Ruff

Hallo Sergej,

wenn man Indikatorvariablen nutzt, ersetzt man die fehlenden Werte durch einen konstanten Wert. Dadurch wird dann die wahre Variabilität der Variablen unterschätzt. Besonders problematisch ist das, wenn die Variable mit fehlenden Werten mit einer anderen korreliert. Dadurch gehen dann ggf. Informationen der Korrelation verloren.

Zu Indikatorvariablen, und wann man diese eben doch verwenden kann, finde ich diesen Link ganz gut:
https://statisticalhorizons.com/is-dummy-variable-adjustment-ever-good-for-missing-data/
"Unfortunately, as Jones showed, the coefficient estimates tend to be biased. Without going into the details of his proof, the intuition for this bias is pretty clear. If you substitute any constant value for the missing values of X, you’re reducing the true variability in that variable. If X and Z are correlated, that will lead to bias in the coefficient of Z. And that further induces bias in the coefficient of X."

Wie man tatsächlich mit den fehlenden Werten umgeht ist leider immer sehr abhängig von den Daten. Je nachdem, wie die vorhandenen Werte verteilt sind und welche Korrelationen die jeweilige Spalte mit anderen Spalten hat, kann es ggf notwendig sein, die Zeilen/Samples einfach zu löschen, z. B. weil man sonst "unnatürliche" Peaks an Werten einführen würde und Korrelationen beeinflusst/verändert. Oft kann man aber diese Korrelationen und das Wissen über die Verteilungen ausnutzen und die fehlenden Werte imputieren. Besonders, wenn man nicht sehr viele Daten hat und möglichst viele Samples behalten will. Wichtig ist es in jedem Fall, darüber nachzudenken, welche Auswirkungen die jeweilige Methode auf das Ergebnis haben kann.

Liebe Grüße
Sarah

Frage zu MICE
Hallo,

ich wollte fragen, ob ich MICE richtig verstanden habe. Vor allem, ob die Variable mit den wenigsten Fehlern die anfängliche Imputation verliert und dann als
response variable benutzt wird.

Hier sind meine Zusammenfassungen zum Thema Mice. habe ich irgendetwas falsch verstanden:

Multiple Imputaton by Chained Equations (MICE) funktioniert bei willkürlichen Mustern. Es handelt sich also um Multiple Imputation durch aneinander gekettete Gleichungen.
Hier soll die Methode erklärt werden (aber nicht klausurrelevant). MICE (Multiple Imputation by Chained Equations) ist ein Verfahren zur Schätzung von fehlenden Daten in einem Datensatz. Der Prozess beginnt damit, dass fehlende Werte einer Variable zufällig aus den vorhandenen Werten derselben Variable ersetzt werden. Wenn also in einer Beobachtung Daten fehlen, werden diese durch zufällig aus den vorhandenen

Daten ausgewählte Werte ersetzt. Dies stellt sicher, dass die Imputation realistische Werte verwendet, die bereits im Datensatz vorhanden sind.
In einem weiteren Schritt wird eine Variable ausgewählt, die im Vergleich zu den anderen Variablen die geringste Anzahl von fehlenden Werten aufweist. Diese ausgewählte Variable wird dann in Beziehung zu den anderen Variablen mit fehlenden Werten gesetzt und mithilfe einer Regressionsanalyse modelliert. Das Hauptziel dieses Schritts besteht darin, die Beziehungen zwischen den fehlenden und vorhandenen Variablen zu erfassen und die verfügbaren Daten bestmöglich zu nutzen.
Die Grundidee hinter MICE ist, dass die verbleibenden fehlenden Werte in einer Variable wahrscheinlich von den Werten anderer Variablen beeinflusst werden. Durch die Regression können diese Abhängigkeiten erfasst werden, was zu genaueren Schätzungen der fehlenden Werte führt, basierend auf den Beziehungen zwischen den Variablen.
Es ist wichtig zu beachten, dass die Regression auf Beobachtungen beschränkt ist, bei denen keine fehlenden Werte in der Variable mit den wenigsten fehlenden Daten vorhanden sind. Wir nehmen also die Variable mit den wenigsten Fehlern und wir entfernen die zufällig eingesetzten Werte aus Schritt 1 wieder.
Unsere beobachtete Variable hat wieder Missing Data. Die anderen Variablen werden jetzt zur Feature-Matrix, während die Variable mit den wenigsten Fehlern zum Target/ Response wird. Ich sage durch die Regressionsanalyse voraus, welcher Wert in die beobachtete Spalte eingesetzt werden muss. Ich muss dafür aber im ersten Schritt zufällige Zahlen einsetzen, weil ansonsten die Feature-Matrizen keine Werte für die Vorhersage haben würden (man muss die NA-Werte für die Regressionsanalyse entfernen). Dieser Prozess wird dann für alle Variablen im Datensatz durchgeführt.Wenn man dann eine Variable – nennen wir Sie x1 – durch die Regression mit neuen Werten imputiert hat, macht man mit der nächsten Variable x2 weiter. Für x2 macht man wieder eine Regressionsanalyse, wobei man diese auf alle anderen Variablen regressiert – einschließlich der imputierten Werte aus x1. Man nutzt bei x2 aber nur die Beobachtungen, die nicht fehlen. Das führt man für alle Variablen durch. Wenn man alle Variablen regressiert hat gegen alle anderen Variablen, dann hat man einen Zyklus.
Diese Zyklen von Imputation und Regression werden iterativ wiederholt. Jede Iteration erzeugt unterschiedliche Ergebnisse aufgrund der zufälligen Auswahl von Werten und der Verwendung der zuvor imputierten Daten in den Regressionen. Mit zunehmenden Iterationen konvergieren die Imputationen schließlich zu stabilen Werten.
Man macht es also so, dass man die vorhandenen Daten jeweils nutzt, um die fehlenden Daten zu modellieren. Dann werden in mehreren Iterationen auch die imputierten Daten genutzt, um die anderen Variablen vorherzusagen.

MfG Sergej Ruff

Hallo Sergej,

sorry, für die späte Antwort. Das ist irgendwie untergegangen.

Ich finde, Du hast das sehr gut erklärt. Mir sind keine Missverständnisse aufgefallen. Ich habe das auch vor einigen Jahren während meiner Doktorarbeit mal zusammengefasst und jeweils mit Quellen versehen. Vielleicht hilft das noch zusätzlich, auch weil ich da noch ein paar weitere methodische Detailaspekte behandelt habe. Hier mein Text von damals:

MI by chained equations employ an iterative algorithm to fill the missing values. This algorithm was described by White et al as follows [1]. First, all missing values are replaced by random draws with replacement from the observed values in that variable. Then the variable with the smallest number of missing values of all variables that have missing values, here called x1, is regressed on all other variables. This regression is restricted to observations that have no missing values in x1. The missing values in x1 are then replaced by simulated draws from the posterior predictive distribution of x1 derived from the regression.

Thereafter, the next variable, here called x2, is regressed on all other variables, including the imputed values in x1 and restricted to observations with no missing values in x2. Again, the missing values in x2 are replaced by simulated draws from the posterior predictive distribution of x2 derived from the regression. This is continued until all variables with missing values have been regressed on the other variables, which constitutes a full cycle. These cycles are iteratively repeated. Each iteration yields different results because of the draws from the posterior predictive distribution and because the chained regression models use the newly generated imputations from the previous cycle as independent variable. Therefore, the iterative process should at least be continued until a

convergence of the imputed values has been reached. After a predefined number of iterations have been carried out, a single imputation is achieved and the process is repeated several times to generate multiple imputations.

Several methodological decisions are required to configure the process of MI by chained equations. These are a decision for the variables included in the imputation models, a decision for the number of imputations and a decision for the number of iterations in each imputation cycle. Van Buuren et al provided a four-step advice for the decision of which variables to include in the imputation models in case of many potential covariates [2]. First, the imputation models should include all variables that will be included in the analyses models. This includes the dependent variable of the analyses models, because its exclusion could lead to biased imputation [3,4]. Second, the imputation models should include all variables that are related to the missingness of other variables. These variables should be identified by analysing the correlation between the potential variables and an indicator variable that represents the missingness of the other variables. Third, the imputation model should include all variables that explain a considerable amount of variance in the observed values of variables that contain missing data. Fourth, all variables should be removed from the imputation model that are not part of the analyses model and that have a large number of missing values themselves.

The number of imputations is another relevant decision in configuring the process of MI by chained equations. In the past, a relatively small number of three imputations were considered to be sufficient [5]. Maybe due to increased computational resources, the recommended number of imputations has become much higher. Graham et al analysed the number of imputations in relation to the fraction of missing data, the derived standard errors and the power falloff in analyses models [6]. Their results showed that 20 imputations should be sufficient for a fraction of missing data of up to 30%.

The number of iterations at each imputation cycle is another relevant decision in configuring the process of MI by chained equations. It is important because the values imputed at each iteration need to converge before the end of the imputation cycle. Van Buuren et al suggested that 5 - 20 iterations should usually be sufficient. Furthermore, they suggested to use so-called trace plots to follow the convergence of imputed values and to adjust the number of iterations, if necessary [7].

All methods of MI require the data to be MAR [8]. MAR depends on the true values of the missing data. Therefore, it ultimately remains an assumption [9]. However, MAR is not an assumption inherited in the dataset but in the imputation models, since data are MAR if the missingness can be controlled for by other variables, as explained at the beginning of this section. Therefore, Sterne et al recommend inclusion of as many variables as possible in the imputation models to meet the MAR assumption [10]. Janssen et al pointed out that it should usually be possible to meet the MAR assumption and reliably impute missing values in data with many covariates, in his case patient characteristics, because the probability of missingness should mainly be controlled by the covariates [11]. Furthermore, they stated that even in cases where missing values are not precisely MAR, MI still tends to perform better than other missing data methods and that in many real-world cases the failure to meet the MAR assumption would have only a minor impact on results.

1 White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Statist Med. 2011;30:377–99.

2 van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. Stat Med. 1999;18:681–94.

3 von Hippel PT. Regression with missing Ys: An improved strategy for analysing multiply imputed data. Forthcoming from Sociological Methodology. 2007;1–54.

4 Moons KGM, Donders RART, Stijnen T, et al. Using the outcome for imputation of missing predictor values was preferred. Journal of Clinical Epidemiology. 2006;59:1092–101.

5 Rubin DB, Schenker N. Multiple imputation in health-are databases: An overview and some applications. Statist Med. 1991;10:585–98.

6 Graham JW, Olchowski AE, Gilreath TD. How Many Imputations are Really Needed? Some Practical Clarifications of Multiple Imputation Theory. Prev Sci. 2007;8:206–13.

7 van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. Stat Methods Med Res. 2007;16:219–42.

8 Allison PD. Multiple imputation for missing data - A cautionary tale. Sociological Methods & Research. 2000;28:301–9.

9 Baraldi AN, Enders CK. An introduction to modern missing data analyses. Journal of School Psychology. 2010;48:5–37.

10 Sterne JAC, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009;338:b2393.

11 Janssen KJM, Donders ART, Harrell Jr. FE, et al. Missing covariate data in medical research: To impute is better than to ignore. Journal of Clinical Epidemiology. 2010;63:721–7.

Frage zu SVM
Hallo,

ich weiß, dass man SVm nutzen kann, um Klassifikationen zu machen.

Ich glaube ind er Vorlesung zu SVm haben wir auch gelernt, dass Daten linear separierbar ode r nicht linear separierbar sein können. Das kann über die Performance
des Modells entscheiden.
Ich wollte jetzt wissen, ob man SVM auch nutzen kann, um Daten auf lineare separierbarkeit zu testen, bevor man ein anderes Modell nutzt.
Ich habe nämlich ein Problem damit zu verstehen, wie man sonst sehr hoch dimensionale Daten (sagen wir mal 100 + features/Dimensionen) auf lineare separierbarkeit testen kann.
Einfach plotten und Sie anhand von Graphiken zu vergleichen wird sicherlich schwer bei sehr hohen Dimensionszahlen.

MfG Sergej Ruff

Hi Sergej,
ja, das ginge. Ist aber Zeitintensiv. Hier noch eine übersicht anderer Methoden:
https://www.tarekatwan.com/index.php/2017/12/methods-for-testing-linear-separability-in-python/

LG
Dominik

Frage zu XAI
Hallo,

mir sind ein paar Fragen zum Thema XAI eingefallen.

1. Wie stabil ist LIME? Ich muss doch Permutationen nutzen. Sind diese permutationen nicht random generiert? Sollte ich deshalb nicht jedesmal verschiedene Werte
für die Erklärung bekommen bei LIME, auch wenn das Modell bei jedem Run gleich performed? Wenn ja: Wie gehe ich Sicher, dass Ergebnisse reproduzierbar sind
(für z.B. eine Veröffentlichung meiner Ergebnisse)?

2. Gibt es Beispiele, wo das Modell nicht lokal durch ein lineares Modell aproximiert werden kann und deshalb nicht durch LIME erklärt werden kann? Nutzt man in diesen
Fällen SHAP oder gibt es Möglichkeiten dann doch LIME zu nutzen?

3. Zu welchen Zeitpunkt im Workflow benutzt man XAI? Erst, wenn man mit einem ersten Modell zufrieden ist oder nutzt man es mit jeder Iteration des Modells als einen "Debugger",

um zu verstehen, was am Modell verändert werden kann?

4. Mir ist nicht ganz klar, welche Aussagen SHAP und LIME jeweils machen oder eher, worin sie sich unterscheiden. LIME - so wie ich es verstanden habe - kann Erklärungen für einen
Datenpunkt machen und dann nur für diesen einen Datenpunkt. Wenn ich also Patienten A und B anschaue mit Variablen X,Y und Z und beide Patienten werden im Rahmen einer Klassifikation
in Klasse 1 zugeordnet, dann kann es sein, dass Variable X die Zuweisen für patient A in Gruppe 1 erklärt. LIME ist aber Datensatzspezifisch, so wie ich es verstanden habe. variable X muss also
nicht unbedingt die Erklärung für Patient B liefern. patient b kann auch wegen variable Y in Gruppe 1 eingeordnet werden sein. ich muss also LIME für jeden patienten einzeln laufen lassen.
SHAP hingegen macht Aussagen, die global gültig sind, weil er lokale Fälle aggregiert? Also wäre Variable X für Patient 1 und 2 als Erklärung, wieso beide in Gruppe 1 eingeteilt wurden, gültig.
SHAp würde demnach Muster in den features erkennen, die eine generelle Zuordnung in eine Gruppe erklären würden?
ist das richtig oder ist SHAP genau auf die gleiche Weise wie LIME nur für einzelne Datensätze gültig?
LIME habe ich glaube ich gut verstanden. Die Frage bezieht sich eher darauf, dass ich die Ergebnisse von SHAP nicht ganz verstehe.

MfG Sergej Ruff

Frage zu XAI
Ich würde gerne noch eine weitere Frage zu XAI und LIME ergänzen:

Wenn ich ein lineares Modell generiere, gibt es dann bestimmte Grenzen in denen das gilt oder gilt das dann nur genau an der Stelle meines ersten Datenpunktes?

Dankeschön und viele Grüße
Tabea

Frage zu XAI
Liebe Tabea,
das Model "gilt" in der Nachbarschaft des zu erklärenden Datenpunktes. Die Größe der Nachbarschaft wird durch die Permutationen und deren Abstand zum zu erklärenden Datenpunkt festgelegt. Bei dem Modell handelt es sich aber nur um eine Approximation!

Frage zu XAI
Lieber Sergej,

zu 1. Genau, die Permutationen werden zufällig generiert. Damit kommt eine gewisse Streeuung in die Ergebnisse die erstens durch die Anzahl an verwendeten Permutationen und zweitens durch die Gewichtung des Einflusses jeder einzelnen Permutation über den Abstand zum zu erklärenden Datenpunkt versucht wird zu verringern. Wenn du genug Datenpunkte nimmst, die die Nachbarschaft abbilden, sollte die Streeung der Ergebnisse im Bezug auf die Feature sehr sehr gering sein. Das ganze ist und bleibt aber eine Approximation.

zu 2. Es handelt sich bei LIME immer um eine Approximation, die besser oder auch schlechter sein kann, und keine mathematisch genaue Lösung. Auch die Reduktion auf ein (lokales) lineares Modell bringt einen gewissen Fehler ein, falls die wirkliche Trennlinie der Klassen nicht linear ist. Es ist also ein Näherungsverfahren, wie sie in der Mathematik häufig genutzt werden (zB in der numerischen Integration). Die Alternative hier wäre eine Blackbox (oder ein anderes XAI Verfahren). Dies zeigt aber auch eine Schwäche von LIME auf: es wird nicht das Modell an sich erklärt sondern ein Surrogatmodell genutzt, dass vom echten Modell abweichen kann.

zu 3. Man kann es als Debugger nutzen, aber die Interpretation von Erklärungen von schlecht funktionierenden Modellen ist häufig Kaffeesatzlesen, kann aber auch zu guten Ergebnissen führen. Des Weiteren nimmt man XAI um die Validität des Modells zu überprüfen, wenn man zufieden ist (siehe Urban Legend Beispiel aus der Vorlesung), und zu guter Letzt um die Prädiktion dem Nutzer zugänglicher zu machen und so da Vertrauen zu erhöhen.

zu 4. "LIME - so wie ich es verstanden habe - kann Erklärungen für einen Datenpunkt machen und dann nur für diesen einen Datenpunkt." - Ja

SHAP basiert auf einer spieltheoretischen Annahme, dass man durch Mittlung den Anteil eines Features am Gesamtergebnis bestimmen kann. Es gibt also immer Veränderungen im Bezug auf die durchschnittliche Vorhersage an. Es wird im Endeffekt betrachtet, was mit der Vorhersage passiert, wenn man ein bestimmtes Feature weglassen würde (indem der Wert des Features durch den Durchschnitt dieses Features ersetzt wird). Das kann man für jedes Feature und jeden Datenpunkt machen. Da es sich dabei dann für alle Datenpunkte um den Abstand zum Mittelwert handelt, kann man diese aggregieren und so eine globale Erklärung erhalten, die nicht nur einen einzigen Datenpunkt betrifft. Bei LIME ist das nicht der Fall, da pro Datenpunkt ein neues Surrogatmodell trainiert wird und diese nicht so einfach kombinierbar sind (bzw ihre Linearität bei der Kombination und damit auch die Tauglichkeit als Erklärung verlieren).

Frage zu XAI
Danke für die Antworten.

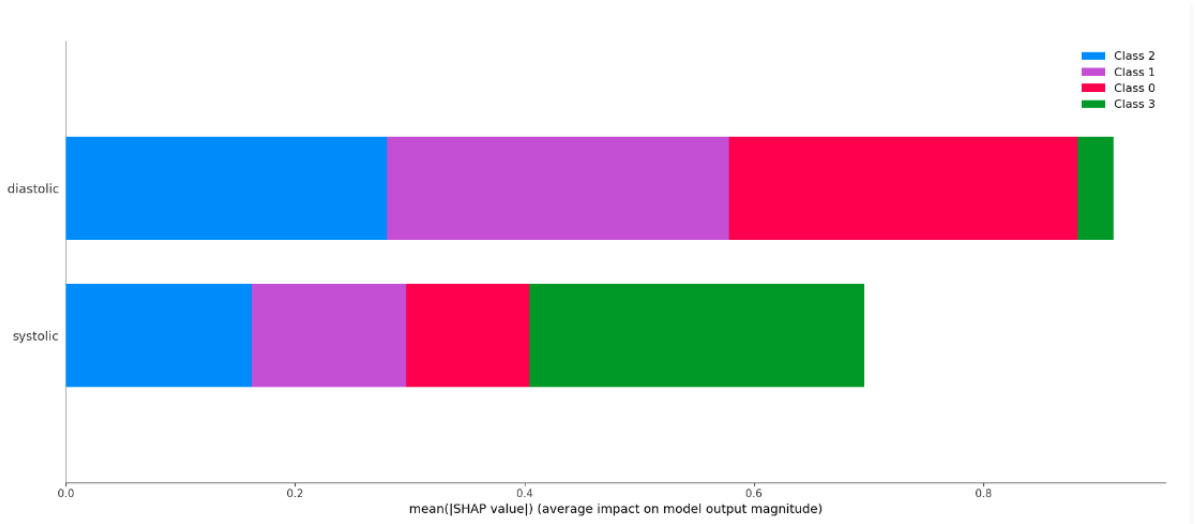
Punkt 4 hat noch keine Antwort. Bei mir steht da nur "zu 4.". Ist das ein Fehler von Ilias?

Zu 1. Kann man in LIME einen Seed setzen. Ich habe überall im Code außer in SHAP und LIME Seeds gesetzt. Ich kriege jedesmal diesen Ergebnisse für SHAP und
LIME, aber die genauen Zahlenwerte unterscheiden sich leicht - ich gehe davon aus , dass es auf die zufällig generierte Permutation zurückzuführen ist.

Frage zu XAI
Können wir vielleicht morgen in der besprechung auch drauf eingehen, wie man die Plots zu interpretieren hatß Das haben wir uns in der Gruppe gefragt.
Vor allem der SHAp Balkenplot. Ich habe 4 Klassen für den Datensatz mit dem Bloodpreassure Datensatz.
Der Grüne Balken (Klasse 3) ist größer als die restlichen Klassen in Systolic. Was hat das zu bedeuten für die Interpretation der Egebnisse?
Wie interpretiere ich, dass der grüne Balken in diastolic viel kleiner ist.

ich habe es so verstanden, dass Diastolic in meinem Datensatz den größten Einfluss auf die Zuteilung in eine Klasse hat, gefolgt von Systolic.
Und wenn man in Klasse 3 eingeteilt wird, dann liegt es wahrscheinlich an den systolischen Werten, weil der Balken für Klasse 3 so viel größer ist. Oder umgekehrt,
wenn man aufgrund der systolischen Werte irgendwo eingeteilt wird, dann in Gruppe 3 und wenn man wegen seiner diastolischen Werte in eine Gruppe eingeteilt wird,
dann wird es wahrscheinlich nicht in Gruppe 3 sein. ist das richtig?

In Lime waren wir nicht sicher, wie man die zahlenwerte auf der X-Achse interpretieren soll und was die roten und grünen Balken zu bedeuten haben, vor allem bei einem
Datensatz mit mehreren Klassen.



Frage zu XAI
zu zu 1. ja, das geht in der init des Explainers über die "randomstate" Variable:
https://github.com/marcotcr/lime/issues/119
https://github.com/marcotcr/lime/blob/master/lime/lime_tabular.py#L117
Genau es liegt an den zufälligen Perutationen, bei der realen Anwendung setzt man den Seed dann aber nicht per Hand.

Hi Sergej,
das können wir uns gerne morgen angucken. Es macht mehr sinn sich einen Plot pro Klasse erstellen zu lassen, da das Stacken der Bars die Wahrnehmung verzerren kann.
Insgesamt ist in deinem Modell der diastolische Wert wichtiger als der Systolische (reicht ja, bei den dinjunkten Klassen idR aus). Bei Klasse 3 ist es dann aber umgekehrt.
Hier nochmal ein Link bzgl der Interpretation:
https://towardsdatascience.com/using-shap-values-to-explain-how-your-machine-learning-model-works-732b3f40e137

Fragen
Hallo,

Wie kann ich denn in einem Plot mit z.B. 6 Graphen einen vorher erzeugte Graphik als subplot einfügen?
Ich habe z.B. einen Barplot erzeugt, den ich jetzt als einen subplot einügen möchte.

Viele Grüße und danke

Christoph

Frage: Unterschied Konkatentation und Parameter Aufgabe A0/A1 S.20
Hi,
beim Nacharbeiten hat sich bei der Aufgabe A0/ A1 aud S,20 bei mir eine Frage ergeben.
Was ist der genaue Unterschied zwischen Konkatentation und mehren Parameter .
Meine Lösung wäre folgende : Konkatention erfolgt über das Pluszeichen --> print("Heute" + " " + "ist" + " " + "schönes" + " " "Wetter"),
mehrere Parameter erfolgt über Kommatasetzung print("Heute" , "ist" , "schönes" ,"Wetter") . Ist dies korrekt ?

Frage: Unterschied Konkatentation und Parameter Aufgabe A0/A1 S.20
Hallo Katharina,
in beiden Fällen wird eine Konkatenation durchgeführt, aber zu unterscheidlichen Zeitpunkten.

print("Heute" + " " + "ist" + " " + "schönes" + " " + "Wetter")
Beim "+" werden alle angegebenen Strings erst zusammengeführt, bevor ein einzelner String ("Heute ist schönes Wetter") an das print übergeben wird.

print("Heute" , "ist" , "schönes" ,"Wetter") übergibt vier Strings an die Print-Funktion, welche dann innerhalb diese Konkatiniert bzw als Junks ausgegeben werden.

Wenn man sehr sehr viele diese Operationen durchführt ist die parametrisierte Variante zeitlich überlegen (= schneller). Das gilt aber nur bei wirklich vielen Strings, die zusammengeführt werden sollen.

Unter folgendem Link nochmal mehr Informationen:
https://stackoverflow.com/questions/52354518/difference-between-and-when-printing-output

Liebe Grüße
Dominik

Hallo Christoph,
das ist meines Wissens nach in Python so nicht möglich. Falls du zwei einzelne Funktionen verwendest, kannst du es zusammenfügen, indem du das Axes-Obejekt des Subplots übergibst
(https://stackoverflow.com/questions/65618035/adding-a-figure-created-in-a-function-to-another-figures-subplot):

```
def my_plot_1(x, y, ax):
    ax.plot(x, y)
def my_plot_2(x, y, ax):
   ax.plot(x, y)
fig, axs = plt.subplots(ncols=2, nrows=1)
# pass the Axes you created above
my_plot_1(x, y, axs[0])
my_plot_2(x, y, axs[1])
```

Ansonsten einfach den Subplot erstellen und dann die einzelnen Figures.

Liebe Grüße
Dominik

Datenbereinigung
Hallo,
ich schaffe es leider heute nicht in die Sprechstunde, daher stelle ich meine Fragen hier mal rein:

Da bei der Datenbereinigung ja tendentiell immer auf die gleichen Dinge geprüft wird, frage ich mich, ob es hier nicht eine Art Standardsammlung von Befehlen gibt, die ich zu Beginn anwenden könnte?
Außerdem tue ich mich noch ein bisschen schwer, mir zu Beginn erstmal einen Überblick über die Daten geben zu lassen und beispielsweise alle verwendeten Antwortmöglichkeiten in einer Spalte anzeigen zu lassen um dort ggf. zu erkennen, wo ich ansetzen kann. Die Möglichkeit mir im Debug-Modus unten den Dataframe anzeigen zu lassen, und dort nach Auffälligkeiten zu suchen, scheint mir nicht sehr effizient :D Gibt es hier noch eine andere Idee?

Dankeschön!

Hallo Tabea,

bitte entschuldige die späte Antwort.
Leider lässt es sich meist nicht vermeiden, sich erstmal einen Überblick über die Daten verschaffen zu müssen. Um Daten korrekt bereinigen und aufbereiten zu können, muss man verstehen, was sie bedeuten und was die Manipulationen, die man darauf anwenden möchte, für Auswirkungen haben könnten.
Allerdings gibt es ein paar Befehle in Python, die einem das Leben etwas leichter machen:
info()
Anzahl nicht-Null Werte und Typen der Spalten
describe()
Statistische Kennzahlen (z. B. min, max, mean bei numerischen Spalten)
value_counts()
Wie oft kommen welche Werte vor?
nunique()
Anzahl einzigartiger Werte
sum()
Summe der Spalte
count()
Anzahl vorhandener Werte
head(), tail()
Die ersten/letzten Zeilen ausgeben
df.columns
Spaltennamen
shape()
Anzahl Zeilen und Spalten

Außerdem gibt es noch das ydata-profiling Package. Das beschreibt einem die Daten ganz gut und findet sogar schon Korrelationen. Allerdings funktioniert es leider nur für relativ kleine Datensätze.

Liebe Grüße

Sarah

Frage Figure speichern
Guten Tag,

ich hätte eine Frage zu meiner Funktion. Ich möchte gerne meine ROC-Kurve speichern. Hierbei werden aber zwei figures geöffnet, zuerst eine leere und dann eine zweite mit meiner ROC-Kurve. Gespeichert wird leider die leere. Wahrscheinlich liegt es daran, dass ich zweimal einen plot aufrufe (display.plot() und plt.plot() und hierbei die Verbdingung fehlt).

```
roc = m.roc_curve(test_set["Churn"], y_probabilities)

fpr, tpr, thresholds = m.roc_curve(test_set["Churn"], y_probabilities)
roc_auc = m.auc(fpr, tpr)
display = m.RocCurveDisplay(fpr=fpr, tpr=tpr, roc_auc=roc_auc, estimator_name='estimator')
display.plot()
plt.plot([0,1], [0,1], "--")
plt.show()

plt.savefig("C:/Users...")
```

Vielen Dank

Christoph

Hallo Christoph,

das Problem liegt in den Aufrufen von show() und savefig(). Die show Funktion ruft immer die Liste mit aktuellen figures auf und arbeitet diese sozusagen ab. Nach dem Aufruf von show ist diese Liste dann eben leer.
Hier sind nochmal etwas mehr Infos dazu: https://stackoverflow.com/questions/21875356/saving-a-figure-after-invoking-pyplot-show-results-in-an-empty-file

Es sollte sich lösen lassen, indem du vor dem Aufruf von show() den Plot zwischenspeicherst mit plt.gcf() und dann auf der figure savefig() aufrufst. Siehe die beiden fetten Zeilen:

```
roc = m.roc_curve(test_set["Churn"], y_probabilities)

fpr, tpr, thresholds = m.roc_curve(test_set["Churn"], y_probabilities)
roc_auc = m.auc(fpr, tpr)
display = m.RocCurveDisplay(fpr=fpr, tpr=tpr, roc_auc=roc_auc, estimator_name='estimator')
display.plot()
plt.plot([0,1], [0,1], "--")
fig = plt.gcf()

plt.show()

fig.savefig("figure.png")
```

Liebe Grüße
Sarah

Fragen zu Feature Selection und Tuning
Hallo,

ich habe ein paar Fragen, die ich mal hier aufschreibe und vielleicht später ergänzen werde:

1.Gehört zur Feature Selection die Eliminierung von unnötigen Featuren oder auch die Senkung ihrer Wirkung? Ich weiß, dass Lasso zur Feature Selection dient (Embedded), weil es unwichtige Features

0 setzen kann. Ich wollte wissen,ob Ridge Regression auch Feature Selection durchführt, da es Feature nicht auf 0 setzen kann. Es kann nur ihre Wirkung/Einfluss senken.

2. Welche Entscheidungskriterien gibt es für Forward/ Backward und Recursive Feature Selection? ich verstehe, wie Sie im Prinzip funktionieren, aber ich frage mich, welche Entscheidungskriterien ich nutzen soll, um zu entscheiden, welche von den 3. ich für ein Problem nutzen soll. Hat eine eine bessere performance oder geringere Rechenzeit als die andere? Ist jeder von ihnen von der Struktur und Form des Datensatzes abhängig und vom Datentyp und Größe? Und welchen Unterschied macht es, ob ich eine Sache vorwärts oder rückwärts durchführe?

3. So wie ich es verstanden habe nutzen wir das Surrogat Model bei Bayes, weil wir die Beziehung zwischen hyperparameter als Inputs und dem Ergebnissen der Zielfunktionen wissen wollen. Wir behandeln die Zielfunktion (Folie 29) als eine Blackbox und wir wissen Input und Output. Um die optimale Hyperparameterkombination zu bestimmen, müsste man für die Zielfunktion das originale Model mehrmals laufen lassen für alle möglichen Kombinationen an hyperparametern. Das ist spätestens bei Deepl Learning nicht möglich, weil bereits ein Durchlauf sehr lange dauert. Deep Learning Modelle durchlaufen zu lassen, wäre also viel zu aufwendig. Deshalb nutzt man Surrogat Modelle, weil man hier das ursprüngliche Modell nicht mehrmals durchlaufen lassen muss?
Oder täusche ich mich und man muss das Modell doch jedesmal laufen lassen?

4. Werden Grid und random Search für komplexe Modelle (Deep learning) genutzt?

MfG Sergej Ruff

Hallo Sergej,

1.
Zur eigentlichen Feature Selection zählen tatsächlich nur Methoden, die Feature auswählen, bzw. eliminieren. Ridge an sich ist eine Form der Regularisierung. Es kann allerdings zur Feature Selection genutzt werden, indem man über die Feature Importance bzw. die Koeffizienten, die weniger einflussreichen Feature entfernt. Das kann man prinzipiell mit allen Modellen machen, bei denen eine Art der Feature Importance erlernt wird.

2.
Recursive Feature Elimination wird zwar recht häufig verwendet, ist aber nicht unbedingt die beste Wahl. Bei RFE werden die Feature, die eliminiert werden, mehr oder weniger univariat ausgewählt, während zum Beispiel bei der Forward Selection nach und nach immer "multivariater" ausgewählt wird. Also, um bei RFE ein Feature zu eliminieren, schaut man sich wirklich nur die Feature Importance der einzelnen Feature an und entfernt das schlechteste. Bei der Forward Selection nimmt man nach und nach immer das Feature hinzu, dass in Kombination mit den bisher ausgewählten Feature die beste Performance hat. Hier wird also multivariater ausgewählt, in dem eben die gemeinsame Performance evaluiert wird.
RFE ist eine Art der Backward Elimination, aber noch etwas mehr "greedy", da in jedem Eliminierungsschritt nur ein Modell trainiert wird, und nicht für jedes der verbleibenden Feature ein Modell eben ohne das jeweilige Feature.
Eine konkrete Regel, wann man welche Methode verwendet gibt es nicht. Das ist immer abhängig von den Daten, der Problemstellung und den Ressourcen.
Bei sehr großen Datensätzen fängt man oft mit einer einfachen Filtermethode an, da diese effizient zu berechnen sind. Man sollte dabei aber nicht zu viele Feature entfernen, da man eben die Interaktionen zwischen den Feature nicht beachtet. Tatsächlich sind Filtermethoden häufig auch stabiler als andere Methoden.
Wenn man erwartet, dass relativ viele Feature am Ende im Modell verbleiben, dann kann RFE im Vergleich zu Forward Selection die schnellere Methode sein. Das heißt aber nicht, dass die resultierenden Feature dann unbedingt besser sind, als die die mit der Forward Selection ausgewählt worden wären.
Im Allgemeinen, sollte man für die jeweilige Problemstellung in der Literatur suchen, welche Methoden dort bereits gut funktioniert haben. Oder, wenn man ausreichend Ressourcen hat, selbst verschiedene Methoden vergleichen.

3.
Das klingt soweit korrekt. Das Surrogatmodell dient dazu, die Anzahl tatsächlicher Trainingsdurchläufe auf dem eigentlichen Modell zu reduzieren, indem nur für vielversprechende Hyperparameterkombinationen tatsächlich das eigentlichen Modell trainiert wird. Für diese aus dem Surrogatmodell ausgewählten Kandidaten muss dann

aber tatsächlich wieder das eigentliche Modell trainiert werden, damit man weitere Informationen über das tatsächliche Verhalten der Zielfunktion erhält, also der Performance der Hyperparameter im Modell, und das Surrogatmodell aktualisieren kann. Das eigentliche Modell muss also dennoch trainiert werden, aber deutlich seltener, da das Surrogatmodell die Hyperparametervektoren so auswählt, dass die hoffentlich eine gute Performance aufweisen.

4.
Beim Deep Learning kann man tatsächlich nochmal unterscheiden zwischen Hyperparametern des Model Designs, also Anzahl Layer, etc., und Hyperparametern des Trainings, also z. B. dem Optimzer, der Learning Rate, etc.
Bei "kleineren" Projekten wird häufig das Babysitting angewendet: Verschiedene Hyperparameter werden manuell getestet.
Tatsächlich wird auch die Grid Search recht häufig für kleinere Suchräume verwendet. Sie kann auch gut parallelisiert werden, da die einzelnen Durchläufe unabhängig von einander sind. Grid Search ist aber eher nur zu empfehlen, wenn man bereits genug Erfahrung hat, um einen guten aber kleinen Suchraum definieren zu können.
Für Random Search sieht es recht ähnlich aus, wobei diese oft tatsächlich bessere Ergebnisse liefert, besonders, wenn Hyperparameter nicht gleichverteilt sind und man weniger Erfahrung hat. Man kann beide Methoden aber auch dazu verwenden, um erst einmal recht grob einen geeigneten Suchraum zu finden, also herauszufinden, welche Parameter man in welchem Bereich vielleicht noch tunen sollte, oder welche eben keinen großen Einfluss haben.
Bei komplexeren Methoden wie der Bayes'schen Optimierung wird in der Regel nicht so viel Vorwissen und Erfahrung über die Hyperparameter benötigt. Sie zählen auch meist zu den state-of-the-art Methoden bzw. Ergebnissen.
Wie immer ist es aber abhängig von der Problemstellung und den Ressourcen.

Liebe Grüße
Sarah

Fragen zu KNN
Hallo,

ich habe ein paar Fragen zum Thema KNN, die man heute in dem Meeting besprechen kann 8schreibe es hier, falls ich vergesse, selbst zu fragen):

Wozu Pooling bei CNN oder besser, was wäre wenn ich es weglassen würde
-> Wozu brauche ich eine Dimensionsreduktion bei CNN

Wäre es besser, wenn man Daten vorher skaliert und auf die selbe Skala bringt. Welchen Einfluss haben unterschiedliche Skalen auf KNN, wenn Sie
Einfluss haben.

Nehmen KNN nur numerische Werte an?
Wenn ja, wie funktionieren Sequenzielel Daten wie Texterkennung und Vorhersagen?
z.b. Wenn ich RNA-Daten mit RNA-Modifikationen als Sequenzieller Input nutzen will, dann habe ich 4 Nukleotide + ca. 170 Modifikationen.
Müsste ich alle 174 Input-Werte in numerische Werte umwandeln, um z.b. RNN nutzen zu können?

Was genau ein Tensor ist -> oft als mehr dimensionales Array vorgestellt.
Wenn es ein Array ist, wozu braucht man Tensoren? wieso nutzt man nicht einfach die Matrizen, die man hat (Numpy, pandas).
Gibt es weitere unterschiede, außer, dass Tensoren vielleicht GPU optimiert sind.

Wenn ich CNN auf ein 2d bild anwende, welche Dimensionen hat der Tensor?
Was wird dadurch abgebildet (Größe, Breite des Bildes und Channels?)

MfG Sergej Ruff

Fragen zu KNN
>Wozu Pooling bei CNN oder besser, was wäre wenn ich es weglassen würde
>-> Wozu brauche ich eine Dimensionsreduktion bei CNN
Das Pooling dient zur Dimensionsreduktion um die Performance (im Endeffekt hauptsächlich die Geschwindigkeit) zu erhöhen. Die Wahl des Pooling kann aber auch einen gewissen Einfluss auf den Trainingserfolg haben. Ein Min Pooling ist zb eher bei Hellen Hintergründen in Bildern geeignet, da dann dunkle Pixel mehr berücksichtigt werden. Hier auch nochmal eine schöne Erklärung:
https://www.educative.io/answers/what-are-some-deep-details-about-pooling-layers-in-cnn

>Wäre es besser, wenn man Daten vorher skaliert und auf die selbe Skala bringt. Welchen Einfluss haben
>unterschiedliche Skalen auf KNN, wenn Sie Einfluss haben.
Definitiv! Um die Performance zu verbessern, werden die Feature (manchmal auch die Ausgaben) idR in den Wertebereich [0,1] transformiert. Hier findest du eine gute Beschreibung warum, das so für die verschiedenen Verfahren ist: https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/#Why_Should_We_Use_Feature_Scaling?
https://scikit-learn.org/stable/auto_examples/preprocessing/plot_scaling_importance.html

>Nehmen KNN nur numerische Werte an?
Ja, wie fast alle ML-Modelle.

>Wenn ja, wie funktionieren Sequenzielel Daten wie Texterkennung und Vorhersagen?
Man muss sich eine vernünftige Repräsentation der Daten (Embedding) überlegen. Bei der Texterkennung werden die Worte idR auf Vektoren abgebildet, die die Anzahl oder das Vorkommen jedes Wortes in einem Satz/Dokument/... wiederspiegeln. Dafür wird ein zuvor definiertes Alphabet verwendet, dass idR aus den Trainingsdaten abgeleitet wird (plus einem Term der alles Unbekannte beinhaltet). Ein Beispiel dafür ist Word2Vec, welches Wörter in Vektoren verwandelt, die bestimmte Maße erfüllen, dh arithmetische Operationen in sinnvoller Art zulassen. So könnte man zb das Embedding von "König" minus das Embedding von "Mann" nehmen und würde sehr wahrscheilich bei "Königin" landen. (König - Mann = Königin).

>z.b. Wenn ich RNA-Daten mit RNA-Modifikationen als Sequenzieller Input nutzen will, dann habe ich 4
>Nukleotide + ca. 170 Modifikationen.
>Müsste ich alle 174 Input-Werte in numerische Werte umwandeln, um z.b. RNN nutzen zu können?
Du musst die Werte irgendwie in Zahlen transformieren. Da ich gerade nicht genau weiß was mit dem letzten Satz gemeint ist, schlage ich vor hier nachher drüber zu reden.

>Was genau ein Tensor ist -> oft als mehr dimensionales Array vorgestellt.
>Wenn es ein Array ist, wozu braucht man Tensoren? wieso nutzt man nicht einfach die Matrizen, die man hat (Numpy, pandas). Gibt es weitere unterschiede, außer, dass Tensoren vielleicht GPU optimiert sind.

Ein Tensor ist die Verallgemeinerung eines Vektors auf n-Dimensionen. Sie werden bei KNNs genutzt um die Daten und Gewichte des Netzes zu repräsentieren. "Tensors provide a unified way to train neural networks for more complex data sets." In der CPU sind die hochdimensionalen Computations recht langsam. NVIDIA hat gegen 2011 angefangen GPUs herzustellen, die extra für die Manipulation von Tensoren ausgelegt sind und somit das Training deutlich beschleunigen. Siehe auch:
https://medium.datadriveninvestor.com/what-is-the-tensor-in-deep-learning-77c2af7224a1
https://en.wikipedia.org/wiki/Tensor_(machine_learning)

>Wenn ich CNN auf ein 2d bild anwende, welche Dimensionen hat der Tensor?
Wenn du den DataTensor meinst typischerweise 3D (height, width, channels).
>Was wird dadurch abgebildet (Größe, Breite des Bildes und Channels?)
genau.

Bis gleich Dominik

Hinweis zu skopt BayesSearchCV - np.int Fehlermeldung

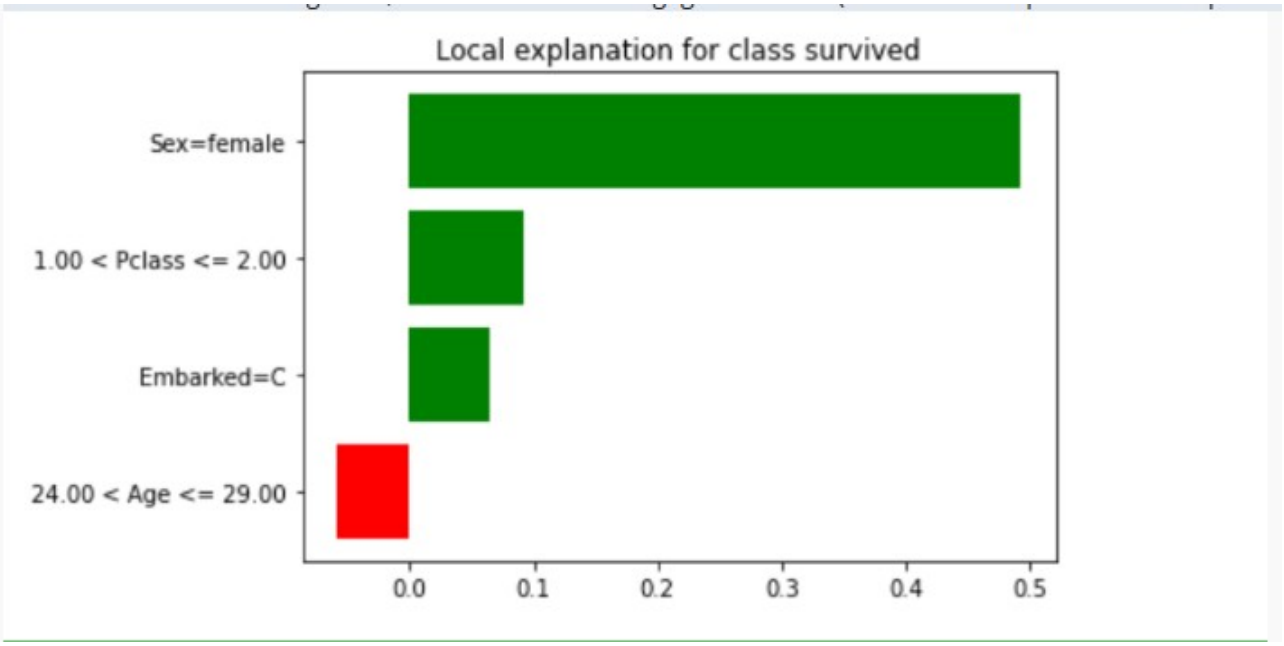
Hallo zusammen,

bei dem Package skopt kann es zu einer Fehlermeldung kommen, dass np.int nicht mehr verwendet wird.
Wenn die Fehlermeldung auftritt, am Besten einfach auf die Datei transformer.py in der Fehlermeldung klicken und dort alle Vorkommen von "np.int" durch "int" ersetzen (natürlich ohne Anführungszeichen).
Alternativ kann man auch im verwendeten Environment nach dem Package skopt suchen und dort im Unterordner "space" die Datei transformers.py korrigieren.

Liebe Grüße
Sarah

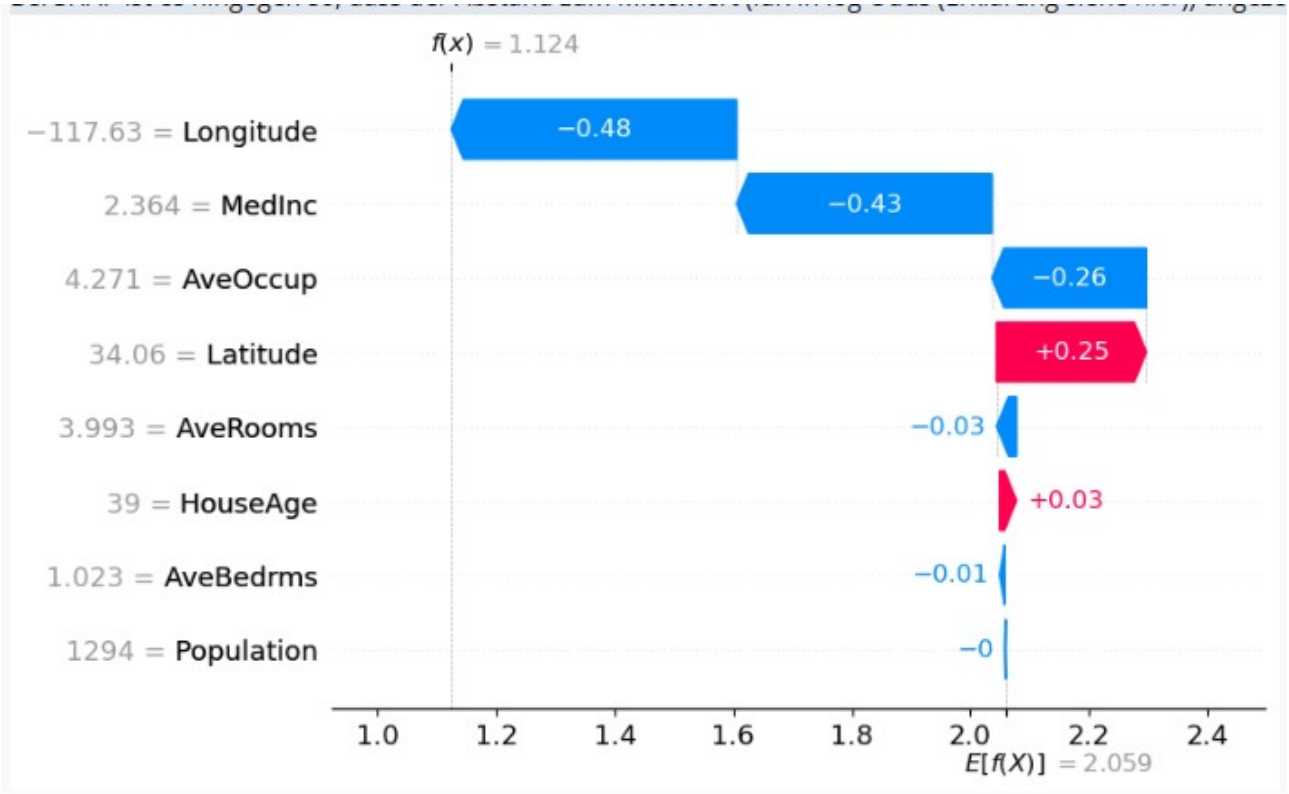
Hinweis zu skopt BayesSearchCV - np.int Fehlermeldung
np.int muss in Zeile 262 und in Zeile 275 ersetzt werden --> dann müsst ihr nicht so lange suchen ;-)

Interpretation von LIME und SHAP
Liebe Studierende,
ich bin ebnd noch die Antwort auf die Frage, was die Werte in der X-Achse des LIME plots bedeuten, schuldig geblieben.
Der Wert auf der X-Achse gibt an, wie stark sich der ausgegebene Wert (in unseren Beispielen die Classprobability) ändert durch die Ausprägung des Features.



In diesem Beispiel hat die Ausprügung des Geschlechts als weiblich einen positiven Anteil von 0,5 am ausgebenen Klassenscore (und das Alter einen leicht Negativen).
Diese Werte können über die Permutationen bestimmt werden, bei denen das Feature einen anderen Wert annimmt und sich das Klassifikationsergebnis änder (in diesem simplen Beispiel das Geschelcht "männlich").

Bei SHAP ist es hingegen so, dass der Abstand zum Mittelwert (idR in log Odds (Erklärung siehe hier)) angezeigt wird.



In dieser Abbildung sieht man den Erwartungswert (E[f(X)]) eingezeichnet (= durchschnittliche Vorhersage für den Datensatz) und wie stark die Ausprägungen der Feature im Bezug auf den Erwartungswert dazu beigetragen haben, den Output zu generieren.

Weitere informationen findet ihr auch nochmal unter:
https://towardsdatascience.com/using-shap-values-to-explain-how-your-machine-learning-model-works-732b3f40e137
https://shap.readthedocs.io/en/latest/example_notebooks/overviews/An%20introduction%20to%20explainable%20AI%20with%20Shapley%20values.html
https://vishesh-gupta.medium.com/understanding-your-model-with-lime-d7704d984696

Korrektur Imputation in Musterlösung - Wichtig!
Hallo zusammen,

bei mir hat sich tatsächlich ein Fehler in der Musterlösung bei der Imputation eingeschlichen.
Wenn man die Werte des Testdatensatzes imputieren muss (einige Methoden können die Prädiktion mit fehlenden Werten machen) ist Folgendes korrekt:

```
imp = IterativeImputer()
x_train = imp.fit_transform(x_train)
x_test = imp.transform(x_test)
```

Die Imputation wird also nur auf den Trainingsdaten erlernt, und dann auf den Trainings- wie auch den Testdaten angewendet.
Ich hatte in der Musterlösung fälschlicherweise jeweils für Trainings- und Testdaten separat eine Imputation berechnet.
Das ist zwar kein Problem hinsichtlich Data Leakage, aber in der Realität unsinnig.
Wenn man zum Beispiel ein trainiertes Modell hat und nun für einen einzelnen neuen Patienten die Vorhersage machen möchte, kann man an diesem einzelnen Patienten natürlich keine Imputation trainieren. Wenn ein Wert bei dieser "neuen Testinstanz" fehlt, kann man den nicht einmal durch den Mittelwert der neuen "Testdaten" ersetzen, da die "Testdaten" ja nur aus diesem einen Patienten bestehen, bei dem eben der Wert fehlt.

Die Imputation gehört also mit zu dem Modell, dass auf den Trainingsdaten trainiert und dann auf den Testdaten evaluiert wird. Daher trainiert oder definiert man die Imputation anhand der Trainingsdaten und wendet sie dann auf Training und Test an. In einer Pipeline geschieht das automatisch.
Das ist tatsächlich kein Problem mit Data Leakage. Data Leakage geschieht zum Beispiel, wenn die Testdaten Einfluss auf die Trainingsdaten haben, also die Imputation zum Beispiel auf den gesamten Daten erlernt/berechnet wird. Wenn man die Imputation wie oben durchführt, werden Training und Test also trotzdem separat behandelt. Man kann sich das mit der Analogie vorstellen, dass die Vergangenheit nicht von der Zukunft lernen darf, die Zukunft aber durchaus von der Vergangenheit.

Liebe Grüße
Sarah

Literaturempfehlungen
Hallo,

welche Literatur ist den zu empfehlen, um sich mit dem Thema ML und Deep Learning auseinander zu setzen?

Mich interessiert vor allem die Theorie und nicht unbedingt die direkte Implementierung (das wäre ein Bonuspunkt).
Es kann auch Mathe-lastig sein, wenn die Mathematik verständlich erklärt ist.

Auf der MHH Seite werden folgende Bücher empfohlen:
Artificial Intelligence - A Modern Approach Stuart J. Russell and Peter Norvig

Deep Learning with Keras Antonio Gulli

Einführung in Machine Learning mit Python Sarah Guido, Andreas C. Müller

Interpretable Machine Learning - A Guide for Making Black Box Models Explainable Christoph Molnar

Welcher dieser Bücher wäre gut für den Anfang? Und gibt es weitere Empfehlungen?

MfG Sergej Ruff

Hallo Sergej,
die Bücher sind alle zu empfehlen haben aber einen unterscheidlicehn Fokus.

Artificial Intelligence - A Modern Approach ist eine klassische Einführing in KI. Neben den Grundlagen zu Machine Learning, liegt ein starker Fokus auf symbolischer KI, welche wir in unserem Modul nicht behandeln.

Deep Learning with Keras ist ein super Standardwerk für künstliche Neuronale Netze und Deep Learning mit diesen.

Einführung in Machine Learning mit Python ist von deiner Beschreibung her das passenste Buch. Es bietet einen guten Überblick und Einstieg in die Themen der VL.

Interpretable Machine Learning - A Guide for Making Black Box Models Explainable befasst sich ausschließlich mit Explainable AI, welches wir im letzen Drittel der VL durchgehen.

Des Weiteren möchte ich noch die englischen Wikipediaseiten als kostenfreie Alternative erwähnen. Auch wenn es sich hierbei um keine wissenschaftliche Quelle handelt, bieten sie einen sehr hochqualitative Übersicht zu allen Themen des Bereichs, sowie oft auch sehr gute vertiefende Informationen.

Liebe Grüße
Dominik

Nachtrag "OpenSSL Appears to Be Unavailable on This Machine Anaconda"
Liebe Studierende,

als Nachtrag zu dem oben genannten Fehler von gestern von Conda hier der Link zur schrittweisen Lösung: https://www.youtube.com/watch?v=-6puHFu8zDY

Alternativ kann es ein Problem bei mit der Kompatibilität der Anacondaversion und eurem Chipsatz sein. Hier würde entweder ein Update oder Downgrade (Neuinstallation einer älteren Version) von Conda helfen.(https://stackoverflow.com/questions/45197777/how-do-i-update-anaconda)

Falls jemand es ausprobiert, würed ich mich über eine kurze Rückmeldung freuen.

LG
Dominik

Woche 7 _ Wert Lernrate
In der Vorlesung wurde bei der Berechnung der Trainingsschritte zum logischen AND wurde bei der Korrektur ein Wert von 0,2 für die Lernrate angenommen. Wie komme ich auf diesen Wert?

Woche 7 _ Wert Lernrate
In der Vl wurde es einfach ausgewählt. Es soll aber auch Verfahren geben, um ideale Lernraten zu bestimmen.

Hier ist ein Medium Artikel zu einem Verfahren: https://towardsdatascience.com/how-to-decide-on-learning-rate-6b6996510c98

Am Ende des Artikels steht noch folgendes:

If you don’t want to perform hyper-parameter search using different values, which can take ages, you have two options left: pick initial values at random (which may leave you with terribly bad performance and convergence but may work great if you are a lucky man though) or use a learning rate finder included in your machine learning framework of choice.

Einige Frameworks scheinen also auch einen eigenen Finder zu enthalten. Keras solle vielleicht auch sowas haben.
Liebe Katharina,
wie Sergej schon schrieb: der Wert wurde in der VL einfach festgelegt. Die Lernrate ist an sich frei wählbar und gibt die "Schrittweite" beim inkrementellen Updaten der Gewichte an. Wenn sie zu groß ist, ist es sehr wahrscheinlich, dass man den Tiefpunkt verpasst, und wenn sie zu klein ist, dauert das Training sehr lange. ADAM ist beispielsweise ein Optimizer, der die Lernrate während des Trainings automatisch anpasst, aber auch einen initialen Lernraten-Wert braucht.

@Sergej: Vielen Dank für die gute Antwort und den super Artikel!!