Machine Learning
Part II

Impacto
Edinson Tolentino
email: et396@kent.ac.uk
Twitter: @edutoleraymondi

Content

Proimpacto

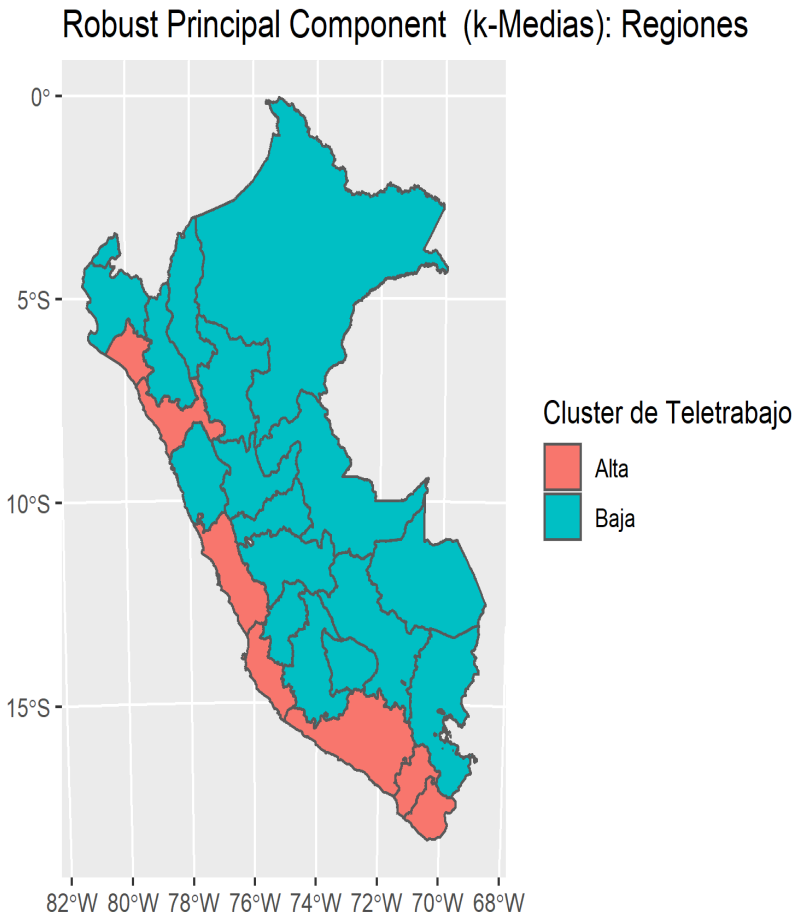
Unsupervised Learning

Proimpacto

- We where observe only the features $X_1, X_2, \cdots, X_p$. We are not interested in prediction, because we do not have an associated response variable Y.
- The goal is to discover interesting things about the measurements:
  - is there an informative way to visualize the data?
  - Can we discover subgroups among the variables or among the observations?
- We discuss two methods:
  1. principal components analysis
  2. clustering

PCA y Cluster

Proimpacto

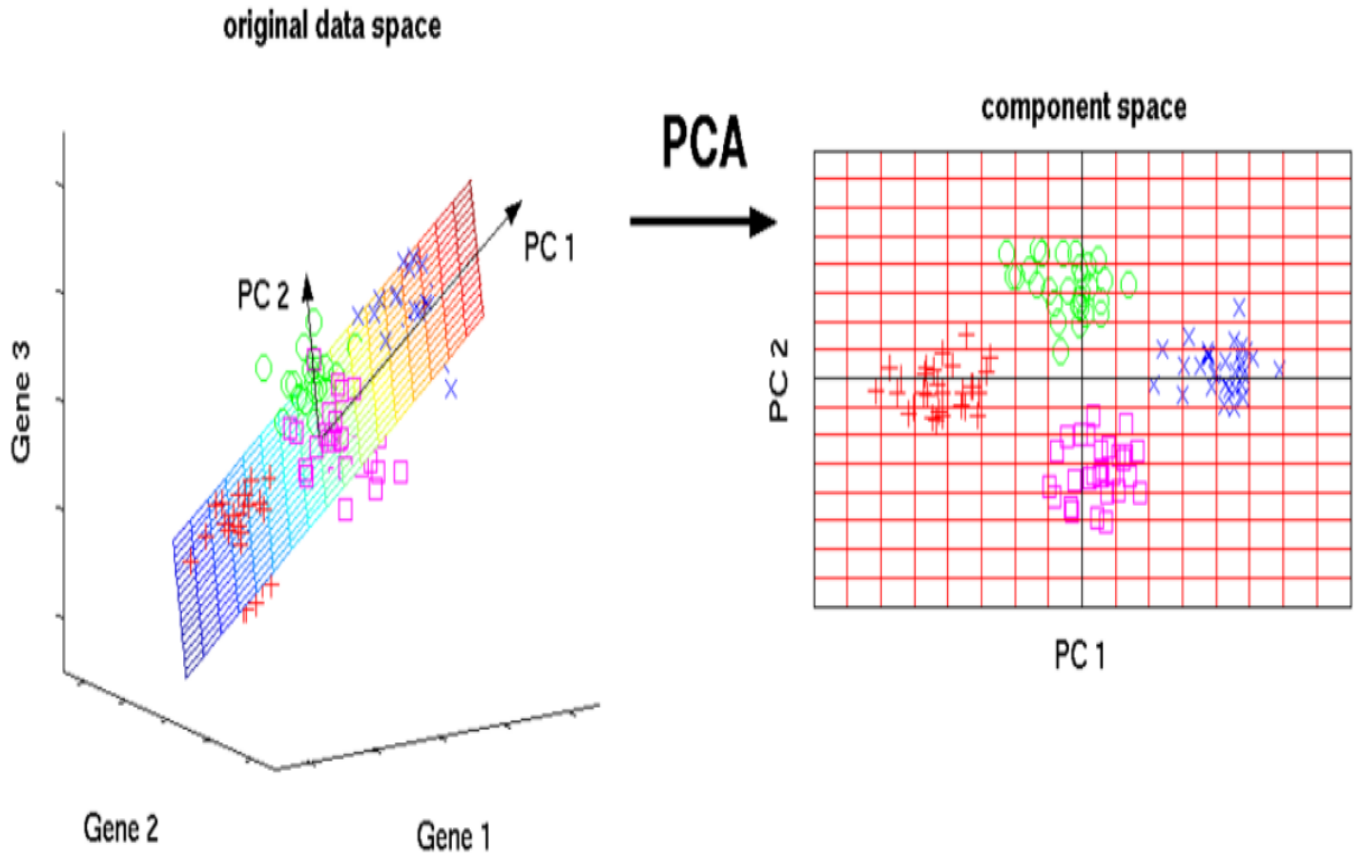


- **Machine Learning**:
  - Análisis supervisado
  - Análisis no supervisado
- **Análisis no supervisado**
  - Tecnicas de reducción de dimensionalidad (`PCA`)
  - Tecnicas de agrupamiento o clustering (`k-means`)

PCA

Proimpacto

- Los métodos de reducción de dimensionalidad consisten en resumir y visualizar la información más importante contenida en un dataset.
- Existen varias técnicas alrededor de esta temática, tanto si asumimos que existen patrones lineales como no lineales en los datos.



PCA : dimension

Proimpacto

- PCA produce una representación de baja dimensión de un conjunto de datos. Encuentra una secuencia de combinaciones lineales de las variables que tienen varianza máxima y no están correlacionadas entre sí.
- El primer componente principal de un conjunto de características X_1, X_2, $\cdots$,X_p es la combinación lineal normalizada de las características

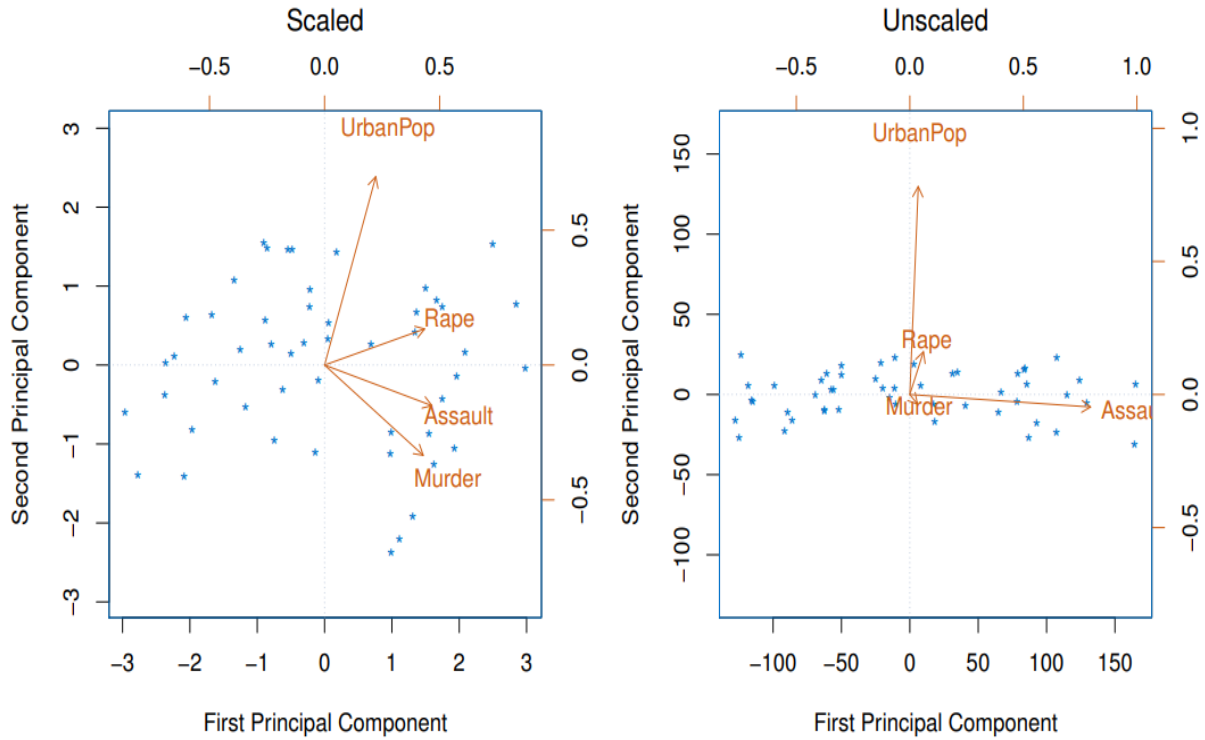
$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \cdots + \phi_{p1}X_p$$

  la cual posee una varianza grande, por tanto al `normalized`, nos hacemos referencia $\sum_{j=1}^{p} \phi_{j1}^2 = 1$
- Restringimos las cargas para que su suma de cuadrados sea igual a uno, ya que de lo contrario, establecer estos elementos en valores absolutos arbitrariamente grandes podría dar como resultado una varianza arbitrariamente grande.

PCA : escala

Proimpacto

- Si las variables son de diferentes unidades de medida, se debe realizar una escala o estandarizacion de las variables (recomendado)
- Si las variables tienen la misma unidad de medida, no es recomendable realizar dicha estandarizacion



Cluster

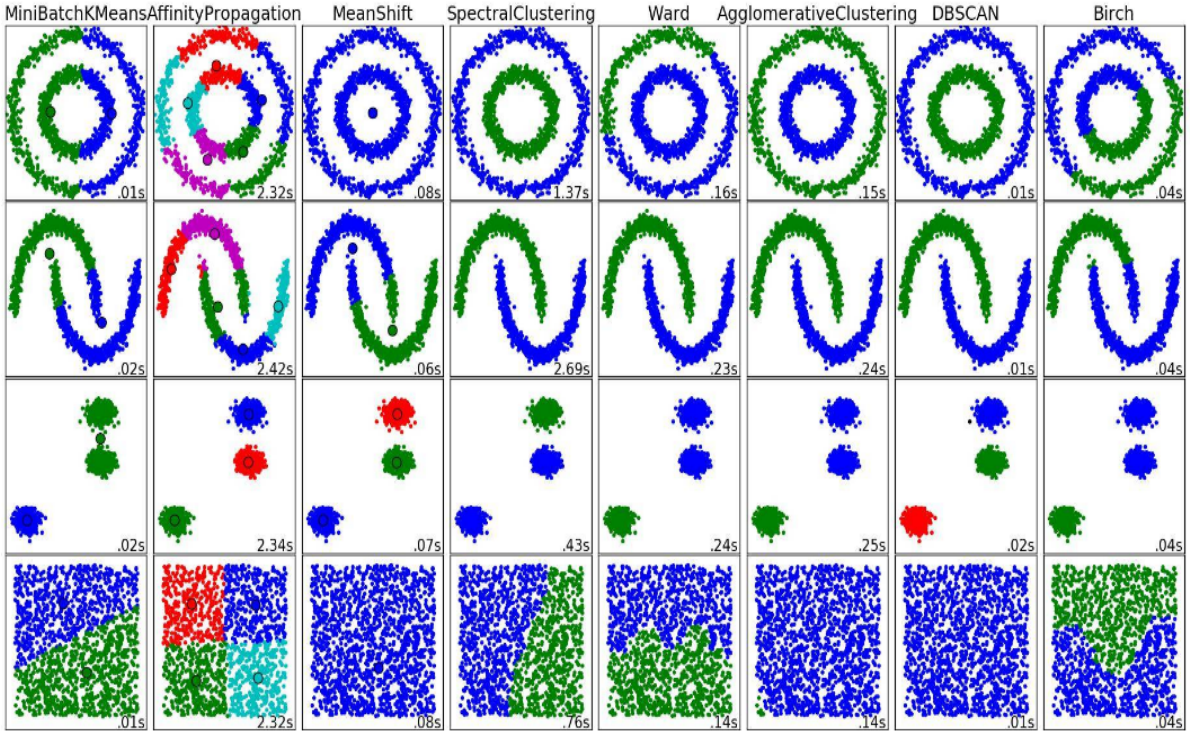
Proimpacto

- El concepto de cluster esta relacionado con tecnicas que permitan encontrar `subgrupos` o `patrones de grupos` en la data usada
- Para concretar esto, debemos definir qué significa que dos o más observaciones sean similares o diferentes.
- De hecho, ésta suele ser una consideración específica de un dominio que debe realizarse basándose en el conocimiento de los datos que se están estudiando.

Cluster

Proimpacto

- Las técnicas de agrupamiento o clustering nos permiten obtener conocimiento a partir del descubrimiento de patrones existentes en los datos.
- Específicamente, el objetivo de los métodos de clustering yacen en la identificación de grupos de objetos similares en un conjunto de datos de interés a través de una medida de similaridad entre puntos (e.g la euclideana).



Cluster vs PCA

Proimpacto

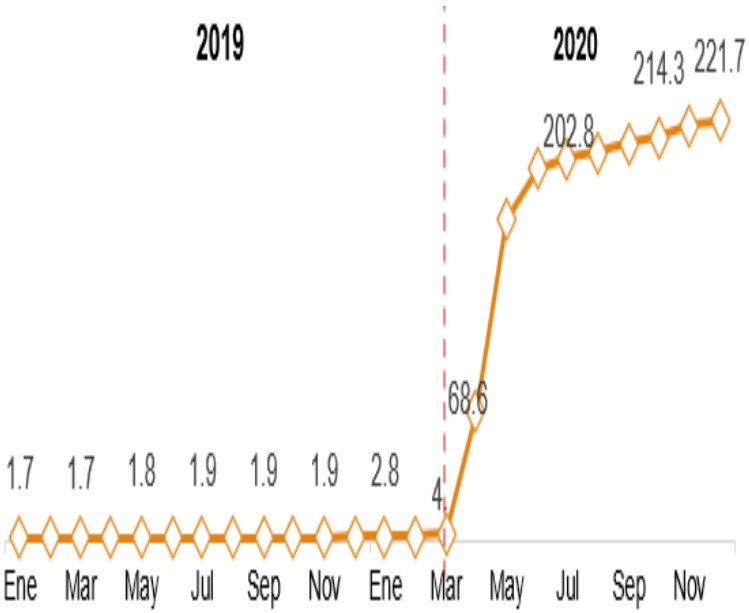
- **PCA**: busca una representación de baja dimensión de las observaciones que explique una buena fracción de la varianza.
- **Cluster**: busca subgrupos homogéneos entre los observaciones.

Aplicacion: Teletrabajo

Proimpacto

Teletrabajo: COVID-19 y los trabajos que se pudieron adaptar

Figura 1: PEA Ocupada bajo modalidad de teletrabajo /remoto



- Durante el 2020, la pandemia del COVID-19 causo un cambio brusco en la tendencia del uso de teletrabajo a nivel mundial
- Por la pandemia también, en el 2020, se perdieron 2.2 millones de empleos (INEI) a nivel nacional y 1.1 millones de empleos fueron perdidos en Lima (IEP).
  - **¿Cuántos de los trabajos se pudieron adaptar a la modalidad de teletrabajo?** ;
  - Solo 18.27% de la PEA ocupada mayor de edad a nivel nacional (667,926 personas). ENAHO

Aplicación: Teletrabajo

Proimpacto

- Para la presente sección se buscará agrupar los departamentos, en función a algunas de sus caracteristicas socieconomicas, informalidad, PBI, entre otras variables
- Para resolverlo compararemos los resultados de usar ACP y ACP Robusto sobre nuestro conjunto de datos para luego agruparlos con k-medias
- Data:
  - Información de la ENAHO para estimar el teletrabajo (información secundaria), PBI, etc
  - Toda la información a nivel de departamento
- Informacion: **Codigo QR** para las bases de datos