UDACITY

PROJETO

Creating Customer Segments

Uma parte do Machine Learning Engineer Nanodegree Program

REVISÃO DO PROJETO

REVISÃO DE CÓDIGO

COMENTÁRIOS

COMPARTILHE SUA REALIZAÇÃO!

Requires Changes

3 ESPECIFICAÇÕES NECESSITAM DE MUDANÇAS

Parabéns por uma ótima primeira versão do seu projeto! Peço especial atenção à revisão da resposta à questão sobre componentes principais - rever os vídeos sobre o assunto pode ser uma boa ideia. As outras revisões necessárias são menores, e acredito que a próxima versão do seu projeto será a última. Bom trabalho e mantenha o ritmo!

Exploração dos Dados

Três amostras diferentes dos dados são escolhidos e o que elas representam é proposto com base na descrição estatística dos dados.

Excelente

Ótimo trabalho incluindo visualizações para informar e incrementar sua resposta à **Questão 1**.

Sugestão

Aqui é uma boa ideia ser mais específico nas observações sobre cada cliente-exemplo, usando as estatísticas do conjunto de dados como um todo. Por exemplo: em vez de escrever que "o cliente gasta muito mais do que a média com o produto X", você poderia escrever que "os gastos do cliente com o produto X são maiores do que o terceiro quartil dos clientes como um todo".

A pontuação do atributo removido foi corretamente calculada. A resposta justifica se o atributo removido é relevante.

Excelente

> Por Detergents_Paper e Grocery apresentarem o melhor coeficiente de determinação imaginei que seriam os melhores atributos a serem utilizados na identificação dos hábitos de consumo. Porêm depois disso entendi que como eles serão os mais faceis de prever com o balanço dos dados, então são os menos necessarios no momento da identificação dos hábitos de consumo.

Exato! Ótima revisão do seu entendimento do significado de uma boa pontuação nessa situação. :)

Comentário

É um pouco confuso a variável `detergents_array` ser a coluna `Grocery`. :)

Sugestão

Você poderia usar um loop para calcular o R^2 para todos os atributos!

Atributos correlacionados são corretamente identificados e comparados ao atributo previsto. A distribuição dos dados para esses atributos é discutida.

Alteração necessária

Ficou faltando a resposta para o último item da questão: como os dados de cada *feature* são distribuídos? Trata-se de uma distribuição normal?

Duas dicas adicionais:

- Os dados apresentam assimetria?
- Pode ser interessante pensar em distribuições log-normais. :)

Sugestão

Acho que uma boa ferramenta para visualizar as correlações entre variáveis é um mapa de calor anotado:

```
import seaborn as sns
sns.heatmap(data.corr(), annot=True);
```

Pré-processamento dos Dados

Os valores aberrantes extremos são identificados, e discute-se se eles deveriam ser removidos. A decisão de remover quaisquer dados é corretamente justificada.

Excelente

Ótimo trabalho identificando programaticamente as observações que são *outliers* para mais de um atributo. Parabéns!

Comentário

- No contexto de uma análise como *clustering* ou PCA, remover *outliers* pode ser particularmente benéfico.
- A presença de *outliers* pode levar a atributos com uma variância significativamente maior, o que pode afetar os resultados de uma análise de componente principal.
- *Outliers* também podem tornar mais difícil o trabalho de encontrar *clusters* nos dados - dependendo do algoritmo, pode-se terminar com um *cluster* que inclua apenas um *outlier*!

O código de dimensionamento de atributos tanto para os dados quanto para as amostras foi corretamente implementado.

Transformação de Atributos

O código do PCA foi corretamente implementado e aplicado tanto para os dados dimensionados quanto para as amostras dimensionadas no caso bidimensional.

A variância explicada total para duas e quatro dimensões dos dados do PCA é corretamente relatada. As primeiras quatro dimensões são interpretadas como uma representação dos gastos do cliente com justificativa.

Alteração necessária

Na **Questão 5**, é necessário analisar cada componente principal individualmente, considerando os pesos de cada atributo. Tenha em mente que o sinal dos pesos pode ser invertido sem afetar o significado do componente principal: o que interessa é a magnitude dos pesos e os sinais comparativos.

- Atributos com peso positivo elevado terão valor alto caso o componente principal tenha um valor alto.
- Atributos com peso negativo elevado terão valor baixo caso o componente principal tenha um valor alto.
- Atributos com peso próximo de zero não afetarão significativamente o componente principal.
- Comparando dois atributos:
    - Se ambos apresentarem pesos elevados com o mesmo sinal, os atributos são positivamente correlacionados *para observações com valor elevado desse componente principal*
    - Se ambos apresentarem pesos elevados com o sinal invertido, os atributos são negativamente correlacionados *para observações com valor elevado desse componente principal*

Levando isso em conta, analise cada componente principal individualmente, e busque também identificar *tipos de consumidores* que apresentariam valores significativos para cada componente principal. Por exemplo, que tipo de cliente você acha que teria valores elevados para o primeiro componente principal?

> Acredito que a dimensão que melhor representa as despesas dos clientes é a 4, pois dentre as 6 dimensões é a que apresenta a maior quantidade de produtos como positivo em Feature Weight.

Lembre-se de que a variância explicada por componente principal diminui à medida que componentes principais são acrescentados - ou seja, o primeiro componente principal é o que mais explica a variação de valores no conjunto de dados original.

Clustering

Os algoritmos GMM e K-Means são comparados em detalhes. A escolha do aluno é justificada com base nas características do algoritmo e dos dados.

Alteração necessária

Uma diferença importante entre os dois algoritmos não foi mencionada na sua resposta: como eles lidam com a **incerteza**, ou com observações próximas da "fronteira" entre *clusters*?

Amostras dos dados são corretamente relacionadas aos segmentos da clientela, e o grupo a que pertence cada ponto da amostra é discutido.

Os grupos representados por cada segmento da clientela são propostos com base na descrição estatística do conjunto de dados. O código de transformação e dimensionamento inversos foi corretamente implementado e aplicado para os centros dos grupos.

Comentário

Algo para se ter em mente ao responder essa pergunta é que a fronteira entre cada segmento é praticamente uma linha vertical:

Portanto, pertencer a um segmento ou ao outro parece depender quase exclusivamente da Dimensão 1. E o que a Dimensão 1 representa?

Diversas pontuações são corretamente relatadas, e o número ótimo de grupos é escolhido com base na melhor. A visualização escolhida mostra o número ótimo de grupos baseado no algoritmo de clustering escolhido.

Sugestão

Aqui você também pode usar um loop para reportar as pontuações de acordo com o número de *clusters* utilizado.

Conclusão

Os segmentos da clientela e os dados em `Channel` são comparados. Os segmentos identificados pelos dados de `Channel` são discutidos, inclusive se essa representação é consistente com resultados anteriores.

O estudante discute e justifica como os dados de clustering podem ser usados em um modelo de aprendizagem supervisionada para fazer novas estimativas.

Excelente

Boa resposta, indicando que a segmentação de criadas gerada pelo algoritmo de *clustering* pode ser usada como um novo **atributo** para nos auxiliar em futuros projetos de aprendizado supervisionado.

O estudante corretamente identifica como um teste A/B pode ser feito com a clentela após uma mudança no serviço de distribuição.

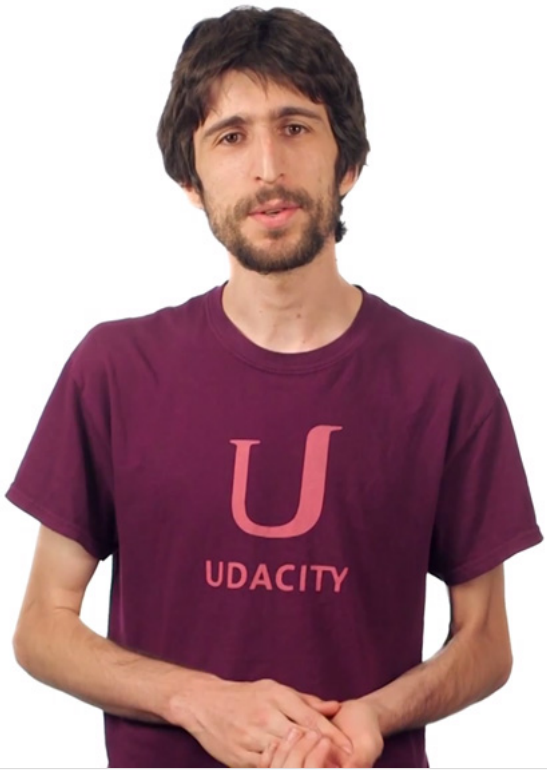
Excelente

Você entendeu perfeitamente a questão mais importante aqui: qualquer teste a ser realizado com esses clientes, incluindo um teste A/B, precisa levar em conta o fato de que há dois segmentos distintos de clientes. Misturar esses segmentos pode levar a resultados indesejáveis no teste.

Por exemplo, suponha que apenas um segmento fosse acolher favoravelmente uma mudança no sistema de entregas, e que a maior parte dos clientes escolhidos para uma mudança no sistema (em um teste A/B) fossem do segmento *contra* a mudança. A conclusão do estudo poderia ser que os clientes *em geral* se opõem a uma mudança de sistema, o que não seria verdade.

REENVIAR

BAIXAR PROJETO



Melhores práticas para sua resubmissão do projeto

Ben compartilha 5 dicas úteis para a revisão ressubmissão do seu projeto.

Assistir Vídeo (3:01)

RETORNAR

Avalie esta revisão

FAQ do Estudante