Review Alex 29/12/2017

> l120 wirklich drought? dry period nehme ich an

Ja, drought period ist der meteor. Ausdruck für Trockenperiode

> Ich finde immer noch, dass dieser Satz hier aus dem Rahmen fällt. Es kann ja nicht davon ausgegangen werden, dass der Leser hinreichend vertraut mit Bayesscher Statistik ist, um mit dieser Analogie etwas anfangen zu können.

Ok, ja - habs rausgenommen und für die Definition von hyperparamter vs parameter danach eine Referenz eingefügt.

> da kernel ja nun kein HYperparam. mehr ist, müsste das hier wohl rausgelassen werden

Fragt sich dann der Leser nicht, welcher kernel benutzt wurde?
Aber ja, das steht ja im SVM Abastz nochmal explizit drin.

> warum nicht auch distance=1, die Manhattan-Distanz?
> Sind distance-Wert bis 80 wirklich sinnvoll, oder wäre es ausreichend, bei 10 Schluss zu machen?
> https://en.wikipedia.org/wiki/Minkowski_distance
> Ich habe das Gefühl dass ich das schonmal gefragt habe, aber ehe die gleiche Frage vom Reviewer kommt, frage ich lieber nochmal nach…
> Wenn ich mir die plots so anschaue, sieht man, dass hohe distance-Werte manchmal optimal sind, aber wahrscheinlich ist das Optimum sehr sehr flach, so dass es dann auch wurscht ist… naja von mir aus lass die obere Grenze so, aber dass distance=1 fehlt, wurmt mich jetzt etwas.

Mit welcher Begründung willst du das Limit bei 10 setzen?
Ja, $distance = 1$ sollte mit drin sein.
Habs nochmal mit reingenommen und das *distance* upper limit auf 100 gesetzt. Hat sich kaum was getan in der performance aber ist sauberer.
$Distance = 1$ wurde auch ein paar mal als Optimum gewählt.
Kann nicht mehr sagen, warum ich es rausgelassen hatte…hätte ich mir mal notieren sollen.

> Ich finde dass das ganze grid search / random search Argument erstens am Thema vorbei geht und zweitens nicht besonders belastbar ist (wie schon bei der letzten Iteration angemerkt). Denn ob ich nun z.B. 50 random search Punkte oder 7^2 = 49 grid search Punkte abklappere, weder die Punktdichte noch die Rechenzeit werden sich unterscheiden. Im Gegenteil kann es bei der random search ja sogar passieren, dass irgendwo eine große Lücke im search space bleibt.
> Also sag einfach kurz, dass random search von den ein oder zwei Autoren als bessere Alternative zum grid search vorgeschlagen wurde und gut is.
> So jetzt hab ich mir aber auch mal das Bengio paper angeschaut. Knackpunkt ist, dass es um recht hochdimensionale Optimierungen geht und es eine “low

> effective dimensionality” gibt, weil manche Hyperparameter einfach wurscht sind. In dieser Situation ist es auch anschaulich klar, dass man in den relevanten Dimensionen besser abgedeckt hat.
> Im letzten bullet item auf S. 283 sagen die Autoren auch, dass grid search in niedrigdimensionalen Suchräumen verlässlich sind, und auf der gleichen Seite sagen sie, dass es ihnen darum geht, die Überlegenheit von random search in hochdimensionalen Suchräumen mit low effective dimensionality zu zeigen.

Ja da hast du Recht, uns geht es ja um den Unterschied zwischen SpCV und non SpCV und nicht grid search vs random search.
Ich hab es jetzt aufs folgende heruntergebrochen:
“We used a random search with a varying number of iterations (10, 50, 100, 200, 1000) for all machine learning models in this study to analyze the difference of varying tuning iterations.
Random search has been shown to be superior to grid search by \citep{Bergstra2012}.”

> [l182] Dieser Absatz lässt sich zu einem Satz kürzen, wenn du wie oben vorgeschlagen zuerst die CV-Fehelrschätzung einführst

Jop, hab ich so geändert :)

> [l197] "Model performance estimates will most
> often be overoptimistic due to the similarity of training and test data in a non-
> spatial partitioning setup when using any kind of cross-validation for tuning or
> validation (Brenning, 2012). "

Wolltest du noch umschreiben laut deiner Note?

> Dieser Satz gehört hier nicht hin. Die Info über hyperparameter ranges gehört in den Abschnitt zum hyperparameter tuning oder eben zu den Modellbeschreibungen. Die computing times sind ERGEBNISSE und haben hier nichts zu suchen, auch nicht als Tabellenverweis.

Ok, thanks for the heads up :)

> at the fold level of the outer CV?? (Bezüglich der Parallelisierung)

Nein, die folds des outer loops laufen sequentiell, nur das tuning wurde parallelisiert. Man kann immer nur ein Level parallelisieren.

> Dieser Absatz passt hier nicht so richtig rein. Vielleicht ließe er sich wie folgt umformulieren:
> An exemplary selection of widely-used statistical and machine-learning techniques was compared in this study. While the following sections describe the used models and their settings, a justification of the choice

> of specific implementations in the R statistical software is included in Appendix A.

Ja, schön formuliert. Hach, bei dir wirkt das immer so easy ;)

> sigma, C?

Hab noch eine Zeile eingefügt.

> Es wäre aus meiner Sicht angemessener, durchgehend von Weighted k-NN zu sprechen, da die meisten Leser unter k-NN eine nicht-gewichtete k-NN verstehen/erwarten werden. Das package verwendet auch diese Bezteichnung. → “WKNN”

Ja, durch die Kernel Sache ist das sonst schwammig. Habs geändert!

> Entweder bei allen Methoden das Package nennen und kurz begründen oder bei keiner. Einleitend wurde in 3.4 ja auch Appendix A verwiesen → ggf. Hinweis abändern.

Ja, habs geändert hin zu WKNN überall und die package Nennung nur im Appendix A.

> Vielleicht einfach diesen letzten Satz weglassen, sonst fragt sich der geneigte Leser noch, wo denn die Breite der Kernfunktionen herkommt, wenn nicht aus einem Hyperparameter. (Vermutlich adaptiv über die Größe der lokalen Nachbarschaft definiert.)

Ok.

> Dieser Absatz gehört aber nicht in Abschnitt 3.4.5 → eigener Abschnitt 3.4.6 oder in den Über-Abschnitt 3.4

Ja, habs nach oben zu 3.4 geschoben, wo auch der Verweis zu den package choices steht.

> [Figure 3] Hier müsste die y-Achse auch logarithmisch dargestellt werden, sonst wird eine nichtlineare Beziehung vorgegaukelt. Eigentlich muss das ja super linear sein, insofern finde ich, dass man sich die Abbildung dann auch schenken kann.

Ja, der Anstieg sollte linear sein solange nichts am setup sich geändert hat. Meine Idee war einfach, einen graphischen Vergleich der runtime zu haben, da man damit schneller den relativen Unterschied der Modelle sieht als an nackten Zahlen.
Jedoch hängt der ja auch wieder stark von den settings ab bei RF und BRT.
Die Darstellung der x-Achse ist nicht logarithmisch, die tuning iteratiosn sind einfach als Faktor codiert. Ist natürlich bissl unschön, wenn man numerische Werte als nominal ansieht…
Wenn ich sie numerisch codiere, sitzen die ersten alle aufeinander weil der Sprung zu der 1000 so groß ist. Wenn ich dazu schreibe, dass die x-Achse nominal skaliert ist, wäre das dann in Ordnung?

> Ich finde diese Kürzel ja etwas schwer zu merken, zumal “sp” und “nsp” beide links wie rechts des / auftreten können.
> Wäre es wirklich so mühsam, z.B. die Spalten zu überschreiben mit “Non-spatial CV [Zeilenumbruch] Non-spatial tuning” etc.?
> Mit den runtimes bin ich hier auch nicht so richitg happy. So richtig viel Info steckt da ja nicht drin, und es hat auch nichts mit den AUROCs zu tun. Auch zwischen den drei nested CVs war ja keine runtime-Differenz erwartet worden. Reicht es nicht, die Spannweiten im Text zu erwähnen, z.B. allgemein 0.20-0.31 ohne tuning, und mit tuning für SVM 128-134. In der Tabelle fehlen ja auch runtimes von GLM und GAM. A propos, wieso nicht die AUROCs von GLM und GAM in diese Tabelle hineinstecken? Ich weiß es geht dir hier um tuning, aber in der sp/not-Spalte könnte man doch GLM und GAM aufführen.

Mein Gedanke war, dass man die Kürzel aus dem Text dann auch wieder in der Tabelle findet. Aber ja, vll ist es am Besten, eine Kombination zu machen in der Tab., z.B. Non-spatial/Non-spatial (nsp/nsp)?
GLM und GAM hab ich rausgelassen, da es mir ums tuning ging, ja.
Aber ich kann Sie im setting sp/not unterbringen, stimmt.
Genau genommen würde dann noch nsp/not fehlen fürs GLM und GAM alleine.
Meinst du, dass eine note in der table caption reicht, dass diese stat fehlt?

> Spaltenreihenfolge ist anders als in den Abbildungen.
> Modellreihenfolge auch.
> nsp/nsp Spaltenbeschriftung ist einmal falsch.

Ja, einfach nur schlampig. Habs bereinigt und nun nach Alphabet gemacht für die ML models.

> “tradeoff” lässt sich nur mit antagonistischen Sachen verwenden, also z.B. tradeoff zwischen Trainingsdauer und Trainingsintensität.

Ok, habs mal umformuliert. Hoffe, dass der Antagonismus jetzt so stimmt.

> Das stand bei Bergstra & Bengio aber nicht so explizit geschrieben? Dort war von hochdimensional und low effective dimensionality die Rede.
> Vielleicht eher folgendes:
> Random search algorithms are particularly promising in multidimensional hyperparameter spaces with possibly redundant or insensitive hyperparameters (low effective dimensionality; Bergstra & Bengio, 2012). These as well as adaptive search algorithms offer computationally efficient solutions to these difficult global optimization problems in which little prior knowledge on optimal subspaces is available. In this study, a random search with at least 50 iterations was sufficient for all algorithms considered.

Hm ja, ich glaube, ich hatte das mit der “low effective dimensionality” möglicherweise falsch verstanden.

Beeindruckend, wie gut und verstädnlich du das immer in so wenigen Worten ausdrücken kannst!

> [l368] hyperparameter “combination” instead of hyperparameter “setting”

Jakob Richter aus Dortmund meinte, dass er den Begriff Hyperparameter combination eher vermeiden würde und immer von setting sprechen würde. Kann man sich sicherlich auch drüber streiten aber da ichs jetzt kontinuierlich “setting” genannt hab im paper, bleibe ich mal der Kontinuität wegen dabei :)

> RF ist einfach ziemlich insensitiv gegenüber ntree und mtry, da isses dann auch wurscht ob du zB mit räumlichem tuning nen mtry=1 oder 2 oder nichträumlich 2-4 bekommst. Es wäre halt mal schön, so einen Querschnitt durch die zu optimierende Funktion anzuschauen, dann kann man etwas informierter darüber reden, wie wurscht diese Parameter sind…
> Und n.tree ist ja eh wurscht und sehr intuitiv klar, dass man mit zu vielen Bäumen nichts kaputt machen kann (im Gegensatz zu BRT). Das Modell wird durch mehr Bäume ja nicht komplexer, nur eine bessere Zickzack-Näherung einer glatten Funktion.
> Das einzige was wirklich zu einer stärkeren Generalisierung bei RF führen würde, wäre ein kleinerer maxdepth oder größerer min.nodesize oder wie das heißt.

Hmm, die Grundaussage, die ich machen will, ist Folgende: Die generell ja relativ guten defaults von RF sind nicht “so gut” für spatial data sets und daher ist ein spatial tuning eben doch wichtig, auch wenn es nicht direkt zu einer besseren performance beiträgt.
Ja, wir haben kaum einen performance Unterschied aber das kann bei anderen Datensätzen ja anders aussehen?
Denkst du, es wäre interessant, mal eine Vergleichststudie zu machen, die verschiedene spatial data sets analyisert und eine neue default Funktion für $mtry$ für spatial data sets ermittelt?
Oder lohnt sich der Aufwand nicht, da der neue optimal $mtry$ Wert für spatial datasets evtl. sogar eine performance Verschlechterung bringen würde?
Ich dachte halt schon, dass ein höherer $mtry$ ein bisschen von der sp. autocor. profitieren kann, ansonsten würde ja ein nsp tuning setting nicht so oft höhere Werte wählen, oder?
Und ja nur $mtry$, dass $num.trees$ nicht so viel ausmacht/kaputt macht ist ja klar bei der Struktur von RF.

Und wenn $max.depth$ und $min.node.size$ auf die Generalisierung Einfluss haben, warum werden die dann eigentlich nie getuned? Sie haben ja auch wirklich kaum einen Einfluss auf eine performance Änderung (hatte das mal probiert…).

Eiei, Fragen über Fragen…

> Generell kann aber eben doch sagen, dass die default parameters von RFnahe am Optimum für diesen Datensatz liegen und gerade n.tree keinen großen Einfluss haben dürfte, solange er >100 ist.

Ja, aber kann das nicht Zufall sein? So groß ist die Range von 1-11 ja nicht und wenn wir da um 2 (absolut) im spatial abweichen ist das ja relativ auch schon ne ganze Ecke. Evtl. würde sich das bei einem Datensatz mit 100 + Variablen dann viel stärker herauskristallisieren?

> so war das sicher nicht gemeint. Der Bias kann ja auch sein Vorzeichen wechseln, so gesehen könnte der räumliche Schätzer auch zu konservativ sein! “bias-reduced” ist also nur eine betragsmäßig geringere Verzerrung – und schlichtweg eine vorsichtigere Aussage, als zu behaupten, dass nun alles unverzerrt zugeht. Im Buch von David Hand sind ein paar schöne Seiten zu bias in bootstrap estimators, wo man auch staunt, dass sich theoretisch beweisen lässt, dass intuitiv einleuchtende Schätzer immer noch verzerrt sind…

Oke ja, das liest sich einseitig. Habs geändert und schau mir das Buch von David Hand mal an!

> wird im Text einfach nur als ‘hail’ bezeichnet??
> Aber ich dachte das ist deine Hagelpotentialvorhersage, hat die nicht Werte zwischen 0 und 1 (vorhergesagte Wahrscheinlichkeiten)??

Ja, habs geändert zu “hail_new” und italic geschrieben.
Ja richtig, aber ich hab sie auf 0/1 codiert und als Faktorvariable eingebaut. Wäre numerisch besser gewesen?.. Weil mehr Information? :(

> was ist denn eine soil horizon difference?
> was für eine Bodenart ist limited space for roots?
> Können wir bitte nochmal diese Bezeichnung zusammen durchgehen?

Ja, das ist natürlich nichtssagend. Das war sogar nicht mal die aktuellste Benennung *facepalm*.
Habs updated, mit der jeweiligen Übergruppe und den eigentlichen Bodennamen in Klammern.