Práctica 4. Análisis de Componentes Principales

Nombre: Gregorio Alberto Alvarez Alvarez

Materia: Análisis estadístico multivariable

Introducción

Actualmente con el avance de la tecnología es posible distribuir energía eléctrica a un gran número de industrias e individuos, sin embargo, existe una limitante para almacenar la energía sobrante, por lo que el excedente no es aprovechado. Este remanente no solo significa perdidas monetarias importantes, sino que representa emisiones de carbono perjudiciales al medio ambiente y consecuentemente a múltiples organismos vivos 1. Al crear modelos, que aproximen sus predicciones a los valores esperados de consumo de energía, se espera reducir la problemática antes planteada.

Objetivo

Encontrar las variables que representan la mayor cantidad de la varianza del conjunto de las variables independientes, mediante un análisis posterior a la aplicación de Análisis de Componentes Principales (PCA por sus siglas en ingles), así como encontrar las componentes principales obtenidas con el modelo de PCA.

Objetivos específicos

Analizar la correlación entre las variables en cuestión y la varianza de cada una de ellas.

Utilizar el modelo de PCA para obtener las componentes transformadas e indagar sobre el numero óptimo de variables que se pueden utilizar para utilizar en modelos de predicción.

Obtener el nivel de importancia que tiene cada variable de la base de datos original, con respecto a su repercusión en la creación de variables principales.

Descripción de la base de datos

La base de datos fue obtenida de 2. Los datos están basados en los datos de consumo de energía, de en una ciudad imaginaria. Se cuenta con datos de 23 días de cada mes, desde julio del 2013 a junio del 2017 (26496 entradas). El fin de estos es predecir el consumo de energía para optimizar la producción de energía para esta ciudad.

En total existen 7 variables independientes, que se describen a continuación.

- ID: índice que indica número de muestra.
- datetime: valor de fecha y hora en la que se midieron las variables que se describieran enseguida. Tiene granularidad por hora.
- temperatura: Temperatura ambiental. No existe información acerca de la escala utilizada. Los datos parecen estar medidos en grados Celsius.
- var1: variable cuantitativa, de origen desconocido. Se utiliza esta variable con fines meramente
- pressure: Presión ambiental. No existe información de la escala utilizada.
- windspeed: Velocidad del viento. No existe información acerca de la escala utilizada.
- var2: Variable categórica de origen desconocido. Contiene 3 valores A, B y C.

La variable dependiente contiene información de consumo de energía. Las unidades de energía no se conocen.

Desarrollo y resultados

Se removieron las variables dependientes no categóricas para el adecuado análisis mediante PCA. con lo que se obtuvo una tabla como la Tabla 1.

temperature	var1	pressure	windspeed
-11.4	-17.1	1003	571.91
-12.1	-19.3	996	575.04
-12.9	-20	1000	578.435

Tabla 1. Muestra con las primeras 5 entradas de la tabla de las variables a analizar

Se sabe que, a mayor correlación entre las variables, mayor el nivel de reducción que se puede obtener de esta; por tanto, se obtendrá la tabla de correlación (Tabla 2) para tener una primera idea sobre el alcance del PCA en estos datos.

	temperature	var1	pressure	windspeed
temperature	1	0.811421	-0.723939	-0.162093
var1	0.811421	1	-0.680821	-0.292305
pressure	-0.723939	-0.680821	1	0.171369
windspeed	-0.162093	-0.292305	0.171369	1

Tabla 2. Correlación entre las variables a estudiar.

De la tabla 2, se puede observar que las variables temperatura, var1 y presión, se encuentran más correlacionadas entre sí. Debido a que este grupo representa la mayoría de las variables a analizar, se espera que su varianza represente la mayoría de la varianza del primer componente principal.

Se sabe que por la comparación de variables que hace PCA, es necesario que todas las variables tengan media igual a cero, así como que la varianza entre las variables sea igual o similar. A continuación, se revisará las varianzas y medias de cada variable (Tabla 3)

	mean	var
temperature	5.09899	75.39
var1	-1.91623	108.68
pressure	986.451	144.06
windspeed	23.96	2330.99

Tabla 3. Varianza y Media de variables continuas independientes.

De la tabla 3, se puede observar que ninguna variable tiene media cero y la diferencia entre las varianzas de cada variable es significativa. Por este motivo se estandarizan las variables mediante la clase `StandardScaler`

de la librería `Scikit learn`.

Después de estandarizar las variables, se aplicó PCA mediante la clase de `Sckit learn`, con lo que se obtuvo la siguiente tabla con los pesos de las componentes de cada componente principal.

# Componente	W_temperature	W_var1	W_pressure	W_windspeed
Componente 1	0.571961	0.574549	0.540063	0.22602
Componente 2	0.195795	0.0161708	0.1789	0.964053
Componente 3	0.31278	0.484254	0.813255	0.0792691
Componente 4	0.732594	0.659644	0.122241	0.115038

Tabla 4. Peso de cada variable para cada componente principal. Número de componente principal (Fila). Peso de variable en componente principal (Columna)

De la tabla 4, se puede observar que la temperatura, var1 y la presión tienen mayor importancia para la creación del primer componente principal. Debido a que este grupo contiene 3/4 del total de variables, se puede concluir que la combinación de estas variables representa una parte importante de la varianza del total del grupo de variables cuantitativas. La segunda componente principal está basada en su mayoría en los valores de la variable de velocidad del viento.

Con esto se puede observar que los resultados de las componentes principales son congruentes con los resultados obtenidos de la tabla de correlación, en donde se observó que existen 2 grupos de variables según el nivel de correlación entre ellas; sin embargo, el remanente de las varianzas (Componente 3 y 4) no se puede explicar tan fácilmente con los resultados de la correlación.

Del porcentaje acumulado de la varianza (Tabla 5), obtenido de las variables de PCA. Se encontró que son necesarias 2 variables principales para representan 87.4 % de la varianza total, siendo el remanente de 12.6 %.

# CPs	varianza explicada
1 CP	64.1 %
2 CP	87.4 %
3 CP	95.7 %
4 CP	100.0 %

Tabla 5. Acumulado de porcentaje de varianza explicada.

Conclusión

Se encontró que existen 2 grupos principales de variables debido a la correlación entre ellas. Después de aplicar PCA se encontró que 2 variables principales representan el 87.4 % de la varianza explicada total y estas variables se encuentran relacionadas con los 2 grupos antes encontrados.

En un trabajo futuro se planea calcular la correlación de las variables que componen al primer componente principal, con el fin de seleccionar la más importante de este grupo, y así, reducir el número de variables independientes, sin el problema de perdida de explicabilidad de las variables, que surge al utilizar componentes principales.

Así mismo, mediante el análisis de componentes principales y pruebas ad-hoc se planea crear una matriz de diseño más robusta utilizando las variables categóricas de la base de datos, con el fin de probar la eficiencia de cada subconjunto para la predicción de consumo de energía eléctrica.

Referencias

literatura

- [1] https://www.researchgate.net/profile/Alexander-Verl/publication/229019188_Energy_Consumption_Forecasting_and_Optimisation_for_Tool_Machines/links/00b4952a58d0031191000000/Energy-Consumption-Forecasting-and-Optimisation-for-Tool-Machines.pdf
- [2] https://www.kaggle.com/datasets/utathya/electricity-consumption?select=train.csv

Código

- [git] https://github.com/Gegori1/analisis-estadistico-multivariado/tree/master/practica_4