MINDCRAFT

CRAFTING INTELLIGENT MINDS

Disclaimer: This content is generated by AI.

मॉड्यूल 5: डेटा प्रीप्रोसेसिंग आणि फीचर इंजिनिअरिंग

:

या मॉड्यूलमध्ये आवश्यक डेटा प्रीप्रोसेसिंग तंत्रांचा समावेश आहे, ज्यामध्ये डेटा क्लीनिंग, हरवलेली मूल्ये हाताळणे, फीचर स्केलिंग आणि मशीन लर्निंग मॉडेलसाठी डेटा तयार करण्यासाठी वर्गीय व्हेरिएबल्स एन्कोड करणे समाविष्ट आहे.

डेटा क्लीनिंग तंत्र

डेटा क्लीनिंग, मशीन लर्निंगमधील एक महत्त्वपूर्ण प्रीप्रोसेसिंग टप्पा, यात चुकीचा, अपूर्ण, असंबद्ध, डुप्लिकेट किंवा विसंगत डेटा ओळखणे आणि दुरुस्त करणे (किंवा काढणे) समाविष्ट आहे. स्वच्छ डेटा तुमच्या मशीन लर्निंग मॉडेलची विश्वासार्हता आणि अचूकता सुनिश्चित करतो. घाणेरड्या डेटामुळे पक्षपाती मॉडेल, चुकीचे अंदाज आणि संसाधने वाया जाऊ शकतात.

गहाळ मूल्ये हाताळणे

गहाळ डेटा ही एक सामान्य समस्या आहे. तंत्रांमध्ये हटवणे (गहाळ मूल्यांसह पंक्ती किंवा स्तंभ काढून टाकणे), आरोपण (गहाळ मूल्यांना सरासरी, मध्य, मोड सारख्या अंदाजे मूल्यांसह बदलणे किंवा -नजीकच्या शेजारी सारख्या अधिक अत्याधुनिक पद्धती वापरणे) किंवा गहाळ डेटा मूळतः हाताळणारे अल्गोरिदम वापरणे समाविष्ट आहे. सर्वोत्तम दृष्टीकोन डेटासेट आणि गहाळ डेटाच्या प्रमाणावर अवलंबून असतो. उदाहरणार्थ, मोठ्या डेटासेटमध्ये मूल्यांची लहान टक्केवारी गहाळ असल्यास, आरोप लावणे योग्य असू शकते. तथापि, डेटाचा महत्त्वपूर्ण भाग गहाळ असल्यास, हटवणे आवश्यक असू शकते किंवा अधिक मजबूत आरोप तंत्र आवश्यक असू शकते.

व्यवहार

आउटलयिर हे डेटा पॉइंट्स आहेत जे इतर नरीक्षणांपेक्षा लक्षणीय भन्नि आहेत. ते परणिाम कमी करू शकतात आणि मॉडेलची अचूकता कमी करू शकतात. शोध पद्धतीमध्ये बॉक्स प्लॉट, स्कॅटर प्लॉट आणि झेड-स्कोअर गणना यांचा समावेश होतो. आउटलयिर्स हाताळण्यामध्ये काढून टाकणे (त्या तरुटी असल्यास कवि असामान्य परस्थितीमुळे), परविर्तन (उदा., अत्यंत मूल्यांचा प्रभाव कमी करण्यासाठी लॉगरदिमकि ट्रान्सफॉर्मेशन वापरणे) कवि आउटलायर्ससाठी कमी संवेदनशील असलेले मजबूत अल्गोरदिम वापरणे समावष्टि आहे. उदाहरणार्थ, घराच्या कमितीच्या डेटासेटमध्ये, 1 दशलक्षपेक्षा कमी कमितीच्या घरांमध्ये 10 दशलक्ष कमितीचे घर बहुधा आउटलायअर आहे आणि त्यासाठी पुढील तपास कवि काढण्याची आवश्यकता असू शकते.

डुप्लकिेट ओळखणे आणि काढणे

डुप्लकिेट डेटा पॉइंट वशिष्टि वैशष्ट्यांचे महत्तत्व वाढवू शकतात आणि चुकीचे मॉडेल होऊ शकतात. पंक्ती क्रमवारी लावणे आणि तुलना करणे यासह ववििधि तंत्रे वापरून डुप्लकिेट ओळखले जाऊ शकतात कवि ' सारख्या प्रोग्रामगि भाषांमध्ये वशिष कार्ये वापरणे. एकदा
ओळखल्यानंतर, प्रत्येक अनन्य डेटा पॉइंटचा एकच प्रसंग ठेवून डुप्लकिेट काढले जाऊ शकतात.

डेटा ट्रान्सफॉर्मेशन

डेटा ट्रान्सफॉर्मेशनमध्ये डेटाचे रूपांतर मशीन लर्नगि अल्गोरदिमसाठी अधकि योग्य स्वरूपात करणे समावष्टि आहे. सामान्य परविर्तनांमध्ये स्केलगि (उदा. कमिान-अधकितम स्केलगि, मानकीकरण), सामान्यीकरण आणि एन्कोडगि वर्गीय चल (उदा. एक-हॉट एन्कोडगि, लेबल एन्कोडगि) यांचा समावेश होतो. स्केलगि हे सुनश्चिति करते की मोठ्या मूल्यांसह वैशष्ट्ये मॉडेलवर असमानतेने प्रभाव पाडत नाहीत. एन्कोडगि अल्गोरदिम समजू शकणाऱ्या संख्यात्मक प्रेझेंटेशनमध्ये स्पष्ट डेटा (उदा. रंग, श्रेणी) रूपांतरति करते.

वसिंगती हाताळणी

समान माहतिीचे प्रतनिधित्वि करूनही डेटा वेगळ्या पद्धतीने रेकॉर्ड केला जातो तेव्हा डेटा वसिंगती उद्भवते (उदा. 'यूएसए' वि 'यूएस' वि 'युनायटेड स्टेट्स'). मानकीकरण महत्त्वाचे आहे; यामध्ये डेटा एंट्रीसाठी एक सुसंगत स्वरूप तयार करणे समावष्टि आहे. यामध्ये डेटा फॉरमॅट्स एकत्रति करण्यासाठी स्ट्रगि मॅनपिुलेशन तंत्र वापरणे कवि मानक मूल्यामध्ये वसिंगत नोंदी मॅप करण्यासाठी डेटा शब्दकोश वापरणे समावष्टि असू शकते.

Reference:

https://www.kaggle.com/learn/data-cleaning

https://towardsdatascience.com/data-cleaning-with-python-and-pandas-a-step-by-step-tutorial-d2f200271f0a

https://r4ds.had.co.nz/tidy-data.html

Video Links:

https://www.youtube.com/watch?v=LI7s_lyooO8

https://www.youtube.com/watch?v=qxpKCBV60U4&pp;=ygUMI2RhdGFjbGVhbnNl

https://www.youtube.com/watch?v=i5ryMeGDnHg

https://www.youtube.com/watch?v=2Jw5S5EbpwA

https://www.youtube.com/watch?v=FbFQH_RNMu0

https://www.youtube.com/watch?v=qGIYA04ZlWc

https://www.youtube.com/watch?v=mZFKy2K8sP8

https://www.youtube.com/watch?v=kNl7YDN-_js

https://www.youtube.com/watch?v=_6a1AZ8R7cI

https://www.youtube.com/watch?v=oT4emh72fuA

गहाळ मूल्ये हाताळणे

रअिल-वर्ल्ड डेटासेटमध्ये गहाळ मूल्ये ही एक सामान्य समस्या आहे. योग्यरत्या हाताळले नाही तर ते मशीन लरनगि मॉडेलच्या कार्यक्षमतेवर लक्षणीय परणिाम करू शकतात. हे उप-मॉड्यूल तुमच्या मॉडेल्सची अचूकता आणि विश्वासार्हता सुनश्चिति करण्यासाठी गहाळ डेटा शोधण्यासाठी, समजून घेण्यासाठी आणि संबोधति करण्यासाठी ववििधि तंत्रे एक्सप्लोर करते.

गहाळ डेटाचे प्रकार

योग्य हाताळणी धोरण नविडण्यासाठी गहाळ डेटाचे स्वरूप समजून घेणे महत्वाचे आहे. प्राथमकि प्रकार आहेत: * **यादृच्छकिपणे पूर्णपणे गहाळ ():** डेटा पॉइंट गहाळ होण्याची संभाव्यता डेटासेटमधील इतर व्हेरएिबल्सशी संबंधति नाही. उदाहरण: अपघाती डेटा एंट्री त्रुटी. * **यादृच्छकिपणे गहाळ ():** डेटा पॉइंट गहाळ होण्याची संभाव्यता इतर नरीक्षण केलेल्या चलांवर अवलंबून असते. उदाहरण: पुरुषांच्या तुलनेत महलिांनी त्यांच्या उत्पन्नाची तक्रार करण्याची शक्यता कमी असते, परंतु हे केवळ आमच्याकडे लगि डेटा असल्यामुळेच ज्ञात आहे. * **मसिगि नॉट ॲट रँडम ():** डेटा पॉइंट गहाळ होण्याची शक्यता गहाळ मूल्यावरच अवलंबून असते. उदाहरण: जास्त उत्पन्न असलेल्या व्यक्तींना त्यांच्या उत्पन्नाचा अहवाल देण्याची शक्यता कमी असू शकते, ज्यामुळे गहाळपणा अनरीक्षति उत्पन्नावर अवलंबून असतो. हाताळण्यासाठी हा सर्वात आव्हानात्मक प्रकार आहे.

गहाळ मूल्ये हाताळण्यासाठी पद्धती

गहाळ डेटा हाताळण्यासाठी अनेक तंत्रे अस्तित्वात आहेत. सर्वोत्तम दृष्टीकोन गहाळ डेटाचा प्रकार, डेटासेटचा आकार आणि व्हेरएिबल्सची वैशष्ट्ये यावर अवलंबून असतो: * **हटवणे:** * **सूचीनुसार हटवणे:** कोणतीही गहाळ मूल्ये असलेल्या संपूर्ण पंक्ती काढून टाका. सोपे परंतु लक्षणीय डेटाचे नुकसान होऊ शकते, वशिषत: अनेक गहाळ मूल्यांसह कवि लहान डेटासेटसह. * **जोडीनुसार हटवणे:** प्रत्येक गणनेसाठी उपलब्ध डेटा वापरा. यामुळे वशिषत: सांख्यकिीय वश्लिेषणामध्ये वसिंगती आणि गुंतागुंत होऊ शकते. * **आकलन:** गहाळ मूल्ये अंदाजे मूल्यांसह पुनर्स्थति करणे. * **मध्य/मध्य/मोड इम्प्युटेशन:** गहाळ मूल्ये मध्य (संख्यात्मक डेटासाठी), मध्यक (आउटलाअरसह संख्यात्मक डेटासाठी) कवि मोड (वशिष्ट डेटासाठी) सह पुनर्स्थति करा. साधे पण व्हेरएिबलचे वतिरण वकिृत करू शकते आणि भन्निता कमी करू शकते. * **के-नजीकचे शेजारी () इम्प्युटेशन:** समान डेटा बदिूंच्या मूल्यांवर आधारति गहाळ मूल्यांचा अंदाज लावा. साध्या आरोप पद्धतीपेक्षा अधकि परष्कृत. * **मल्टपिल इम्प्युटेशन:** एकापेक्षा जास्त प्रशंसनीय आरोपति डेटासेट तयार करा आणि आरोपति मूल्यांमधील अनश्चितितेसाठी परणिाम एकत्रति करून त्यांचे स्वतंत्रपणे वश्लिेषण करा. अधकि संगणकीयदृष्ट्या गहन परंतु सामान्यत: अधकि मजबूत दृष्टीकोन मानला जातो. * **प्रेडकि्शन मॉडेल:** इतर व्हेरएिबल्सवर आधारति गहाळ मूल्यांचा अंदाज लावण्यासाठी एक मॉडेल तयार करा. यासाठी मॉडेलच्या गृहीतके आणि संभाव्य पूर्वाग्रहांचा काळजीपूर्वक वचिार करणे आवश्यक आहे. * **एक वैशष्ट्य म्हणून मसिगिनेस वापरणे:** गहाळपणाचा नमुना माहतिीपूर्ण असल्यास, मूल्य गहाळ आहे की नाही हे दर्शवणारे नवीन वैशष्ट्य तयार करा. हे वशिषत: डेटासाठी उपयुक्त आहे.

योग्य पद्धत नविडणे

इष्टतम पद्धत संदर्भावर अवलंबून असते. वचिार करा: * **गहाळ झालेल्या डेटाची रक्कम:** एक लहान रक्कम हटवण्याची परवानगी देऊ शकते, तर मोठ्या रकमेसाठी आरोप कवि इतर धोरणे आवश्यक असतात.

* **गहाळ झालेल्या डेटाचा प्रकार:** डेटा साध्या आरोपाने हाताळला जाऊ शकतो, तर डेटासाठी अधिक अत्याधुनिक तंत्रांची आवश्यकता असते. * **व्हेरिएबल प्रकार:** संख्यात्मक आणि स्पष्ट व्हेरिएबल्सना वेगवेगळ्या आरोप पद्धतींची आवश्यकता असते. * **मॉडेलवरील प्रभाव:** योग्य मूल्यमापन मेट्रिक्स वापरून तुमच्या मशीन लर्निंग मॉडेलच्या कार्यक्षमतेवर विविध पद्धतींचा प्रभाव मूल्यांकन करा.

Reference:

https://scikit-learn.org/stable/modules/impute.html

https://towardsdatascience.com/6-different-ways-to-compensate-for-missing-values-in-a-dataset-c72865531777

Video Links:

https://www.youtube.com/watch?v=P_iMSYQnqac

https://www.youtube.com/watch?v=kIKg5s_jDAk

https://www.youtube.com/watch?v=v0HItX1xhNg&pp;=ygUPI21pc3NpbmdfdmFsdWVz

https://www.youtube.com/watch?v=hHLQSNWOq64

https://www.youtube.com/watch?v=EYySNJU8qR0

https://www.youtube.com/watch?v=E1_pHKOfUxA

https://www.youtube.com/watch?v=_BFMS1IefzE

https://www.youtube.com/watch?v=FPvMBl8LvGA

https://www.youtube.com/watch?v=XcBFxaHmoas

https://www.youtube.com/watch?v=Mz45KDpSRbs&pp;=ygUSI2V4Y2VsZGF0YWhhbmRsaW5n

वैशिष्ट्य स्केलिंग आणि सामान्यीकरण

फीचर स्केलिंग आणि नॉर्मलायझेशन हे मशीन लर्निंगमधील महत्त्वपूर्ण प्रीप्रोसेसिंग टप्पे आहेत. त्यामध्ये तुमच्या डेटासेटच्या वैशिष्ट्यांचे मानक श्रेणीमध्ये रूपांतर करणे, अनेक मशीन लर्निंग अल्गोरिदमचे कार्यप्रदर्शन आणि कार्यक्षमता सुधारणे समाविष्ट आहे. ही प्रक्रिया भिन्न स्केल किंवा युनिट्स असलेल्या वैशिष्ट्यांमुळे उद्भवलेल्या समस्यांचे निराकरण करते, ज्यामुळे मॉडेल प्रशिक्षणादरम्यान पक्षपाती परिणाम किंवा मंद अभिसरण होऊ शकते.

फीचर स्केलिंग म्हणजे काय?

वैशिष्ट्य स्केलिंग स्वतंत्र व्हेरिएबल्स किंवा डेटाच्या वैशिष्ट्यांच्या श्रेणीचे मानकीकरण करण्यासाठी वापरल्या जाणाऱ्या तंत्रांचा संदर्भ देते. हे महत्त्वाचे आहे कारण अनेक मशीन लर्निंग अल्गोरिदम वैशिष्ट्यांच्या प्रमाणात संवेदनशील असतात. अंतर-आधारित गणना वापरणारे अल्गोरिदम, जसे की -जवळचे शेजारी किंवा सपोर्ट वेक्टर मशीन, विशेषतः प्रभावित होतात. स्केलिंगशिवाय, मोठ्या मूल्यांसह वैशिष्ट्ये अंतराच्या गणनेवर वर्चस्व राखतील, लहान मूल्यांसह वैशिष्ट्यांच्या प्रभावाची छाया करेल. सामान्य स्केलिंग पद्धतींमध्ये मिनि-मॅक्स स्केलिंग आणि मानकीकरण (झेड-स्कोअर सामान्यीकरण) यांचा समावेश होतो.

किमान-मॅक्स स्केलिंग

कमिान-अधकितम स्केलगि वैशष्टि्यांचे एका वशिष्टि श्रेणीत रूपांतर करते, सामान्यतः 0 आणि 1 दरम्यान. सूत्र आहे: _ = (- _) / (_ - _). ही पद्धत डेटा पॉइंट्समधील सापेक्ष
नातेसंबंध जतन करते परंतु आउटलायर्ससाठी संवेदनशील आहे. उदाहरण: जर एखाद्या वैशष्टि्याची मूल्ये 10 ते 1000 पर्यंत असतील, तर कमिान-अधकितम स्केलगि 10 ते 0 आणि 1000 ते 1 मध्ये रूपांतरति होईल, इतर सर्व मूल्ये आनुपातकिपणे मोजली जातात.

मानकीकरण (-स्कोर सामान्यीकरण)

मानकीकरण वैशष्टि्यांचे रूपांतर 0 ची सरासरी आणि मानक वचिलन 1 मध्ये करते. सूत्र आहे:
= (- सरासरी) / मानक वचिलन. ही पद्धत मनिमि-मॅक्स स्केलगिपेक्षा आउटलायर्ससाठी कमी संवेदनशील आहे आणि सामान्यत: वतिरति डेटा गृहीत धरणाऱ्या अल्गोरदिमसाठी प्राधान्य दलि जाते. उदाहरण: जर एखाद्या वैशष्टि्याचे मध्यमान 50 असेल आणि मानक वचिलन 10 असेल, तर 60 चे मूल्य (60-50)/10 = 1 मध्ये बदलले जाईल.

वैशष्टि्य सामान्यीकरण म्हणजे काय?

फीचर नॉर्मलायझेशन हा एक वशिष्टि प्रकारचा स्केलगि आहे जथि वैशष्टि्यांना एक युनटि नॉर्म (लांबी 1) म्हणून मोजले जाते. हे वशिषत: युक्लडियिन अंतर कवि कोसाइन समानता वापरणाऱ्या अल्गोरदिमसाठी उपयुक्त आहे जे वैशष्टि्य वेक्टरच्या वशिालतेस संवेदनशील असतात. सामान्य सामान्यीकरण तंत्रांमध्ये 1 आणि 2 सामान्यीकरणे समावष्टि आहे.

1 आणि 2 सामान्यीकरण

1 सामान्यीकरण प्रत्येक वैशष्टि्य सदशि स्केल करते जेणेकरून त्याच्या घटकांच्या नरिपेक्ष मूल्यांची बेरीज 1 असेल. 2 सामान्यीकरण प्रत्येक वैशष्टि्य वेक्टरचे मोजमाप करते जेणेकरून त्याच्या घटकांच्या वर्गांची बेरीज 1 (युक्लडियिन नॉर्म) असेल. 2 सामान्यीकरण अधकि सामान्य आहे आणि अनेकदा प्राधान्य दलि जाते कारण ते 1 सामान्यीकरणाच्या तुलनेत आउटलायर्ससाठी कमी संवेदनशील आहे.

कोणती पद्धत कधी वापरायची?

स्केलगि आणि सामान्यीकरण यामधील नविड वशिष्टि अल्गोरदिम आणि डेटासेटवर अवलंबून असते. वैशष्टि्यपूर्णतेसाठी संवेदनशील अल्गोरदिम सहसा सामान्यीकरणाचा फायदा घेतात. जेव्हा आउटलायर्स उपस्थति असतात कवि जेव्हा डेटा समान रीतीने वतिरति केला जात नाही तेव्हा मानकीकरणास प्राधान्य दलि जाते. जेव्हा डेटाची श्रेणी जतन करणे महत्त्वाचे असते तेव्हा कमिान-अधकितम स्केलगि योग्य असते. दलिेल्या समस्येसाठी सर्वोत्तम दृष्टीकोन नश्चिति करण्यासाठी प्रयोग अनेकदा आवश्यक असतात.

वास्तवकि-जागतकि उदाहरणे

अनेक वास्तवकि-जागतकि अनुप्रयोग वैशष्टि्य स्केलगि आणि सामान्यीकरण वापरतात. इमेज प्रोसेसगिमध्ये, पक्सिेल व्हॅल्यूज 0-1 च्या श्रेणीत सामान्यीकृत केले जातात. नैसर्गकि भाषा प्रक्रयिेत, शब्द एम्बेडगि अनेकदा 2 सामान्यीकृत केले जातात. फायनान्समध्ये, स्केलगिचा वापर वेगवेगळ्या कमिती श्रेणीसह स्टॉकची तुलना करण्यासाठी केला जातो. वैद्यकीय इमेजगिमध्ये, नदिान मॉडेल्सची अचूकता सुधारण्यासाठी वैशष्टि्य स्केलगि अनेकदा लागू केले जाते.

Reference:

https://scikit-learn.org/stable/modules/preprocessing.html

https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-techniques/

https://towardsdatascience.com/scale-standardize-or-normalize-with-scikit-learn-6b8911d55f4a

Video Links:

https://www.youtube.com/watch?v=sxEqtjLC0aM

https://www.youtube.com/watch?v=bqhQ2LWBheQ

https://www.youtube.com/watch?v=mnKm3YP56PY

https://www.youtube.com/watch?v=6eJHk8JYK2M&pp;=ygUTI3B5dGhvbmRhdGFoYW5kbGluZw%3D%3D

https://www.youtube.com/watch?v=CFA7OFYDBQY&pp;=ygUOI3NjYWxlZmVhdHVyZXM%3D

https://www.youtube.com/watch?v=4d58jmAoSdk

https://www.youtube.com/watch?v=EgISGYkGa5A

https://www.youtube.com/watch?v=TrfyVc7Vcv8

https://www.youtube.com/watch?v=Q-45O3b1pO8

https://www.youtube.com/watch?v=IG38XJ2Ewi4

वैशिष्ट्य नविड पद्धती

वैशिष्ट्य नविड ही मशीन लर्नगिमधली एक महत्त्वाची पायरी आहे ज्याचे उद्दष्टि भवष्यिसूचक मॉडेल त्यार करण्यासाठी डेटासेटमधून सर्वात संबंधति वैशष्टि्ये (व्हेरएिबल्स कवि वशिषता) ओळखणे आहे. केवळ सर्वात माहतिीपूर्ण वैशष्टि्ये नविडून, आम्ही मॉडेल अचूकता सुधारू शकतो, संगणकीय जटलिता कमी करू शकतो, मॉडेलची व्याख्या वाढवू शकतो आणि ओव्हरफटिगि टाळू शकतो. या प्रक्रयिेमध्ये प्रत्येक वैशष्टि्याचे महत्त्व मूल्यमापन करणे आणि मॉडेल कार्यप्रदर्शन ऑप्टमिाइझ करणारा उपसंच नविडणे समावष्टि आहे. अनेक पद्धती अस्तत्विात आहेत, प्रत्येकाची ताकद आणि कमकुवतता, आणि सर्वोत्तम नविड अनेकदा वशिष्टि डेटासेट आणि वापरल्या जाणाऱ्या मशीन लर्नगि अल्गोरदिमवर अवलंबून असते.

फलि्टर पद्धती

फलिटर पद्धती ही एक भन्नि तंत्रे आहेत जी त्यांच्यामधील संबंधांचा वचिार न करता स्वतंत्रपणे वैशष्टि्यांच्या प्रासंगकितेचे मूल्यांकन करतात. ते सांख्यकिीय उपायांवर आधारति वैशष्टि्ये रँक करतात आणि या क्रमवारीवर आधारति उपसंच नविडतात. उदाहरणांमध्ये हे समावष्टि आहे: * **ची-स्क्वेअर चाचणी:** स्पष्ट वैशष्टि्ये आणि लक्ष्य व्हेरएिबलमधील अवलंबत्वि मोजते. * **सहसंबंध गुणांक:** संख्यात्मक वैशष्टि्ये आणि लक्ष्य व्हेरएिबलमधील रेखीय संबंध मोजतो. * **म्युचुयअल माहतिी:** व्हेरएिबलच्या प्रकाराकडे दुर्लक्ष करून, एक वैशष्टि्य दुसऱ्याबद्दल प्रदान करते त्या माहतिीचे प्रमाण मोजते. **उदाहरण:** वय, उत्पन्न आणि वापराच्या वारंवारतेवर आधारति ग्राहक मंथन (होय/नाही) बद्दल अंदाज लावण्याची कल्पना करा. फलिटर पद्धत मंथनाशी त्याच्या मजबूत सहसंबंधाच्या आधारावर वापर वारंवारता सर्वात महत्त्वाची वैशष्टि्ये म्हणून रँक करू शकते, तर वय आणि उत्पन्न कमी संबंधति मानले जाऊ शकते.

आवरण पद्धती

मॉडेल त्या वैशष्टि्यांसह कती चांगले कार्य करते यावर आधारति रॅपर पद्धती वैशष्टि्यांच्या उपसंचांचे मूल्यांकन करतात. वेगवेगळ्या वैशष्टि्यांच्या उपसंचांच्या कार्यप्रदर्शनाचे मूल्यांकन करण्यासाठी ते मशीन लर्नगि अल्गोरदिम वापरतात. हे संगणकीयदृष्ट्या महाग आहे परंतु फलिटर पद्धतीपेक्षा चांगले वैशष्टि्य नविडू शकते. * **रकिर्सव्हि फीचर एलमिनिेशन ():** मॉडेलमधून त्यांच्या महत्त्वाच्या स्कोअरवर आधारति वैशष्टि्ये पुनरावृत्तीने काढून टाकते. * **फॉरवर्ड सलिक्शन:** कोणत्याही वैशष्टि्यांशविाय सुरू होते आणि मॉडेलचे कार्यप्रदर्शन सुधारणारे वैशष्टि्य पुनरावृत्तीने जोडते. * **बॅकवर्ड एलमिनिेशन:** सर्व वैशष्टि्यांसह प्रारंभ होते आणि मॉडेलच्या कार्यक्षमतेवर कमीत कमी

परिणाम करणारे वैशिष्ट्य पुनरावृत्तीने काढून टाकते. **उदाहरण:** ग्राहक मंथनाचा अंदाज लावण्यासाठी सपोर्ट वेक्टर मशीन () वापरून, सर्व वैशिष्ट्यांसह (वय, उत्पन्न, वापर वारंवारता) सुरू करू शकते आणि अचूकता वाढवणारा इष्टतम उपसंच सापडत नाही तोपर्यंत कमिान महत्त्वाची वैशिष्ट्ये पुनरावृत्तीने काढून टाकू शकतात. .

एम्बेडेड पद्धती

एम्बेडेड पद्धती मॉडेल प्रशिक्षण प्रक्रियेचा भाग म्हणून वैशिष्ट्य निवड समाविष्ट करतात. या पद्धती बहुधा रॅपर पद्धतींपेक्षा अधिक कार्यक्षम असतात आणि फिल्टर पद्धती चुकलेल्या वैशिष्ट्यपूर्ण परस्परसंवाद शोधू शकतात. * **1 रेग्युलरायझेशन ():** मॉडेलच्या लॉस फंक्शनमध्ये पेनल्टी जोडते जे कमी महत्त्वाच्या वैशिष्ट्यांचे गुणांक शून्यावर संकुचित करते, वैशिष्ट्यांची निवड प्रभावीपणे करते. * **निर्णय वृक्ष-आधारित पद्धती:** निर्णय वृक्ष नैसर्गिकरित्या वैशिष्ट्यांची निवड करतात ज्यामुळे झाडामध्ये चांगले विभाजन होते. **उदाहरण:** घराच्या किमतींचा अंदाज लावण्यासाठी रेखीय प्रतिगमन मॉडेलचे प्रशिक्षण देताना, कमी परिणामकारक (उदा. खिडक्यांची संख्या) दुर्लक्ष करून आपोआप केवळ सर्वात संबंधित वैशिष्ट्ये (उदा. चौरस फुटेज, स्थान) निवडू शकते.

वैशिष्ट्य महत्त्व आणि रँकिंग

वापरलेल्या पद्धतीकडे दुर्लक्ष करून, वैशिष्ट्याचे महत्त्व समजून घेणे महत्त्वाचे आहे. यामध्ये लक्ष्य व्हेरिएबलचा अंदाज लावण्यात कोणती वैशिष्ट्ये सर्वात प्रभावशाली आहेत हे समजून घेण्यासाठी वैशिष्ट्य निवड पद्धतीद्वारे तयार केलेल्या स्कोअर किंवा रँकिंगचा अर्थ लावणे समाविष्ट आहे. व्हिज्युअलायझेशन तंत्र, जसे की बार चार्ट किंवा हीटमॅप, हे परिणाम संप्रेषण करण्यासाठी उपयुक्त ठरू शकतात.

पद्धत निवडण्यासाठी विचार

वैशिष्ट्य निवड पद्धतीची निवड अनेक घटकांवर अवलंबून असते: * **डेटासेटचा आकार:** मोठ्या डेटासेटसाठी त्यांच्या संगणकीय कार्यक्षमतेमुळे फिल्टर पद्धतींना प्राधान्य दिले जाते. * ** वैशिष्ट्यांची संख्या: ** रॅपर पद्धती कमी वैशिष्ट्यांसह डेटासेटसाठी अधिक योग्य असू शकतात. * **संगणकीय संसाधने:** एम्बेडेड पद्धती संगणकीय कार्यक्षमता आणि कार्यप्रदर्शन यांच्यात चांगले संतुलन देतात. * **व्याख्याक्षमता:** फिल्टर पद्धती बहुधा रॅपर पद्धतींपेक्षा अधिक व्याख्या करण्यायोग्य असतात.

Reference:

https://scikit-learn.org/stable/modules/feature_selection.html

https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e

Video Links:

https://www.youtube.com/watch?v=jm7TYGv32zs

https://www.youtube.com/watch?v=EqLBAmtKMnQ

https://www.youtube.com/watch?v=73SEn4TaCxs

https://www.youtube.com/watch?v=hCwTDTdYirg

https://www.youtube.com/watch?v=LTE7YbRexl8

https://www.youtube.com/watch?v=PD6xPC--yeA

https://m.youtube.com/watch?v=5bHpPQ6_OU4&t;=0s

https://www.youtube.com/watch?v=za1aA9U4kbI

https://www.instagram.com/rajistics/reel/DAMJUm4SqxT/

https://www.youtube.com/watch?v=eciABhWBiUc

आयाम कमी करण्याचे तंत्र

डायमेंशनॅलटिी रडिक्शन हे मशीन लर्नगिमधील एक महत्त्वपूर्ण तंत्र आहे ज्याचा उपयोग डेटासेटमधील व्हेरएिबल्सची संख्या (वैशष्ट्ये) कमी करण्यासाठी महत्त्वाची माहिती जतन करण्यासाठी केला जातो. उच्च-आयामी डेटामुळे मतिीयतेचा शाप (वाढलेली संगणकीय कमित, ओव्हरफटिगि आणि व्हजि्युअलायझेशनमध्ये अडचण) सारखी आव्हाने येऊ शकतात. डायमेन्शनॅलटिी रडिक्शन डेटाचे नम्निन-आयामी जागेत रूपांतर करून, वशि्लेषण सुलभ करून आणि मॉडेल कार्यप्रदर्शन सुधारून या समस्यांचे नरिाकरण करते. यामध्ये कमी-आयामी प्रतनिधित्वि शोधणे समावष्टि आहे जे मूळ डेटाची आवश्यक वैशष्ट्ये कॅप्चर करते.

रेखीय पद्धती

रेखीय आयाम कमी करण्याचे तंत्र वैशष्ट्यांमधील एक रेषीय संबंध गृहीत धरते. मुख्य घटक वशि्लेषण () ही सर्वात लोकप्रयि पद्धत आहे. हे ऑर्थोगोनल मुख्य घटक ओळखते जे डेटामधील कमाल भनि्नता कॅप्चर करतात. रेखीय भेदभाव वशि्लेषण () डेटामधील ववििधि वर्गांमधील पृथक्करण जास्तीत जास्त करण्यावर लक्ष केंद्रति करते. या पद्धती संगणकीयदृष्ट्या कार्यक्षम आहेत परंतु अ-रेखीय संबंध असलेल्या डेटासाठी योग्य नसू शकतात.

नॉन-लाइनर पद्धती

रेषीय पद्धती चुकू शकतील अशा डेटामधील जटलि संबंध कॅप्चर करण्यास नॉन-लाइनर पद्धती सक्षम आहेत. टी-डसि्ट्रबि्युटेड स्टोकास्टकि नेबर एम्बेडगि (-) हे खालच्या आयामांमध्ये उच्च-आयामी डेटाचे दृश्यमान करण्यासाठी एक लोकप्रयि तंत्र आहे, वशिेषत: अन्वेषण डेटा वशि्लेषणासाठी उपयुक्त. इतर पद्धतींचा समावेश आहे , (), आणि (एक न्यूरल नेटवर्क-आधारति दृष्टीकोन). या पद्धती बहुधा रेखीय पद्धतींपेक्षा संगणकीयदृष्ट्या अधकि महाग असतात.

वैशष्ट्य नविड

वैशष्ट्य नविड हे एक आयाम कमी करण्याचे तंत्र आहे ज्यामध्ये नवीन तयार करण्याऐवजी मूळ वैशष्ट्यांचा उपसंच नविडणे समावष्टि आहे. पद्धतींमध्ये फलि्टर पद्धती (उदा. सहसंबंध वशि्लेषण), रॅपर पद्धती (उदा. पुनरावर्ती वैशष्ट्य नरि्मूलन), आणि एम्बेडेड पद्धती (उदा., रेखीय मॉडेल्समध्ये 1 नयिमतीकरण) समावष्टि आहेत. जेव्हा स्पष्टीकरण महत्त्वाचे असते तेव्हा वैशष्ट्य नविडीला प्राधान्य दलि जाते, कारण ते थेट मूळ वैशष्ट्यांचा वापर करते.

सराव मध्ये आयाम कमी

आयाम कमी करण्याच्या तंत्राची नविड अनेक घटकांवर अवलंबून असते, ज्यात डेटाचे स्वरूप (रेषीय वि. नॉन-रेखीय संबंध), आयाम कमी करण्याची इच्छति पातळी, उपलब्ध संगणकीय संसाधने आणि वशि्लेषणाचे उद्दष्टि (व्हजि्युअलायझेशन, मॉडेल सुधारणा) यांचा समावेश होतो. , इ.). अनेक तंत्रांसह प्रयोग करणे आणि योग्य मेट्रकि्स वापरून त्यांच्या कार्यक्षमतेचे मूल्यमापन करणे अनेकदा फायदेशीर ठरते. उदाहरणार्थ, इमेज प्रोसेसगिमध्ये, चा वापर वर्गीकरणापूर्वी वैशष्ट्य काढण्यासाठी केला जाऊ शकतो, तर - चा वापर प्रतमिांच्या क्लस्टर्सची कल्पना करण्यासाठी केला जाऊ शकतो.

उदाहरण: इमेज कॉम्प्रेशनसाठी

कल्पना करा की तुमच्याकडे उच्च-आयामी व्हेक्टर (प्रत्येक पिक्सेल एक वैशिष्ट्य आहे) म्हणून प्रस्तुत केलेल्या प्रतिमांचा डेटासेट आहे. या वेक्टर्सची परिमाण कमी करण्यासाठी लागू केले जाऊ शकते. मुख्य घटक प्रतिमांमधील सर्वात महत्त्वाची विविधता कॅप्चर करतात, कमी परिमाणांसह (कमी वैशिष्ट्यांसह) प्रतिमांची पुनर्रचना करण्यास परवानगी देतात, बहुतेक दृश्य माहिती राखून ठेवत प्रतिमा प्रभावीपणे संकुचित करतात.

Reference:

https://scikit-learn.org/stable/modules/decomposition.html

https://distill.pub/2016/misread-tsne/

https://www.cs.cmu.edu/~efros/courses/36-463/lectures/lecture10.pdf

Video Links:

https://www.youtube.com/watch?v=ioXKxulmwVQ

https://www.youtube.com/watch?v=6XGlqR6rcpU

https://www.youtube.com/watch?v=jc1_yPYmspk

https://www.youtube.com/watch?v=THu9yHnpq9I

https://www.youtube.com/watch?v=UgOHupaIfcA

https://www.youtube.com/watch?v=embks9p4pb8

https://www.youtube.com/watch?v=ZqXnPcyIAL8

https://www.youtube.com/watch?v=ne6vnKoTHwk

https://www.youtube.com/watch?v=SsYhTPkRdLA

https://www.youtube.com/watch?v=cowHdW2-RkU

वैशिष्ट्य अभियांत्रिकी धोरणे

फीचर अभियांत्रिकी ही मशीन लर्निंग अल्गोरिदम अधिक चांगले कार्य करणारी वैशिष्ट्ये तयार करण्यासाठी डोमेन ज्ञान वापरण्याची प्रक्रिया आहे. मशीन लर्निंग पाइपलाइनमधील हे एक महत्त्वपूर्ण पाऊल आहे, कारण तुमच्या वैशिष्ट्यांच्या गुणवत्तेचा तुमच्या मॉडेलच्या कार्यप्रदर्शनावर थेट परिणाम होतो. हे उप-मॉड्युल प्रभावी वैशिष्ट्य अभियांत्रिकीसाठी विविध धोरणे शोधेल.

व्याख्या आणि महत्त्व

वैशिष्ट्य अभियांत्रिकीमध्ये कच्च्या डेटाचे वैशिष्ट्यांमध्ये रूपांतर करणे समाविष्ट आहे जे अधिक माहितीपूर्ण आणि मशीन लर्निंग मॉडेलसाठी योग्य आहेत. यामध्ये नवीन वैशिष्ट्ये तयार करणे, विद्यमान असलेले बदलणे किंवा सर्वात संबंधित वैशिष्ट्यांचा उपसंच निवडणे यांचा समावेश असू शकतो. प्रभावी वैशिष्ट्य अभियांत्रिकी मॉडेल अचूकता लक्षणीयरीत्या सुधारू शकते, प्रशिक्षण वेळ कमी करू शकते आणि मॉडेलची व्याख्याक्षमता वाढवू शकते.

वैशिष्ट्य अभियांत्रिकीचे प्रकार

फीचर इंजिनिअरिंगसाठी अनेक तंत्रे आहेत. याचे स्थूलमानाने वर्गीकरण करता येईल: * **वैशिष्ट्य निर्मिती:** विद्यमान वैशिष्ट्यांमधून नवीन वैशिष्ट्ये निर्माण करणे. उदाहरणांमध्ये परस्परसंवाद संज्ञा

तयार करणे (उदा. दोन वैशिष्ट्यांचा गुणाकार करणे), बहुपदी वैशिष्ट्ये (उदा. वैशिष्ट्याचे वर्गीकरण किंवा घन करणे) किंवा तारीख/वेळ माहितीवर आधारित वैशिष्ट्ये (उदा. तारखेपासून आठवड्याचा दिवस काढणे) यांचा समावेश होतो. * **वैशिष्ट्य परिवर्तन:** मशीन लर्निंग अल्गोरिदमसाठी त्यांची योग्यता सुधारण्यासाठी विद्यमान वैशिष्ट्यांमध्ये सुधारणा करणे. सामान्य परिवर्तनांमध्ये स्केलिंग (उदा., मानकीकरण, किमान-कमाल स्केलिंग), सामान्यीकरण आणि एन्कोडिंग वर्गीय चल (उदा. एक-हॉट एन्कोडिंग, लेबल एन्कोडिंग) यांचा समावेश होतो. * **वैशिष्ट्य निवड:** आयाम कमी करण्यासाठी आणि मॉडेल कार्यप्रदर्शन सुधारण्यासाठी सर्वात संबंधित वैशिष्ट्यांचा उपसंच निवडणे. तंत्रांमध्ये फिल्टर पद्धती (उदा. सहसंबंध विश्लेषण), आवरण पद्धती (उदा. पुनरावृत्ती वैशिष्ट्य निर्मूलन), आणि एम्बेडेड पद्धती (उदा., L1 नियमितीकरण) समाविष्ट आहेत.

उदाहरणे

परिस्थिती: घराच्या किमतींचा अंदाज लावणे. * **कच्चा डेटा:** घराचा आकार (), शयनकक्षांची संख्या, स्नानगृहांची संख्या, स्थान (पिन कोड). ***वैशिष्ट्य अभियांत्रिकी:** * **वैशिष्ट्य निर्मिती:** घराच्या आकाराला बेडरूमच्या संख्येने विभाजित करून 'आकार_प्रति_बेडरूम' एक नवीन वैशिष्ट्य तयार करा. हे प्रति शयनकक्ष जागा कॅप्चर करते, संभाव्यत: किमतीतील एक महत्त्वपूर्ण घटक. * **वैशिष्ट्य परिवर्तन:** एक-हॉट स्थान (पिन कोड) एकाधिक बायनरी वैशिष्ट्यांमध्ये एन्कोड करते, प्रत्येक पिन कोडसाठी एक. हे स्थानाचे स्पष्ट स्वरूप हाताळते. * **वैशिष्ट्य निवड:** घराच्या किमतीचा अंदाज लावण्यासाठी सर्वात महत्त्वाची वैशिष्ट्ये ओळखण्यासाठी सहसंबंध विश्लेषण वापरा. मॉडेल सुलभ करण्यासाठी कमी सहसंबंध असलेली वैशिष्ट्ये वगळली जाऊ शकतात. **दुसरे उदाहरण:** टेलिकॉम कंपनीमध्ये ग्राहक मंथन अंदाज. * **कच्चा डेटा:** ग्राहकाचे वय, मासिक बिलाची रक्कम, कॉल कालावधी, डेटा वापर. ***वैशिष्ट्य अभियांत्रिकी:** * **वैशिष्ट्य निर्माण:** सरासरी मासिक कॉल कालावधी आणि सरासरी दैनिक डेटा वापराची गणना करा. * **वैशिष्ट्य परिवर्तन:** मासिक बिलाची रक्कम आणि डेटा वापर 0-1 स्केलवर सामान्य करा. * **वैशिष्ट्य निवड:** मंथन अंदाजासाठी सर्वात प्रभावशाली वैशिष्ट्ये निवडण्यासाठी पुनरावर्ती वैशिष्ट्य निर्मूलनाचा वापर करा.

गहाळ मूल्ये हाताळणे

रिअल-वर्ल्ड डेटासेटमध्ये गहाळ मूल्ये ही एक सामान्य समस्या आहे. त्यांना हाताळण्याच्या धोरणांमध्ये हे समाविष्ट आहे: * **आकलन:** गहाळ मूल्ये अंदाजे मूल्यांसह पुनर्स्थित करणे. पद्धतींमध्ये सरासरी/मध्यवर्ती आरोप, -जवळचे शेजारी आरोप, किंवा मॉडेल-आधारित आरोप समाविष्ट आहेत. * **हटवणे:** गहाळ मूल्यांसह पंक्ती किंवा स्तंभ काढणे. हा एक सोपा दृष्टीकोन आहे परंतु माहितीचे नुकसान होऊ शकते. * **इंडिकेटर व्हेरिएबल:** मूल्य गहाळ आहे की नाही हे दर्शवणारे नवीन बायनरी वैशिष्ट्य तयार करणे.

वैशिष्ट्य स्केलिंग

समान श्रेणीमध्ये वैशिष्ट्ये स्केलिंग करणे अनेकदा आवश्यक असते, विशेषत: वैशिष्ट्यांच्या परिमाणांसाठी संवेदनशील अल्गोरिदमसाठी (उदा. -जवळचे शेजारी, सपोर्ट वेक्टर मशीन). सामान्य स्केलिंग पद्धतींमध्ये हे समाविष्ट आहे: * **मानकीकरण (-स्कोअर सामान्यीकरण):** 1 च्या मानक विचलनासह 0 च्या आसपास डेटा केंद्रित करते. * **कमीत-कमाल स्केलिंग:** वैशिष्ट्ये 0 आणि 1 मधील श्रेणीमध्ये मोजतात.

Reference:

https://www.analyticsvidhya.com/blog/2020/07/feature-engineering-techniques-machine-learning/

https://towardsdatascience.com/feature-engineering-for-machine-learning-3a5e293a5114

https://machinelearningmastery.com/discover-feature-engineering-how-to-engineer-features-and-how-to-get-good-at-it/

Video Links:

https://www.youtube.com/watch?v=WElBhXr9B7c&pp;=ygUOI2ZlYXR1cmVmdXNpb24%3D

https://www.youtube.com/watch?v=vCSZWM4y2EU

https://www.youtube.com/watch?v=BFXJqJtNivY

https://www.youtube.com/watch?v=FUB1KlhqH58

https://www.youtube.com/watch?v=vsKNxbP8R_8

https://www.youtube.com/watch?v=GduT2ZCc26E

https://www.youtube.com/watch?v=rf5dGtn4Nkk

https://www.youtube.com/watch?v=4w-S6Hi1mA4

https://www.youtube.com/watch?v=C0_bh_5C5ro

https://www.youtube.com/watch?v=ZT9AG9WgGxg