***ما هو تحليل المكونات الأساسية؟***

تحليل المكون الرئيسي ، أو PCA ، هو طريقة لتقليل الأبعاد تستخدم غالبًا لتقليل أبعاد مجموعات البيانات الكبيرة ، عن طريق تحويل مجموعة كبيرة من المتغيرات إلى مجموعة أصغر لا تزال تحتوي على معظم المعلومات في المجموعة الكبيرة.

إن تقليل عدد متغيرات مجموعة البيانات يأتي بشكل طبيعي على حساب الدقة ، ولكن الحيلة في تقليل الأبعاد تكمن في مقايضة القليل من الدقة من أجل البساطة. لأن مجموعات البيانات الأصغر يسهل استكشافها وتصورها وتجعل تحليل البيانات أسهل بكثير وأسرع لخوارزميات التعلم الآلي بدون متغيرات خارجية للمعالجة.

باختصار ، فكرة PCA بسيطة - تقليل عدد متغيرات مجموعة البيانات ، مع الحفاظ على أكبر قدر ممكن من المعلومات.

***متى يجب استخدام PCA؟***

هل تريد تقليل عدد المتغيرات ، لكنك غير قادر على تحديد المتغيرات لإزالتها تمامًا من الاعتبار؟

هل تريد التأكد من أن المتغيرات الخاصة بك مستقلة عن بعضها البعض؟

هل أنت مرتاح لجعل متغيراتك المستقلة أقل قابلية للتفسير؟

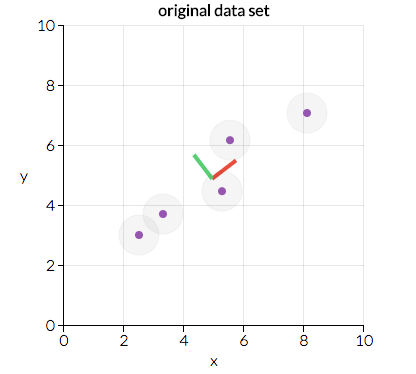
إذا أجبت بـ "نعم" على جميع الأسئلة الثلاثة ، فإن PCA هي طريقة جيدة للاستخدام. إذا كانت إجابتك "لا" على السؤال 3 ، فلا يجب عليك استخدام PCA.

***كيف يعمل PCA؟***

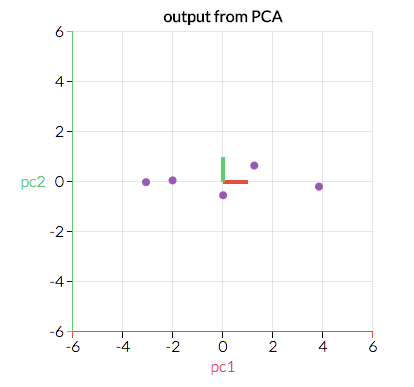
يناقش القسم التالي سبب عمل PCA

سنقوم بحساب مصفوفة تلخص كيفية ارتباط جميع المتغيرات ببعضها البعض.

ثم سنقسم هذه المصفوفة إلى مكونين منفصلين: الاتجاه والحجم. يمكننا بعد ذلك فهم "اتجاهات" بياناتنا و "حجمها" (أو مدى "أهمية" كل اتجاه). تعرض لقطة الشاشة أدناه ، من التطبيق الصغير setosa.io ، الاتجاهين الرئيسيين في هذه البيانات: "الاتجاه الأحمر" و "الاتجاه الأخضر". في هذه الحالة ، "الاتجاه الأحمر" هو الأكثر أهمية. سنتطرق إلى سبب حدوث ذلك لاحقًا ، ولكن بالنظر إلى كيفية ترتيب النقاط ، هل يمكنك معرفة سبب ظهور "الاتجاه الأحمر" أكثر أهمية من "الاتجاه الأخضر؟"



سنقوم بتحويل بياناتنا الأصلية لتتماشى مع هذه الاتجاهات المهمة (وهي مجموعات من المتغيرات الأصلية). لقطة الشاشة أدناه (مرة أخرى من setosa.io) هي نفس البيانات الدقيقة الموضحة أعلاه ، ولكن تم تحويلها بحيث أصبح المحاور x و y الآن "الاتجاه الأحمر" و "الاتجاه الأخضر".

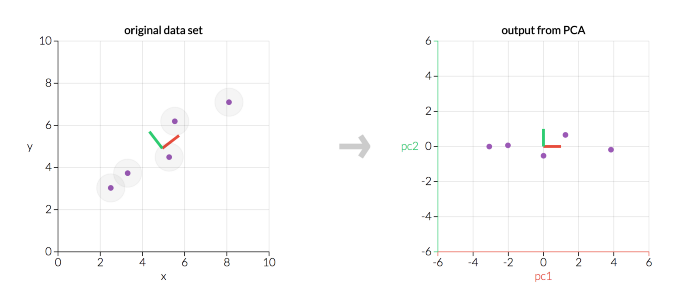


في حين أن المثال المرئي هنا ثنائي الأبعاد (وبالتالي لدينا "اتجاهان") ، فكر في حالة تكون فيها بياناتنا ذات أبعاد أكثر. من خلال تحديد "الاتجاهات" الأكثر "أهمية" ، يمكننا ضغط بياناتنا أو عرضها في مساحة أصغر بإسقاط "الاتجاهات" "الأقل أهمية". من خلال إسقاط بياناتنا في مساحة أصغر ، فإننا نحد من أبعاد فضاء الميزات لدينا ... ولكن نظرًا لأننا قمنا بتحويل بياناتنا في هذه "الاتجاهات" المختلفة ، فقد حرصنا على الاحتفاظ بجميع المتغيرات الأصلية في نموذجنا!

هنا ، أسير عبر خوارزمية لإجراء PCA. أحاول أن أتجنب أن أكون تقنيًا للغاية ، لكن من المستحيل تجاهل التفاصيل هنا ، لذلك هدفي هو استعراض الأمور بأكبر قدر ممكن من الوضوح. يتم تقديم حدس أعمق عن سبب عمل الخوارزمية في القسم التالي.

قبل البدء ، يجب أن تكون لديك بيانات مجدولة منظمة مع عدد n من الصفوف ومن المحتمل أن تكون p + 1 أعمدة ، حيث يتوافق عمود واحد مع المتغير التابع (عادةً ما يُشار إليه Y) وأعمدة p حيث يتوافق كل منها مع متغير مستقل (يُشار إلى المصفوفة عادةً X).

1. إذا كان متغير Y موجودًا وكان جزءًا من بياناتك ، فافصل بياناتك إلى Y و X ، كما هو محدد أعلاه - سنعمل في الغالب مع X. (ملاحظة: إذا لم يكن هناك عمود لـ Y ، فلا بأس بذلك - انتقل إلى النقطة التالية!)
2. خذ مصفوفة المتغيرات المستقلة X ، ولكل عمود ، اطرح متوسط ​​هذا العمود من كل إدخال. (هذا يضمن أن كل عمود له متوسط ​​صفر.)
3. قرر ما إذا كنت تريد التوحيد أم لا. بالنظر إلى أعمدة X ، هل الميزات ذات التباين الأعلى أكثر أهمية من الميزات ذات التباين الأقل ، أم أن أهمية الميزات مستقلة عن التباين؟ (في هذه الحالة ، تعني الأهمية مدى جودة توقع هذه الميزة لـ Y.) إذا كانت أهمية الميزات مستقلة عن تباين الميزات ، فقم بتقسيم كل ملاحظة في عمود على الانحراف المعياري لهذا العمود. (هذا ، جنبًا إلى جنب مع الخطوة 2 ، يوحد كل عمود من X للتأكد من أن كل عمود يعني الصفر والانحراف المعياري 1.) استدع المصفوفة المركزية (وربما الموحدة) Z.
4. خذ المصفوفة Z ، بدّلها ، واضرب المصفوفة المنقولة في Z. (بكتابة ذلك رياضيًا ، سنكتب هذا كـ ZᵀZ.) المصفوفة الناتجة هي مصفوفة التغاير لـ Z ، حتى ثابت.
5. (ربما تكون هذه هي الخطوة الأصعب التي يجب اتباعها - ابق معي هنا.) احسب متجهات eigenvectors وقيمها الذاتية المقابلة لـ ZᵀZ. يتم إجراء ذلك بسهولة في معظم حزم الحوسبة - في الواقع ، فإن التكوين eigendecomposition لـ ZᵀZ هو المكان الذي نحلل فيه ZᵀZ إلى PDP⁻¹ ، حيث P هي مصفوفة المتجهات الذاتية و D هي المصفوفة القطرية ذات القيم الذاتية على القطر وقيم الصفر في كل مكان آخر. سيتم ربط قيم eigenvalues ​​على قطري D مع العمود المقابل في P - أي أن العنصر الأول من D هو λ₁ والمتجه الذاتي المقابل هو العمود الأول من P. وهذا ينطبق على جميع العناصر في D والمتجهات الذاتية المقابلة لها في P. سنكون دائمًا قادرين على حساب PDP⁻¹ بهذه الطريقة. (مكافأة: للمهتمين ، يمكننا دائمًا حساب PDP⁻¹ بهذه الطريقة لأن ZᵀZ عبارة عن مصفوفة شبه متماثلة موجبة.)
6. خذ قيم eigenvalues ​​λ₁ و λ₂ و ... و p وفرزها من الأكبر إلى الأصغر. عند القيام بذلك ، قم بفرز المتجهات الذاتية في P وفقًا لذلك. (على سبيل المثال ، إذا كانت λ₂ هي أكبر قيمة ذاتية ، فاخذ العمود الثاني من P وضعه في موضع العمود الأول.) اعتمادًا على حزمة الحوسبة ، يمكن القيام بذلك تلقائيًا. قم باستدعاء هذه المصفوفة المرتبة للمتجهات الذاتية P \*. (يجب أن تكون أعمدة P \* مماثلة لأعمدة P ، ولكن ربما بترتيب مختلف). لاحظ أن هذه المتجهات الذاتية مستقلة عن بعضها البعض.
7. احسب Z \* = ZP \*. هذه المصفوفة الجديدة ، Z \* ، هي نسخة مركزية / موحدة من X ولكن الآن كل ملاحظة عبارة عن مزيج من المتغيرات الأصلية ، حيث يتم تحديد الأوزان بواسطة المتجه الذاتي. على سبيل المكافأة ، نظرًا لأن المتجهات الذاتية في P \* مستقلة عن بعضها البعض ، فإن كل عمود من Z \* مستقل أيضًا عن بعضه البعض!



لاحظ شيئين في هذا الرسم البياني:

يعرض المخططان نفس البيانات بالضبط ، لكن الرسم البياني الصحيح يعكس البيانات الأصلية التي تم تحويلها بحيث أصبحت محاورنا الآن المكونات الرئيسية.

في كلا الرسمين البيانيين ، تكون المكونات الأساسية متعامدة مع بعضها البعض. في الواقع ، سيكون كل مكون رئيسي دائمًا متعامدًا (ويعرف أيضًا باسم المصطلح الرياضي الرسمي للعمودي) على كل مكون رئيسي آخر. (ألا تصدقني؟ حاول كسر التطبيق الصغير!)

نظرًا لأن مكوناتنا الرئيسية متعامدة مع بعضها البعض ، فهي مستقلة إحصائيًا عن بعضها البعض ... وهذا هو السبب في أن أعمدة Z \* مستقلة خطيًا عن بعضها البعض!

1. أخيرًا ، نحتاج إلى تحديد عدد الميزات التي يجب الاحتفاظ بها مقابل عدد الميزات التي يجب إسقاطها. هناك ثلاث طرق شائعة لتحديد ذلك ، تمت مناقشتها أدناه وتتبعها مثال واضح:

الطريقة الأولى: نختار بشكل تعسفي عدد الأبعاد التي نريد الاحتفاظ بها. ربما أرغب في تمثيل الأشياء بصريًا في بعدين ، لذلك يمكنني الاحتفاظ بميزتين فقط. هذا يعتمد على حالة الاستخدام ولا توجد قاعدة صارمة وسريعة لعدد الميزات التي يجب علي اختيارها.

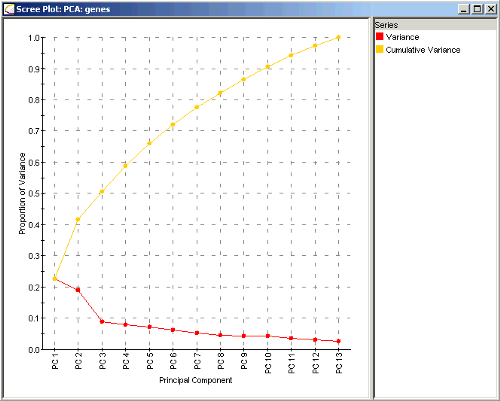
الطريقة الثانية: احسب نسبة التباين الموضحة (الموضحة بإيجاز أدناه) لكل ميزة ، واختر حدًا ، وأضف ميزات حتى تصل إلى هذا الحد. (على سبيل المثال ، إذا كنت تريد شرح 80٪ من إجمالي التباين الذي ربما يفسره نموذجك ، فأضف ميزات ذات أكبر نسبة تباين موضحة حتى توضح نسبة التباين النتائج أو تتجاوز 80٪.)

الطريقة الثالثة: ترتبط ارتباطًا وثيقًا بالطريقة الثانية. احسب نسبة التباين الموضحة لكل ميزة وفرز الميزات حسب نسبة التباين الموضحة ورسم النسبة التراكمية للتباين الموضحة أثناء الاحتفاظ بمزيد من الميزات. (يُطلق على هذه المؤامرة اسم مؤامرة scree ، كما هو موضح أدناه.) يمكن للمرء اختيار عدد الميزات المراد تضمينها من خلال تحديد النقطة التي يكون فيها إضافة ميزة جديدة لها انخفاض كبير في التباين الموضح بالنسبة إلى الميزة السابقة ، واختيار الميزات حتى تلك النقطة . (أسمي هذا أسلوب "العثور على الكوع" ، حيث إن النظر إلى "الانحناء" أو "الكوع" في مخطط الحصاة يحدد مكان حدوث أكبر انخفاض في نسبة التباين الموضحة.)

نظرًا لأن كل قيمة ذاتية هي تقريبًا أهمية ناقلها الذاتي المقابل ، فإن نسبة التباين الموضحة هي مجموع قيم eigenvalues ​​للميزات التي احتفظت بها مقسومة على مجموع قيم eigenvalues ​​لجميع الميزات.

ضع في اعتبارك هذه الصورة للحصول على البيانات الجينية. (المصدر: هنا.) يشير الخط الأحمر إلى نسبة التباين التي توضحها كل ميزة ، والتي يتم حسابها بأخذ القيمة الذاتية للمكون الرئيسي مقسومة على مجموع كل قيم eigenvalues. نسبة التباين التي يتم شرحها من خلال تضمين المكون الأساسي 1 فقط هي λ₁ / (+ λ₂ + ... + p) ، وهي حوالي 23٪. نسبة التباين الموضحة بتضمين المكون الرئيسي 2 فقط هي λ₂ / (+ λ₂ + ... + p) ، أو حوالي 19٪.

نسبة التباين الموضحة بتضمين كل من المكونين الرئيسيين 1 و 2 هي (λ₁ + λ₂) / (λ₁ + λ₂ + ... + p) ، وهي حوالي 42٪. هذا هو المكان الذي يأتي فيه الخط الأصفر ؛ يشير الخط الأصفر إلى النسبة التراكمية للتباين الموضحة إذا قمت بتضمين جميع المكونات الأساسية حتى تلك النقطة. على سبيل المثال ، تشير النقطة الصفراء فوق PC2 إلى أن تضمين المكونين الرئيسيين 1 و 2 سيوضح حوالي 42٪ من إجمالي التباين في النموذج.



الآن دعنا نستعرض بعض الأمثلة:

الطريقة الأولى: نختار بشكل تعسفي عددًا من المكونات الأساسية لتضمينها. لنفترض أنني أردت الاحتفاظ بخمسة مكونات رئيسية في نموذجي. في حالة البيانات الجينية أعلاه ، تشرح هذه المكونات الخمسة الرئيسية حوالي 66 ٪ من التباين الإجمالي الذي يمكن تفسيره من خلال تضمين جميع المكونات الرئيسية الثلاثة عشر.

الطريقة الثانية: لنفترض أنني أردت تضمين ما يكفي من المكونات الأساسية لشرح 90٪ من التباين الإجمالي الموضح بواسطة جميع المكونات الأساسية الثلاثة عشر. في حالة البيانات الجينية أعلاه ، أود تضمين المكونات الرئيسية العشرة الأولى وإسقاط المتغيرات الثلاثة الأخيرة من Z \*.

الطريقة الثالثة: هنا ، نريد "العثور على الكوع". في مخطط الحصيلة أعلاه ، نرى أن هناك انخفاضًا كبيرًا في نسبة التباين الموضحة بين المكون الرئيسي 2 والمكون الرئيسي 3. في هذه الحالة ، من المحتمل أن نقوم بتضمين الميزتين الأوليين وإسقاط الميزات المتبقية. كما ترى ، فإن هذه الطريقة ذاتية بعض الشيء لأن "الكوع" ليس لها تعريف دقيق رياضيًا ، وفي هذه الحالة ، سنقوم بتضمين نموذج يشرح فقط حوالي 42٪ من إجمالي التباين.

(ملاحظة: سيكون لبعض المخططات الحصرية حجم المتجهات الذاتية على المحور Y بدلاً من نسبة التباين. يؤدي هذا إلى نتائج مكافئة ، ولكنه يتطلب من المستخدم حساب نسبة التباين يدويًا. يمكن مشاهدة مثال على ذلك هنا. )

بمجرد أن نتخلى عن المتغيرات المحولة التي نريد إسقاطها ، نكون قد انتهينا! هذا هو PCA.

***لماذا تعمل pca ?***

ي حين أن PCA هي طريقة تقنية للغاية تعتمد على خوارزميات الجبر الخطي المتعمقة ، فهي طريقة بديهية نسبيًا عندما تفكر في الأمر.

أولاً ، مصفوفة التغاير ZᵀZ عبارة عن مصفوفة تحتوي على تقديرات لكيفية ارتباط كل متغير في Z بكل متغير آخر في Z. يعتبر فهم كيفية ارتباط أحد المتغيرات بآخر أمرًا قويًا للغاية.

ثانياً ، القيم الذاتية والمتجهات الذاتية مهمة. المتجهات الذاتية تمثل الاتجاهات. فكر في رسم بياناتك على مخطط مبعثر متعدد الأبعاد. ثم يمكن للمرء أن يفكر في المتجه الذاتي الفردي على أنه "اتجاه" معين في مخطط تبعثر البيانات. تمثل القيم الذاتية الحجم أو الأهمية. ترتبط القيم الذاتية الأكبر بالاتجاهات الأكثر أهمية.

أخيرًا ، نفترض أن المزيد من التباين في اتجاه معين يرتبط بتفسير سلوك المتغير التابع. يشير الكثير من التباين عادةً إلى الإشارة ، بينما يشير التباين القليل عادةً إلى الضوضاء. وبالتالي ، كلما زاد التباين في اتجاه معين ، يكون ، من الناحية النظرية ، مؤشرًا على شيء مهم نريد اكتشافه. (يُعد التطبيق الصغير setosa.io PCA طريقة رائعة للتلاعب بالبيانات وإقناع نفسك بالسبب المنطقي.)

وبالتالي ، فإن PCA هي طريقة تجمع بين:

1. مقياس لكيفية ارتباط كل متغير ببعضه البعض. (مصفوفة التغاير.)
2. الاتجاهات التي تتشتت فيها بياناتنا. (المتجهات الذاتية.)
3. الأهمية النسبية لهذه الاتجاهات المختلفة. (القيم الذاتية).

يجمع PCA بين المتنبئين لدينا ويسمح لنا بإسقاط المتجهات الذاتية غير المهمة نسبيًا

***المصادر***

***التطبيق الصغير PCA من Setosa.io. (برنامج صغير يسمح لك بتصور المكونات الرئيسية وكيف تؤثر بياناتك على المكونات الأساسية.)***

* ***An Introduction to Statistical Learning, 6th printing, by James, Witten, Hastie, and Tibshirani. (PCA is covered extensively in chapters 6.3, 6.7, and 10.2. This book assumes knowledge of linear regression but is pretty accessible, all things considered.)***
* ***Notes from Penn State’s STAT 505 (Applied Multivariate Statistical Analysis) Course. (I’ve found Penn State’s online statistics course notes to be incredible, and the PCA section here is particularly helpful.)***
* ***Linear Algebra and Its Applications, 4th edition, by David Lay. (PCA is covered in chapter 7.5.)***
* ***A Tutorial on Principal Components Analysis, by Jonathon Shlens at Google Research.***
* ***A draft chapter on Principal Component Analysis from Cosma Shalizi of Carnegie Mellon University.***
* ***A chapter on data preprocessing from Applied Predictive Modeling includes an introductory discussion of principal component analysis (with visuals!) in Section 3.3. (h/t to Jay Lucas for the recommendation!)***
* ***Elements of Statistical Learning, 10th printing, by Hastie, Tibshirani, and Friedman. (PCA is covered extensively in chapters 3.5, 14.5, and 18.6. This book assumes knowledge of linear regression, matrix algebra, and calculus and is significantly more technical than An Introduction to Statistical Learning, but the two follow a similar structure given the common authors.)***