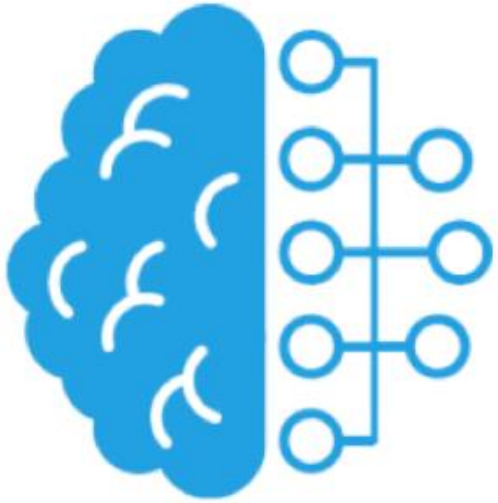
Aprendizaje de Máquina
Clase 7

Claudio Delrieux – DIEC - UNS
cad@uns.edu.ar



Reducción de Dimensionalidad

Trabajar con datasets muy anchos (feature spaces de mucha dimensiones) genera una gran cantidad de problemas, más allá del costo computacional. Esto se conoce como la "**maldición de la dimensionalidad**".

Al aumentar la dimensionalidad del espacio, la densidad de los datos baja exponencialmente, por lo que la estabilidad y robustez de las técnicas de análisis se compromete, y la significatividad estadística de los resultados se debilita. Se requiere recolectar una enorme cantidad de datos de entrenamiento para garantizar un cubrimiento razonable de los posibles casos.

Reducción de Dimensionalidad

Cuando la cantidad de dimensiones es muy alta las métricas de distancia pierden las propiedades intuitivas, y los métodos basados en distancias (k-NN por ejemplo) se sesgan muy rápidamente.

Además, durante las tareas de curado del dataset y análisis exploratorio, muchas veces es necesario visualizar los datos de alguna manera razonablemente accionable.

Finalmente, datos con muchos atributos tienen mucha probabilidad de tener atributos irrelevantes, valores faltantes o contaminados, etc., por lo que se generan modelos con mucha variancia y el sobreajuste es difícil de controlar.

Reducción de Dimensionalidad

La reducción de la dimensionalidad consiste en transformar el dataset original en otro de menor cantidad de atributos (tabla “más angosta”), pero que retenga las propiedades significativas del original con respecto al propósito de análisis.

Muy crudamente, estos métodos pueden catalogarse en *lineales* y *no lineales*, y varias de las técnicas que ya conocemos pueden utilizarse directa o indirectamente para reducir dimensionalidad.

Selección de Atributos

Consiste en retener (siguiendo algún mecanismo y criterio) solamente algunos de los atributos originales. Es un caso particular de *proyección*, donde ésta se hace hacia un hiperplano principal del dataset original.

La expectativa es eliminar del análisis los atributos *irrelevantes* (que más que aportar a la calidad, la bajan) y los atributos *redundantes* (que tienen correlación significativa con otros atributos).

La técnica más fácilmente comprensible es buscar exhaustivamente todas las combinaciones de subconjuntos de atributos y testearlas respecto de alguna métrica de performance, lo cual es intratable.

Selección de Atributos

Los métodos de selección de atributos se pueden agrupar en *filtros*, *wrappers*, y *métodos embebidos*. El tipo de modelo (clasificador, regresor, etc.) y de evaluación de calidad (precisión, error, etc.) determinan fuertemente el método de selección.

Los métodos de *filtrado* son adecuados para clasificadores o regresores, analizan los atributos uno a uno con respecto al atributo target utilizando regresión, información mútua o error. La suposición es que los atributos que modelan individualmente bien, también lo hacen en combinación con otros atributos.

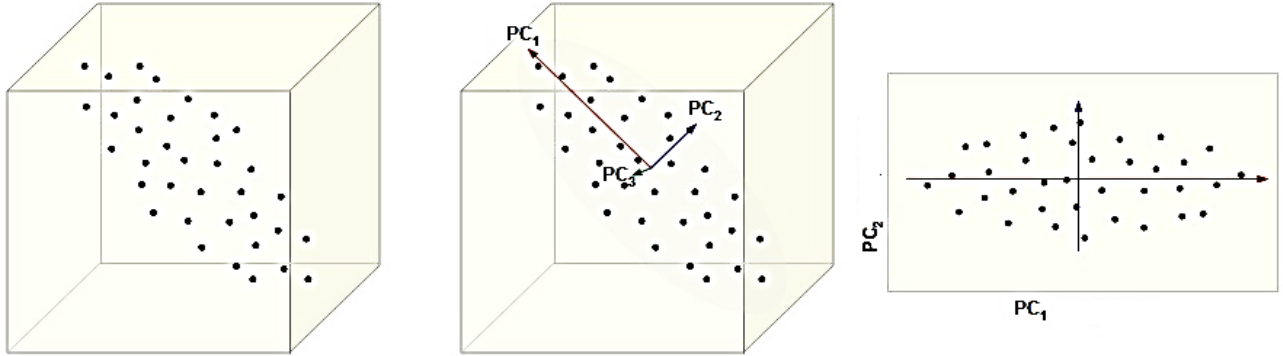
Selección de Atributos

Los métodos de seleción por *wrapper* evalúan modelos con subconjuntos de atributos siguiendo una política determinística o aleatoria, que va reteniendo aquellos más exitosos (o eliminando los que no aportan a la calidad). Se pueden utilizar árboles de decisión para elegir el atributo a agregar (o a eliminar).

Los métodos *embebidos* en realidad pueden pensarse como regularizadores (el modelo tiene parámetros de simplificación), como es el caso de LASSO y ridge que vimos en regresores.

Proyección de Atributos

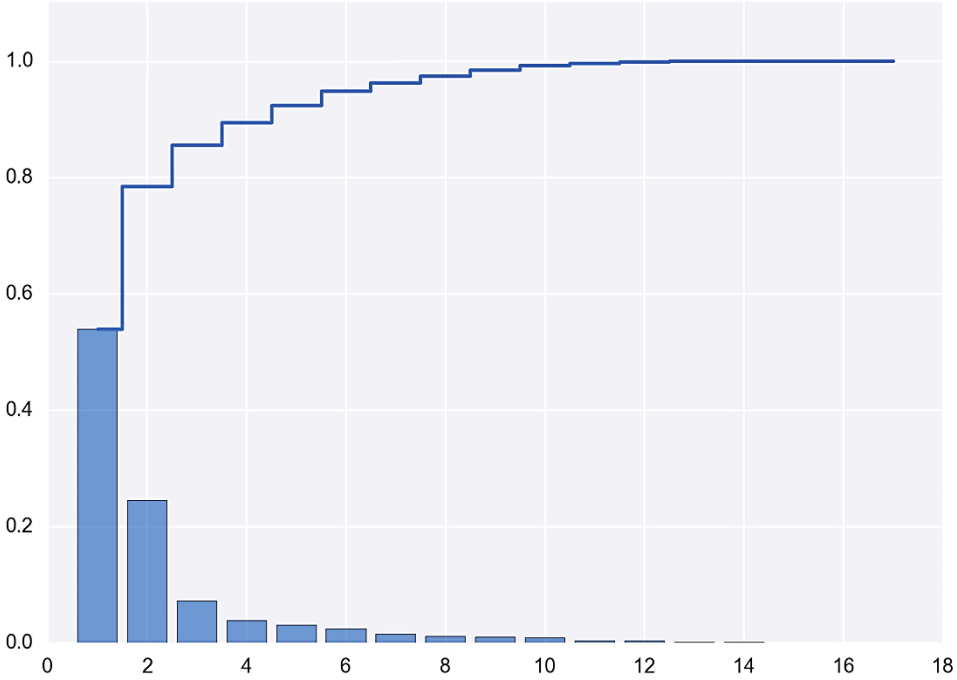
Estos métodos también proyectan el espacio de atributos sobre un subespacio lineal (i.e., un hiperplano), pero en este caso puede estar alabeado respecto de los ejes principales. El más conocido es PCA (implícitamente es lo que utilizamos en regresión lineal). Puede hacerse sobre el espacio de atributos original o sobre un espacio kernel adecuado.



Proyección de Atributos

La matriz de covariancia es semidefinida positiva (todos los AV positvos) y cada AV normalizado respecto de la suma total indica la variancia que se proyecta sobre el autovector asociado.

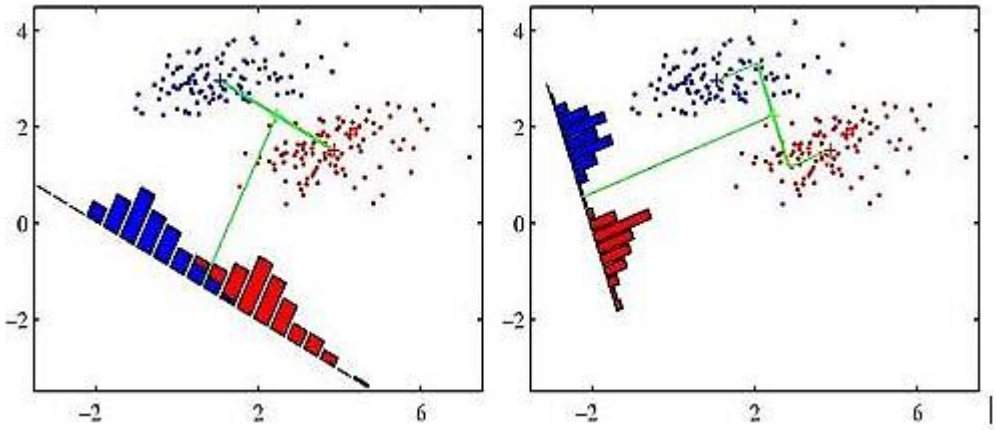
Definimos un espacio de atributos de acuerdo a la cantidad de AVs a retener y a la variancia a explicar.



Proyección de Atributos

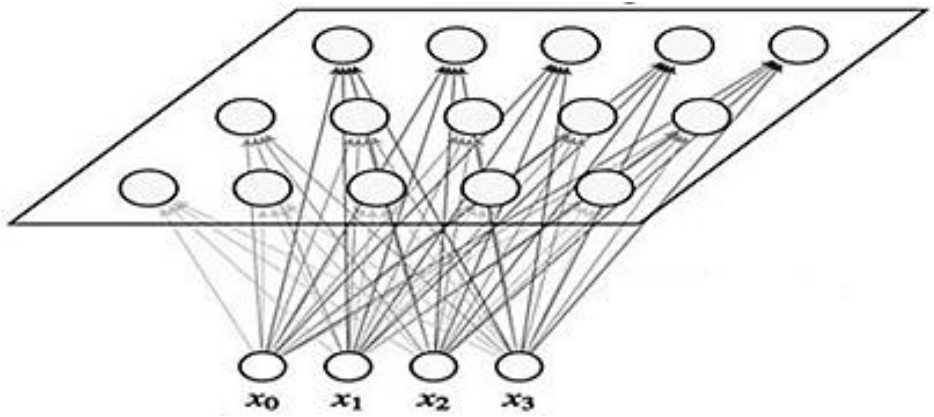
El análisis por discriminantes lineales (LDA) está relacionado con el PCA, pero más orientado a clasificación (atributo target nominal), y para encontrar la frontera lineal en el espacio de atributos independientes.

Este espacio se proyecta hacia el hiperplano separador (discriminante) entre las clases. Es otra forma de entender lo ya visto en las filminas 21 a 23 de la clase 2.



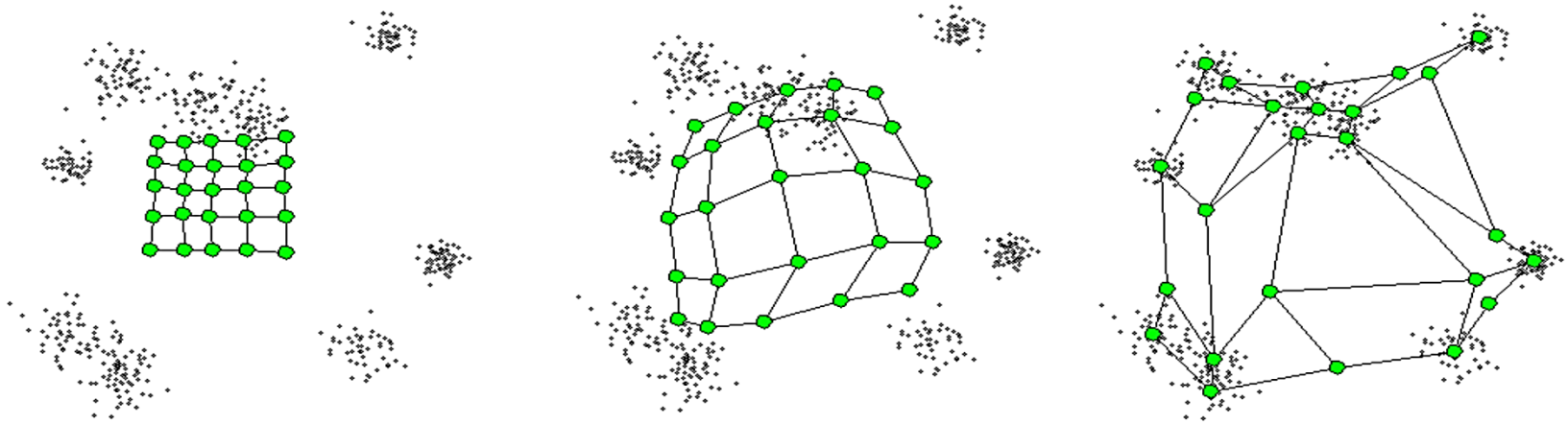
Métodos no Lineales: SOM

Los mapas auto-organizados son un conjunto de celdas (generalmente 2D) tales que, una vez entrenadas, son mapeadas desde un vector de entrada en el espacio de atributos. Luego, todos los datos son asignados a las celdas más cercanas.



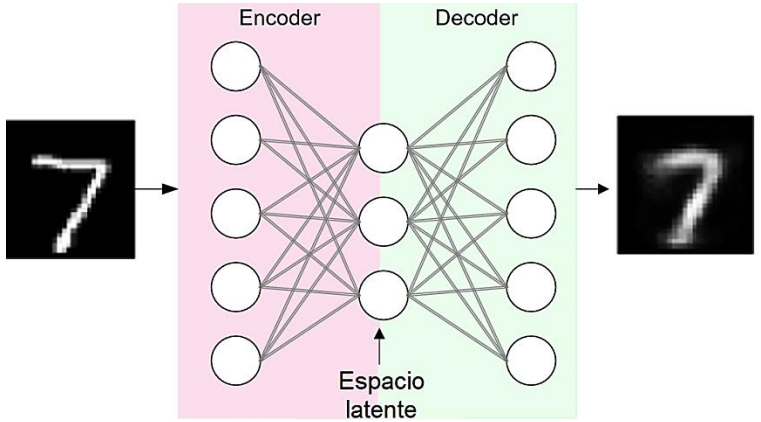
Métodos no Lineales: SOM

El entrenamiento se basa en iterar aleatoriamente los datos, para cada uno encontrar la celda más cercana, y reentrenar sus pesos y los de sus vecinos de acuerdo a determinada política para ajustarse mejor a ese dato.



Métodos no Lineales: Autoencoders

Si bien no es una “tecnica específica” de ML, está siendo utilizada cada vez con mayor frecuencia de la mano del DL. La idea es entrenar una red para copiar datos de entrenamiento en su salida, que sean idénticos a la entrada (o que retengan características específicas), pero pasando por un “cuello de botella” de baja dimensión, para aprender así una representación latente.



Métodos no Lineales: t-SNE

t-Stochastic Neighbor Embedding (a.k.a., *tizni*) es un método diseñado específicamente para visualización. El objetivo es mapear el dataset de muy alta dimensionalidad a 2D en forma útil y fiel. El algoritmo aplica transformaciones no lineales adaptativas a los datos y a diferentes regiones.

Primero, t-SNE construye una distribución probabilística entre los pares de puntos del dataset, de manera que datos similares obtienen mayor probabilidad de elegirse como vecinos que datos dissimilares.

Métodos no Lineales: t-SNE

Luego se asocia esa distribución a la probabilidad inducida en el mapa 2D, y se busca minimizar la divergencia de Kullback–Leibler entre esas dos distribuciones, utilizando un método iterativo.

Una característica notable de t-SNE es su hiperparámetro de “perplejidad”, que permite equilibrar el mapeo de acuerdo a la densidad local de los puntos y su variancia.

Ver el tutorial en https://distill.pub/2016/misread-tsne/.

Métodos no Lineales: UMAP

También es un método desarrollado para visualización, y también se basa en encontrar un “análogo” 2D del dataset. En este caso el fundamento proviene de la topología. Primero se establece una variedad diferenciable (*manifold*) entre los datos originales, la cual se modela con una estructura topológica fuzzy. A partir de esa estructura se busca una proyección 2D que retenga la mayor parte posible de las propiedades equivalentes a la original.

Ver detalles y ejemplos en https://umap-learn.readthedocs.io/en/latest/

Recursos Adicionales

Siempre ver ejemplos en SKLearn:

https://scikit-learn.org/stable/auto_examples/feature_selection/plot_select_from_model_diabetes.html#sphx-glr-auto-examples-feature-selection-plot-select-from-model-diabetes-py

Recomendamos también el video de MLxtend Wrapper Method

https://www.youtube.com/watch?v=zW1SvA0Z-I4

Actividad Práctica

Como en los TPs anteriores, elegir un problema de complejidad razonable y aplicar alguno/s de los métodos aquí presentados.

Pueden tomar los ejercicios anteriores si ya están familiarizados con los datasets, sobre todos los que tienen feature spaces muy grandes.

Presentamos también un par de problemas nuevos.

Actividad Práctica

El objetivo de este proyecto es describir las disparidades entre géneros en la última Encuesta Nacional Argentina a trabajadores sobre Condiciones de Empleo, Trabajo, Salud y Seguridad (ECETSS 2018). El dataset contiene 8966 encuestas y 373 columns (preguntas). Por favor, utilizar y contrastar al menos dos métodos de selección de características durante el desarrollo del proyecto.

https://github.com/manlio99/Materia-de-aprendizaje/tree/master/3_MidtermProjects/trabajadores

Actividad Práctica

Desarrollar el siguiente challenge propuesto IFOOD, una empresa líder en el servicio de delivery. Por favor, utilizar y contrastar al menos dos métodos de selección de características.

https://github.com/manlio99/Materia-de-aprendizaje/tree/master/3_MidtermProjects/trabajadores