Conversaciones en línea sobre Ciencia de Datos, IBM-Bourbaki

Escuela de Matemáticas Bourbaki

Marzo 2020

Contents

1 Introducción

Este curso en línea está organizado en forma de charlas. El objetivo es introducir y utilizar el lenguaje de las matemáticas para estudiar algunas de las ideas y dificultades recurrentes en la Ciencia de Datos.

La intención del curso es ofrecer una perspectiva general sobre diversos aspectos en la Ciencias de Datos. El curso esta dividido en dos partes fundamentales, cada una consta de tres clases:

- Eficiencia computacional y estadística
- Teoría de la información e inteligencia artificial

1.1 Temario

Los temas a tratar son los siguientes:

- ¿La bendición o la maldición de la dimensión en la clasificación lineal?
- Detección de anomalías mediante el análisis topológico de datos.
- Consideraciones matemáticas de la computación cuántica.
- Codificación de la información y Entropía
- Inteligencia artificial v.s. los grandes maestros de los juegos de estrategia
- Programación dinámica, una invitación al Aprendizaje por refuerzo

1.2 IBM

A partir de la segunda clase, IBM es co-organizador de este curso y participará mediante la intervención de diversos profesionales de su equipo, ellos son expertos en temas afines a cada uno de los temas propuestos por la Escuela Bourbaki. El objetivo es aportarle a los estudiantes una visión más completa de los problemas en Ciencia de Datos, por ejemplo que incluya aspectos de negocio y de la implementación. Es un placer trabajar con una compaía como IBM con tantísimos recursos intelectuales a su alcance, este gusto se hace extensivo a compartir la preparación del curso con los ponentes. Es justo agradecer en particular a Isaac Carrada quién hizo posible esta sinergia y a Rubén Pineda cuyo apoyo ha sido esencial.

Los ponentes de IBM y los temas de sus charlas son:

1. Adolfo Ollín Cruz: Actualmente Client Technical Prosesional y Data Science Entusiastic en IBM México, Adolfo Cruz se dedica a crear soluciones de industria explotando información tanto interna como externa con el fin de mejorar sus procesos y la toma de decisiones usando la más alta tecnología del mercado.
2. Vanessa Hernández
3.
4. Isaac Carrada
5.

1.3 Estructura del curso

Cada una de las clases está dividida en dos partes:

- Los primeros 45 minutos enseñaremos los detalles matemáticos de algún modesto aspecto relacionado con el tema a tratar, buscamos que los asistentes se expongan a una explicación minuciosa y esta les sea útil para entender mejor la segunda parte de la clase. Compartiremos con los asistentes unas notas cortas sobre esta sección.
- Los segundos 45 minutos se dedicarán a tratar aspectos más generales y elaborados del tema, incluso algunos más polémicas.
- Los últimos 45 minutos un experto en IBM nos hablará sobre problemas afines y su relevancia en IBM.

2 La maldición v.s. la bendición de la dimensión

Uno de los aspectos más importantes en Machine Learning es la batalla entre la cantidad de ejemplos de una base de datos y la dimensión donde vive alguna vectorización de estos ejemplos.

Las llamadas bendición y maldición de la dimensión se manifiestan cuando la cantidad de ejemplos es menor a la cantidad de dimensiones del problema vectorizado.

2.1 Primera parte: árboles de decisión y separación lineal

2.1.1 Problemas supervisados de clasificación binaria

Supongamos que estamos en un problema clásico de aprendizaje supervisado tipo clasificación binaria en $\mathbb{R}^d$ donde buscamos entrenar a nuestro modelo con un conjunto de ejemplos $S = \{(x_1, y_1), \ldots, (x_N, y_N)\}$ i.e. $x_i \in \mathbb{R}^d, y_i \in \{-1, +1\}$.

El objetivo de un algoritmo que se entrene utilizando a S es encontrar una función $f^* : \mathbb{R}^d \to \{-1, 1\}$ que cumpla lo siguiente:

1. (Entrenamiento) Para la mayor parte de los puntos $(x_i, y_i) \in S$ la función f^* satisfaga $f^*(x_i) = y_i$. Es importante notar que algunas veces no es posible que la igualdad anterior sea para todos los puntos en S sin embargo por lo menos nos gustaría que fuera cierto para buena parte de ellos.

2. (Predicción) Un posible supuesto en un problema de Machine Learning es que exista alguna regla que de manera general pueda aplicarse a una observación x y nos regrese un resultado y. Este supuesto se traduce en que exista una función desconocida $f : \mathbb{R}^d \to \mathbb{R}$ (que nuevamente para la gran mayoría de los $(x_i, y_i) \in S$ satisfaga $f(x_i) = y_i$). El segundo objetivo (y quizás el más importante en Machine Learning) es que f y f^* sean parecidas.

2.1.2 Una maldición de la dimensión en un problema de clasificación

Para simplificar el estudio y poder hablar de árboles de decisión de la manera más sencilla posible vamos a simplificar aún más un problema de clasificación y por el momento supondremos que $x_i \in \{0, 1\}^d$. Esto significa que las características de nuestra base de datos son binarias, pensemos por ejemplo en una imagen que solo tiene pixeles negros o blancos. Eso significa que utilizando nuestro conjunto S buscamos aproximar alguna función $f : \{0, 1\}^d \to \{-1, +1\}$.

Definition 2.1. Un árbol de decisión de profundidad r y dimensión d es un conjunto de r-preguntas ordenadas sobre un vector $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ de la siguiente forma:

- x_{i_1} es igual a cero o a uno?
- x_{i_2} es igual a cero o a uno?
- x_{i_3} es igual a cero o a uno?

...

- x_{i_r} es igual a cero o a uno?

De tal forma que cada hoja formada por posibles respuestas a todas estas preguntas nos arrojan una posible clasificación ya sea -1 o $+1$.

Exercise 2.1. *Sea $S = \{(0, 0, 1), (0, 1, -1), (1, 0, -1), (1, 1, 1)\}$. Encontrar el árbol de clasificación que prediga la clasificación en S. Es posible encontrar un árbol de clasificación de profundidad uno que también prediga la clasificación en S?*

La ventaja y al mismo tiempo gran problema de los árboles de decisión es la siguiente:

Proposition 2.2. *Toda función $f : \{0,1\}^d \to \{-1,+1\}$ corresponde con un árbol de decisión con profundidad d.*

Lo anterior genera un gran problema debido a la posibilidad de incurrir en sobre-ajuste, más adelante en la clase número tres de este curso hablaremos sobre la navaja de Occman que es un resultado que previene el sobre-ajuste, el problema es que estamos considerando árboles de decisión con profundidad demasiado grande. Para explicar con formalidad las desventajas causadas por ese sobre-ajuste podemos enunciar el siguiente resultado.

Theorem 2.3. *La cantidad de ejemplos necesaria para entrenar un árbol de decisión de manera eficiente crece de manera proporcional a 2^d.*

El resultado anterior se puede entender como una maldición de la dimensión pues a medida que d crezca, la cantidad de ejemplos necesarios 2^d podría ser excesivamente grande.

Por el momento no hemos hablado del algoritmo que entrena de manera eficiente un árbol de decisión, solo mencionamos que está relacionado con la idea de simpleza proveniente de Occman. En el curso 4 hablaremos de la teoría de información y su importancia para entrenar un árbol de decisión.

2.1.3 Una bendición de la dimensión en un problema de clasificación

Una de las formas más simples para afrontar un problema de clasificación es mediante la llamada clasificación lineal. Para esto podemos regresar sin problema al caso cuando $x_i \in \mathbb{R}^d$.

Definition 2.2. El producto punto entre dos vectores $x = (x_1, \ldots, x_d), x' = (x'_1, \ldots, x'_d) \in \mathbb{R}^d$ se define como $x \cdot x' = x_1 x_1 + \ldots + x_d x'_d$

Exercise 2.4. *Verificar que el producto punto de dos vectores coincide con multiplicar dos matrices, cuáles?*

Definition 2.3. Dos vectores $x, x' \in \mathbb{R}^d$ son linealmente separables mediante el hiperplano definido por $\beta = (\beta_1, \ldots, \beta_d) \in \mathbb{R}^d$ y $\beta_0 \in \mathbb{R}$ si $x \cdot \beta + \beta_0 > 0$ y $x \cdot \beta + \beta_0 < 0$. Decimos que dos conjuntos S, S' de vectores en $\mathbb{R}^d$ son linealmente separables mediante β, β_0 si todo par de puntos en S y S' son linealmente separables.

Exercise 2.5. *Partiendo de la ecuación clásica de una recta $y = mx + b$ escribir qué significa que dos puntos en el plano $\mathbb{R}^2$ sean separables linealmente, preferentemente utilizando la noción de producto punto.*

En esta sección propondremos una condición algebráica sobre el conjunto S para garantizar que es separable linealmente por un hiperplano sin importar cómo sus elementos estén clasificados. Esto es considerablemente más fuerte a lo que buscábamos en un inicio.

La condición algebraica necesaria se define a continuación:

Definition 2.4.

- Una familia de vectores $x_1, x_2, \ldots x_m \in \mathbb{R}^d$ son linealmente dependientes si existen $\gamma_1, \gamma_2, \ldots, \gamma_m \in \mathbb{R}$ distintos de cero tales que $\gamma_1 x_1 + \ldots \gamma_m x_m = 0$
- Sea $S \subseteq \mathbb{R}^d$ un conjunto arbitrario de puntos de tamaño N (notemos que esta vez no estamos suponiendo que los puntos están etiquetados). Decimos que S está en posición general cuando no existe un subconjunto $S' \subset S$ de tamaño menor a d que sea linealmente dependiente.

Exercise 2.6. *Cuál es la mayor cantidad de puntos en $\mathbb{R}^2$ que pueden estar en posición general? Y en $\mathbb{R}^3$ o $\mathbb{R}$?*

Theorem 2.7. *(Cover) sea $S \subseteq \mathbb{R}^d$ un conjunto arbitrario de tamaño N con $d > N$. Además supondremos que S está en posición general en el sub-espacio generado por ellos. Entonces cualquier etiquetado de S: $\left(x_{i_j}, -1\right)$ si $j \leq m \leq N$ y $\left(x_{i_j}, 1\right)$ si $j > m$ es linealmente separable.*

El trabajo de Cover es mucho más profundo que el resultado anterior y vale la pena consultar sus detalles en el artículo originalGeometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition.

2.2 Segunda parte: cocentración de la medida en altas dimensiones

En la segunda parte nos concentraremos únicamente en las llamadas bendiciones de la dimensión. La base técnica para estos resultados son los llamados Teoremas de Concentración en Probabilidad, comenzaremos esta sección describiendo algunos de estos resultados desde un aspecto púramente teórico.

2.2.1 Teoremas de Concentración

La mayoría de estos resultados son suficientemente antiguos sin embargo nos concentraremos en dos resultados particularmente, uno de Chebyshev y el otro de Lévy.

Lemma 2.8. *(Desigualdad de Chebyshev) Sea X una variable aleatoria con valores en $\mathbb{R}$ y $\mathbb{E}[X] = \mu, Var(X) = \sigma^2$, entonces para cualquier $t > 0$ tenemos:*

$$\mathbb{P}(|X - \mu| > t) < \frac{\sigma^2}{t^2}$$

Esta desigualdad es considerablemente más poderosa cuando la dimensión donde toman valores las variables aleatorias crece considerablemente, aún así en la siguiente sección daremos un ejemplo que nos da una idea de cómo este tipo de desigualdades son ejemplos de la llamada bendición de la dimensión.

Antes de continuar vamos a introducir una clase de funciones súmamente importantes, también en el contexto de Ciencias de Datos.

Definition 2.5. Sea $f : \mathbb{R}^n \to \mathbb{R}^m$ y $\nu > 0$, decimos que f es una función ν-Lipschitz cuando para $x, y \in \mathbb{R}^n$ se tiene:

$$||f(x) - f(y)||_2 \leq \nu ||x - y||_2$$

La intuición de estas funciones para un Científico de Datos es la siguiente:

Utilizando el acercamiento Bayesiano supongamos que tenemos una distribución de probabilidad D sobre $\mathbb{R}^d \times \{-1, +1\}$.

Sean X, Y variables aleatorias en $\mathbb{R}^d$ y $\{-1, +1\}$ cuyas distribuciones coinciden con las distribuciones marginales de D.

Definamos a la función f de la siguiente manera:

$$f(x) = \mathbb{P}_D(Y = 1|X)$$

El acercamiento Bayesiano consiste en aproximar de la mejor manera la función f utilizando una base de datos S.

Exercise 2.9. *¿Qué repercución tiene en Ciencia de Datos que la función f sea ν-Lipschitz?*

En el estudio de la maldición de la dimensión en el contexto de K-NN las funciones Lipschitz son muy importantes.

Exercise 2.10. *Fijemos un vector $\beta \in \mathbb{R}^d$, definimos $f : \mathbb{R}^d \to \mathbb{R}$ de la siguiente forma:*

$$f(x) = \beta \cdot x$$

Demostrar que f es 1-Lipschitz.

Definition 2.6. La esfera de dimensión $d - 1$ es el conjunto de puntos

$$\mathbb{S}_{d-1} = \{x \in \mathbb{R}^d : ||x||_2 = 1\}$$

Proposition 2.11. *(Lévy) Sea X una variable aleatoria con valores en la esfera $\mathbb{S}_{d-1}$, cuya distribución es simétrica respecto a rotaciones y $f : \mathbb{R}^d \to \mathbb{R}$ una función ν-Lipschitz. Entonces para cualquier $t > 0$*

$$\mathbb{P}\left(|f\left(X\right) - \mathbb{E}\left[f\left(X\right)\right]| > t\right) < 2 \cdot exp\left(-\frac{d \cdot t^2}{2c\nu^2}\right)$$

Es importante notar que esta desigualdad será más y más precisa a medida que d sea suficientemente grande dado el decrecimiento de la función exponencial para valores negativos.

2.2.2 Algunas aplicaciones de los teoremas de concentración

- Comencemos con una sencilla pero agradable aplicación de la desigualdad de Chebyshev al estudio probabilista del lanzamiento de una moneda:

  Supongamos que lanzamos una moneda con águila y sol d-veces. ¿Cuál es la probabilidad de obtener águila $\frac{3}{4}d$-veces?

  Si le llamamos A_d al número de veces que obtenemos águila después de lanzar la moneda d-veces entonces entonces es fácil convencerse de que:

  Exercise 2.12. *Definir el espacio de probabilidad adecuado y calcular $\mathbb{E}\left[A_d\right] = \frac{d}{2}$, $Var\left(A_d\right) = \frac{d}{4}$.*

  Al sustituir en la desigualdad de Chebyshev $t = \frac{d}{4}$ obtenemos:

  $$\mathbb{P}\left(|A_d - \frac{d}{2}| > \frac{d}{4}\right) < \frac{4}{d}$$

- Una segunda aplicación del fenómeno de concentración de medidas en altas dimensiones es el llamado lema de Johnson-Lindenstrauss el cuál puede pensarse como un lema para reducir la dimensión.

  Antes de eso recordemos un resultado importante sobre la reducción de la dimensión en Ciencia de Datos.

  Supongamos que tenemos un conjunto $S = \{x_1, \ldots, x_N\} \subseteq \mathbb{R}^d$. La reducción de la dimensión consiste en encontrar puntos $S' = \{x'_1, \ldots, x'_N\} \subseteq \mathbb{R}^p$ con p "considerablemenente" más pequeño que d de tal forma que las dos siguientes condiciones se cumplan:

  1. La geometría del conjunto S' se parezca lo más posible a la geometría del conjunto S.
  2. Exista una transformación lineal M que envíe S a S'.

  El primer punto es clave pues no hemos sido suficientemente claros sobre qué significa la similitud geométrica.

  En el contexto de PCA lo que significa esa similitud geométrica es lo siguiente:

  No solo buscamos una transformación lineal M sino además queremos una transformación M' que "regrese" a S, eso quiere decir que minimice la distancia promedio entre $S'' = M'\left(M\left(S\right)\right)$ y S. No es difícil ver que este acercamiento es equivalente a buscar maximizar la varianza del conjunto S'. Normalmente el resultado cuando p es demasiado pequeño respecto a d es poco satisfactorio pues el conjunto S' será poco representativo del conjunto inicial S.

  Una manera alternativa preservar la geometría del conjunto S que por cierto podría ser relevante cuando se necesite utilizar un algoritmo de proximidad es el llamado lema de Johnson-Lindenstrauss el cuál dice lo siguiente:

Lemma 2.13. *(Johnson-Lindenstrauss) Sea* $S = \{x_1, \ldots, x_N\} \subseteq \mathbb{R}^d$ *y* $\epsilon > 0$. *Si* $p \geq \frac{C}{\epsilon} log\,(N)$ *entonces con alta probabilidad existe un subespacio vectorial (aleatorio)* $\mathbb{R}^p \subseteq \mathbb{R}^d$ *y una transformación lineal(también aleatoria)* $M : \mathbb{R}^d \to \mathbb{R}^p$ *tal que para todo par de puntos* $x_i, x_j \in S$*:*

$$(1-\epsilon)\,||x_i - x_j||_2 \leq ||M\,(x_i) - M\,(x_j)\,||_2 \leq (1+\epsilon)\,||x_i - x_j||_2$$

La precisión en probabilista de este resultado depende de que d sea suficientemente grande utilizando el resultado de Lévy.

- Ahora recordemos el resultado de Cover que era nuestro primer ejemplo de la bendición de la dimensión, si $d > N$ entonces cualquier partición de cualquier conjunto en posición general de tamaño N es linealmente separable por un hiperplano que pasa por el orgien. Una pregunta inmediata es ¿qué pasa cuando la desigualdad anterior no se cumple? Es decir supongamos que $N > d$. Este es nuestro ejemplo final de una bendición de la dimensión: los teoremas estocásticos de separación lineal nos dicen que si d es suficientemente grande entonces con alta probabilidad (dependiendo de lo grande que sea d) un conjunto S generado por variables aleatorias identicamente distribuidas e independientes es linealmente separable. En este caso la noción probabilista de i.d.d. es el análogo de estar en posición general.

- Otra importante aplicación de los teoremas de concentración a la Ciencia de Datos es una modificación de la selección del modelo de Akaike, es decir un tipo de regularización.

3 Análisis topológico de datos y anomalías

La topología es el estudio matemático de las figuras geométricas desde la lupa de sus características más puras. Por lo anterior, el estudio topológico requiere definir aquello que significa pureza, normalmente lo hace vía la abstracción de aquellas propiedades fundamentales para la totalidad de las figuras geométricas.

Pensemos en un símil, en Estadística y Ciencia de Datos desde un punto de vista Bayesiano por ejemplo, este tipo de procedimientos son bastante usuales: supongamos que en lugar de estudiar figuras geométricas buscamos estudiar variables aleatorias (tal como lo hace el análisis Bayesiano para las distribuciones que rigen ciertas bases de datos). Un primer acercamiento para distinguir variables aleatorias es comparar sus momentos de distíntos órdenes, es decir: si dos variables aleatorias tienen distintas esperanzas entonces no hay forma de que sean la misma, por otro lado pueden existir dos variables aleatorias distintas con la misma esperanza, quizás su varianza nos ayude a diferenciar entre una y otra sin embargo nuevamente es fácil encontrar encontrar dos variables aleatorias diferentes con la misma esperanza y la misma varianza. Así sucesivamente.

Exercise 3.1. *En un conjunto finito, encontrar dos variables aleatorias cuyas esperanzas y varianzas sean iguales, sin embargo ellas no sean iguales.*

La topología hace algo semejante con las figuras geométricas: mediante las llamadas "imposibilidades" topológicas encuentra condiciones necesarias para que dos figuras geométricas sean la misma. Concentrémonos en un ejemplo: supongamos que viajamos en una sola dirección a lo largo de una figura uno-dimensional, buscamos la manera de distinguir dos casos fundamentales que podrían ocurrir, el primero es cuando viajamos a lo largo de una línea y le segundo es cuando viajamos a lo largo de un círculo. Para agregar una complicación al problema pensemos que ambos son inmensamente grandes en comparación a la velocidad como nos desplazamos a lo largo de esta figura, en ese caso el problema se vuelve muy complicado; un fenómeno similar a esta complicación ocurre cuando buscamos la manera de demostrar que el planeta tierra no es plano utilizando estrictamente la geometría que nos rodea.

El ejemplo anterior es un síntoma constante en el estudio topológico de la geometría, por tanto los matemáticos han tenido la necesidad de recurrir a métodos más elaborados para resolver este problema

de clasificación. Uno de los métodos más exitosos para este fin es la cohomología (y la homología), la cuál será el centro de la primera parte de esta clase.

En la segunda parte del curso nos concentraremos en el estudio de anomalías en Ciencia de Datos. Tanto la identificación como el proceso de ignorar las anomalías eficientemente es uno de los problemas más complicados en Ciencia de Datos. Existen algunos acercamientos para atacar este problema mediante el llamado Análisis Topológico de Datos que consiste en estudiar la topología probable de una base de Datos. Es MUY IMPORTANTE mencionar que en la actualidad los métodos para el tratamiento de anomalías vía TDA no son más exitosos que otros basados en estudios estadísticos por tanto su eficacia en problemas concretos se presentará como una conjetura optimista.

3.1 Primera parte: Cohomología

En la primera sección estudiaremos con detalle el uso y las interpretaciones cohomológicas de dos problemas sencillos. El primer caso que se estudiará no hace ninguna mención a la topología ni a algún objeto geométrico, sin embargo lo incluimos como una motivación y caso de estudio. El segundo caso sí hace referencia a objetos geométricos sin embargo obviaremos las definiciones topológicas formales.

3.1.1 Cohomología de la suma

Supongamos que debido a restricciones de memoria solo podemos hacer sumas que involucren a lo más dos dígitos, esto significa dos cosas:

- Cuando sumemos dos números ambos pueden tener a lo más dos cifras.
- El resultado de nuestras sumas solo constará de dos cifras.

Exercise 3.2. *Sea x un número entero entre el 0 y el 100. Convencerse de que es posible escribirlo de manera única de la siguiente forma:*

$$x = a + b \cdot 10$$

cuando $0 \leq a, b \leq 9$.

Hint: es necesario utilizar el algoritmo de Euclides.

Gracias al ejercicio anterior es posible "codificar" perfectamente la información de un número $0 \leq x \leq 100$ mediante los enteros $0 \leq a, b \leq 9$.

Ahora regresemos al problema de sumar números $0 \leq x, y \leq 99$, ¿cómo podemos utilizar los códigos proporcionados por el ejercicio anterior para calcular la suma $x + y$? Un primer problema obvio que enfrentamos es el siguiente: existe números $0 \leq x, y \leq 99$ tales que $x + y$ tiene más de dos cifras.

- A partir de este momento solo ignoraremos el tercer dígito de las sumas que hagamos.

Es importante notar que esta hipótesis esconde una simplificación pedagógica profunda, ¿qué pasa si en lugar de considerar números de a lo más 2 dígitos, hubiéramos declarado relevantes para nuestro estudio aquellos números menores o iguales a 11? En ese caso la complicación obvia es la siguiente: ¿cómo representamos aquellos números iguales a la suma de dos $0 \leq x, y \leq 11$ tales que su suma no es menor o igual a 11? La respuesta a este problema nuevamente se encuentra en el ya mencionado algoritmo de Euclides, a saber en el residuo. Por ejemplo pensemos que queremos sumar 6 y 8, la afirmación es que debemos representar a $6 + 8 = 14$ por el número $3 \leq 11$. La explicación de esta representación se encuentra en el siguiente problema: una persona comienza su día laboral de ocho horas a las seis de la mañana, ¿qué número marcará un reloj de manecillas cuando él haya terminado de trabajar? Quizás para números menores o iguales a 11 nuestra intuición proveniente del uso de relojes de manecillas nos ayuda, esta intuición puede ser más complicada para otros sistemas para representar números, afortunadamente no lo es cuando representamos números con solo dos cifras pues nuestro sistema es decimal.

Ahora continuemos con el problema de sumar dígitos x, y menores a iguales a 100. Gracias al ejercicio anterior:

$$x = a_1 + b_1 \cdot 10, y = a_2 + b_2 \cdot 10$$

Para ciertos (y únicos!) $0 \leq a_1, a_2, b_1, b_2 \leq 9$. Buscamos conocer los últimos dos dígitos de $x + y$, llamémosle a partir de ahora a esos dos dígitos $\overline{x + y}$. De nuevo utilizando el ejercicio anterior sabemos que $\overline{x + y} = a_3 + b_3 \cdot 10$ para ciertos $0 \leq a_3, b_3 \leq 9$.

En este momento una observación fundamental es que en la representación $x = a + b \cdot 10$, a pesar de que tanto a y b son números entre 0 y 9, sus roles no son semejantes. Un ejemplo de esto es precisamente el estudio de la suma: la relación que existe por un lado entre a_3 y a_1, a_2, b_1, b_3 y por otro entre b_3 y a_1, a_2, b_1, b_3 es dispar.

Exercise 3.3. *Convencerse de que a_3 depende únicamente de a_1 y a_2 sin embargo sin embargo b_3 depende de a_1, a_2, b_1, b_3. Un ejercicio más complicado aunque muy interesante es formalizar el enunciado anterior. Hint: calcular las covarianzas de ciertas variables aleatorias.*

De hecho calcular a_3 es súmamente fácil en términos de a_1 y a_2, solo repitamos aquello que sabemos hacer para calcular la suma del tiempo en un reloj de manecillas o (de hecho es lo mismo) realicemos el algoritmo de Euclides.

La parte no trivial es encontrar la regla que define a la función que determina a b_3. El matemático Arnold tiene interesantísimas conjeturas sobre las propiedades estadísticas de estas funciones. A partir de ahora denotaremos a esta función misteriosa como F', formalmente $F' : \{0, 1, 2, \ldots, 9\}^4 \to \{0, 1, 2, \ldots, 9\}$ y $F'(a_1, a_2, b_1, b_2) = b_3$ suponiendo lo anterior sobre esos números y utilizando la misma notación.

En realidad hasta el momento hemos sido demasiado pesimistas sobre la función F', decir que no conocemos nada sobre ella es falso pues la suma tiene más estructura de lo que supusimos. No es difícil convencerse de lo siguiente:

$$b_3 = F'(a_1, a_2, b_1, b_2) = b_1 + b_2 + F(a_1, a_2)$$

Exercise 3.4. *Demostrar formalmente la afirmación anterior. Hint: Algoritmo de Euclides (de nuevo!).*

Por tanto en realidad solo vamos a estar interesados en la función $F : \{0, 1, 2, \ldots, 9\}^2 \to$.

De nuevo será necesario agregar un poco más de estructura a nuestro problema, esta vez recordaremos un enunciado fundamental llamado asociatividad:

Proposition 3.5. *Utilizando las notaciones anteriores es posible demostrar que si $0 \leq x, y, z \leq 99$ entonces*

$$\overline{(\overline{x + y}) + z} = \overline{x + (\overline{y + z})}$$

Proof. Sugerimos al lector comenzar con el siguiente caso $0 \leq x, y, z \leq 1$. □

Otra observación importante es que la proposición anterior es cierta también para vectores! Más adelante vamos a hacer uso de esa importante observación.

Exercise 3.6. *Demostrar la siguiente ecuación funcional que satisface F:*

$$F(a_2, a_3) - F(a_1 + a_2, a_3) + F(a_1, a_2 + a_3) - F(a_1, a_2) = 0$$

Hint: utilizar la proposición anterior.

Exercise 3.7. *Demostrar que $F(a, 0) = 0 = F(0, a)$.*

La tarea fundamental de la cohomología de la suma es el estudio de aquellas funciones F que satisfacen la conclusión de los dos ejercicios anteriores.

Definition 3.1. Sea $F : \{0,1,2,\ldots,9\}^2 \to \{0,1,2,\ldots,9\}$, decimos que F es un co-ciclo cuando:

- $F(a,0) = 0 = F(0,a)$ y
- $F(b,c) - F(a+b,c) + F(a,b+c) - F(a,b) = 0$ para cualesquiera a, b, c.

Es entendible que haber introducido la definición anterior en plural y por tanto sugerir la existencia de distintas funciones co-ciclos genere confusión por la manera como desarrollamos la explicación hasta el momento. Parecía que lo que queríamos es adivinar a la función F sin embargo ahora estamos permitiendo la existencia de distintas funciones F. Lo anterior no debe permitir que lleguemos a la siguiente conclusión falsa: el valor de b_3 no está únicamente determinado por b_1, b_2 y $F(a_1, a_3)$.

Para entender correctamente la naturaleza de los co-ciclos es momento de hacer una analogía con el problema de identificar si estamos recorriendo una línea o un círculo, recordemos que este problema puede ser complicado, la cohomoloía topológica nos permitirá explicar en cuál figura caminamos pues distintos co-ciclos corresponderán a distintas figuras geométricas.

La pregunta natural en este momento es la siguiente: en qué otro tipo de "figuras" podríamos estar sumando números que a su vez correspondan con distintos co-ciclos? La respuesta son los espacios vectoriales:

Definition 3.2. Definimos el espacio vectorial $V_{10,2}$ de la siguiente manera:

- $V_{10,2}$ consiste en aquellos vectores con dos entradas $v = (\alpha, \beta)$ tales que $0 \leq \alpha, \beta \leq 9$.
- Definimos la suma de dos vectores $v = (\alpha_1, \beta_1), w(\alpha_2, \beta_2)$ de la siguiente forma: $v + w = (\alpha, \beta)$ donde α es la unidad del número $\alpha_1 + \alpha_2$ y β es la unidad del número $\beta_1 + \beta_2$.
- Para hablar de espacio vectorial es necesario definir el producto con escalares, sin embargo para fines de estas notas no lo necesitamos.

Exercise 3.8. *Demostrar que $V_{10,2}$ tiene 100 vectores.*

El ejercicio anterior en realidad es una provocación pues aunque no lo hemos dicho con claridad a lo largo de esta sección hemos estado estudiando la estructura artimética de la suma en un conjunto de 100-elementos.

Definition 3.3. Llamaremos $\mathbb{Z}_{100}$ al conjunto de los números $0 \leq x \leq 99$ junto a la estructura de suma usual bajo la siguiente regla:

- Solo consideraremos los últimos dos dígitos de la suma $x+y$. Recuerden que esto es lo que hemos estado abreviando como $\overline{x+y}$.

La pregunta fundamental ahora es la siguiente: son $V_{10,2}$ y $\mathbb{Z}_{100}$ iguales (isomorfos)?
Comencemos con un caso mucho más sencillo:

Proposition 3.9. *$\mathbb{Z}_4$ y $\mathbb{Z}_2^2$ no son isomorfos.*

Proof. Es suficiente con notar lo siguiente, en $\mathbb{Z}_4$ existe un elemento x tal que $x + x \neq 0$ sin embargo en $\mathbb{Z}_2^2$ cualquier elemento satisface $x + x = 0$. □

Exercise 3.10. *Utilizando la idea de la proposición anterior demostrar que $V_{10,2}$ y $\mathbb{Z}_{100}$ no iguales.*

Aunque tanto la proposición y el ejercicio anterior sugieren que es muy simple diferenciar entre ambos "espacios" en realidad este problema es súmamente complejo cuando su tamaño es inmenso. Piensen por ejemplo en la maldición de la dimensión o en un zoom muy grande a la línea y al círculo. La cohomología es capaz de resolver clase de problemas utilizando el concepto de co-cliclos y el siguiente resultado:

Theorem 3.11. *Sea F el co-ciclo asociado a la suma de dos elementos en $\mathbb{Z}_{100}$ y G el co-ciclo asociado a la suma de dos elementos en $V_{10,2}$, si ellos son distintos entonces ambos "espacios" no pueden ser isomorfos.*

La demostración formal (e incluso su formulación exacta) de este resultado está fura del alcance de nuestra clase. Es fundamental entenderlo como el análogo para variables aleatorias de lo siguiente: si dos variables aleatorias tienen distintas esperanzas entonces sus leyes de probabilidad deben de ser distintas.

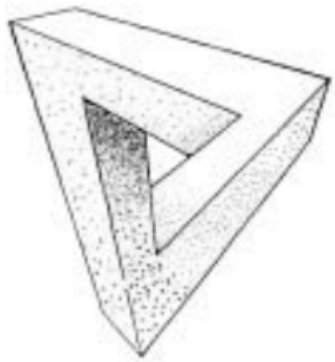

Figure 1: Triángulo de Penrose

Uno de los detalles técnicos que estamos obviando hasta ahora es la importancia del (llamado así por Sir. R. Penrose) grupo de ambigüedad el cuál introduciremos en la siguiente sección.

3.1.2 Cohomoloía y la obra de Escher (después de Penrose)

En la sección anterior explicamos cómo la idea de co-ciclo podría ayudar a diferenciar un espacio 2-dimensional de otro 1-dimensional, en esta sección vamos a utilizar esas ideas para demostrar la imposibilidad de construir en el espacio 3-dimensional (en la vida real) un dibujo (es decir una figura 2-dimensional).

Consideremos el dibujo de la figura 1, llamada Triángulo de Penrose. Esta figura está profundamente inspirada por la obra del artista Escher la cuál no tiene desperdicio.

La afirmación fundamental de esta sección es la siguiente: no existe una figura tres dimensional consistente con aquello que vemos en el trángulo de Penrose. Para comenzar la demostración de esa imposibilidad es necesario partir el dibujo en tres regiones fundamentales Q_1, Q_2, Q_3 como en la figura 2.

A partir de ahora supondremos por contradicción que la figura sí existe en tres dimensiones. Eso implica que también las tres figuras tres dimensionales correspondientes existen. Para continuar podemos pensar en la cohomología como armar un rompecabezas:

- juntar los sólidos Q_i y Q_j en los puntos $A_{i,j}$ de la figura 2, otra forma de decir lo anterior es: juntar los puntos $A_{i,j}$ y $A_{j,i}$.

Antes de comenzar el estudio cohomológico es necesario hacer un repaso de un concepto geométrico fundamental en matemáticas: la simetría.

La idea intuitiva de simetría quizás es clara con la figura 3: los tres sólidos dibujados son simétricos. Es importante notar que esta simetría es un poco más complicada que aquellas a las que estamos acostumbrados, quizás nos gustaría pensar en figuras que crecen de volumen de manera proporcional. Esto no es suficiente para los fines de este problema pues nos interesa considerar la ambigüedad de

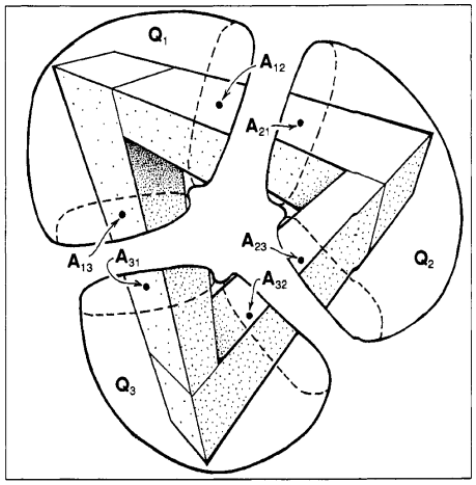

Figure 2: Partición del triángulo de Penrose

poder escalar las piezas Q_1, Q_2, Q_3 del rompecabezas tanto como sea necesario siempre y cuando se puedan "armar" en el espacio tres dimensional.

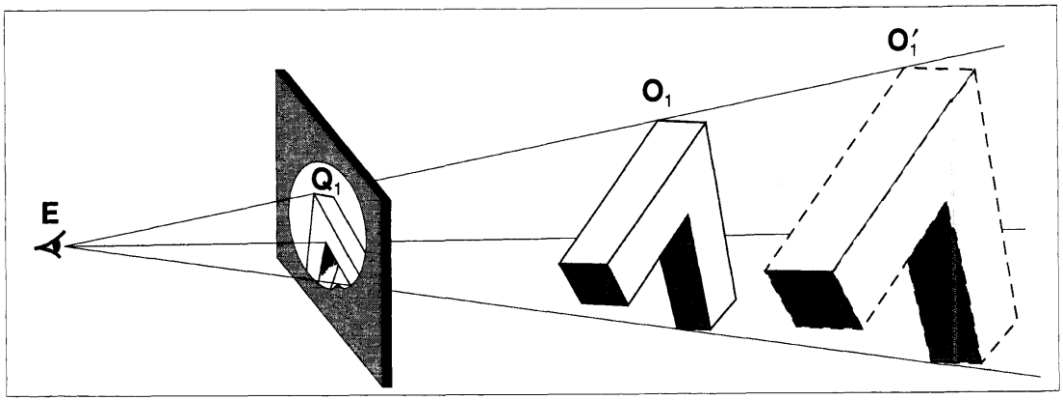

Figure 3: Partición del triángulo de Penrose

Para definir el co-ciclo adecuado es necesario estudiar las distancias entre un observador imaginario y nuestro rompecabezas, para ser más precisos nos interesan las distancias $d_{i,j}$ entre un observador O y los puntos $A_{i,j}$ como en la figura 4. En este lenguaje, armar el rompecabezas significa:

- $d_{i,j} = d_{j,i}$

Ahora definamos $r_{i,j} = \frac{d_{i,j}}{d_{j,i}}$. Es claro que $r_{i,j} = \frac{1}{r_{j,i}}$. Considerar $r_{i,j}$ en lugar de $d_{i,j}$ es lo análogo a la ecuación del ejercicio 3.7.

En este lenguaje armar el rompecabezas significa lo siguiente:

- $F = (r_{1,2}, r_{2,3}, r_{1,3}) = (1, 1, 1)$.

Este elemento F es en realidad un co-ciclo, y recibirá el nombre de co-frontera cuando $F = (1, 1, 1)$. Es decir cuando la figura sea realizable en el espacio.

Theorem 3.12. *El co-ciclo F asociado al triángulo de Penrose nunca es igual a* $(1, 1, 1)$.

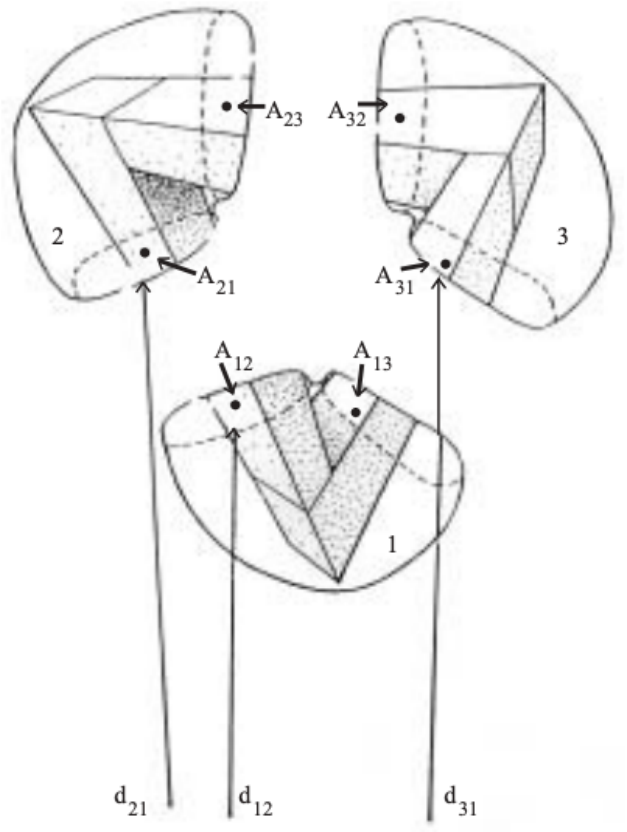

Figure 4: Partición del triángulo de Penrose

3.2 Segunda parte: anomalías

Las técnicas para estudiar anomalías fundamentalmente se dividen en dos: por un lado están aquellos algoritmos que buscan modelar y predecir la presencia de anomaías y por otro lado están aquellos algoritmos que buscan ignorarlas, pues obedecerlas puede ocasionar que caigamos en errores.

Para este segundo caso pensemos por ejemplo en el método de mínimos cuadrados: supongamos que tenemos una base de datos $S\{(x_i, y_i)\}_{i\leq N}$ tal que $x_i \in \mathbb{R}^d$ e $y_i \in \mathbb{R}$, ahora pensemos que buscamos atacar este problema mediante una regresión (digamos lineal para fijar ideas), entonces el problema fundamental es buscar cierto $\beta^* \in \mathbb{R}^d$ de tal forma que $f(x_i) := x_i \cdot \beta^* \sim y_i$. El método de mínimos cuadrados consiste en buscar algún β^* de tal forma que en promedio sobre S, la cantidad $(\beta^* x_i - y_i)^2$ sea lo menor posible. Ahora supongamos que cierto (x_j, y_j) es una anomalía para el modelo lineal, eso significa que para cualquier $\beta \in \mathbb{R}^d$, la distancia $d(\beta \cdot x_j, y_j)$ es demasiado grande. Si la distancia anterior ya es demasiado grande qué pasa con $(\beta \cdot x_j - y_j)^2$?

Una característica importante sobre las anomalías es su virtual imposibilidad de definirlas con alto grado de generalidad. La inteción de esta sección es buscar una relación entre el concepto de anomalía y la topología de un modelo.

Por otro lado hasta ahora hemos hablado de la cohomología como una herramienta para estudiar las propiedades geométricas de un objeto. En análisis topológico de datos se utiliza la homología. La diferencia entre ambos conceptos está fuera de los objetivos de este curso. Gracias a profundos resultados de dualidad y comparación, en muchos casos es posible hablar de cohomología y homología de manera indistinta, cosa que haremos en este curso.

3.2.1 Métodos estadísticos para la detección de anomalías

Comencemos recordando el primer teorema de concentración, el lema 2.8 de Chevyshev nos dice que si X es una variable aleatoria con valores en $\mathbb{R}$ entonces para cualquier $t > 0$ tenemos:

$$\mathbb{P}\left(|X - \mu| > t\right) < \frac{\sigma^2}{t^2}$$

Eso significa por ejemplo que si t^2 es demasiado grande respecto a la varianza σ^2 entonces la probabilidad de que X esté demasiado separado de la esperanza podría ser grande, eso es precisamente la idea de una anomaía y podríamos pensarlo como nuestra definición probabilista. Es importante mencionar que no todas las anomalías deben ser de este tipo, es perfectamente razonable pensar que un valor anómalo (por ejemplo en una encuesta) no proviene de la misma distribución de donde viene el resto de valores no-anómalos. Para estudiar este comportamiento se definen distancias entre leyes de probabilidad.

Ahora vayamos al problema de la detección de anomalías, existen diferentes formas de detectar anomalías utilizando la definición anterior, vale la pena mencionar aquellos tests estadísticos que definen la hipótesis nula como la ausencia de una anomalía, después comparando entre lo observado y lo supuesto por H_0 buscan llegar a contradicciones con cierto grado de confianza, la mayoría de estos métodos tienen su justificación matemática en los Teoremas de los Grandes números y en el Teorema del Límite Central.

Otro método estadístico interesante y por otro lado profundamente relacionado con la reducción de la dimensión vía PCA o el lema 2.13 es el llamado Determinante Mínimo de la Covarianza.

Supongamos que tenemos un conjunto $S\{x_i\}_{i\leq N} \subseteq \mathbb{R}^d$ de puntos, por un momento vamos a pensar en un problema de aprendizaje no-supervisado. Nuestra intuición nos dice que en este sentido no-supervisado un problema de detección de anomalías es muy parecido a un problema de clusterización i.e. buscamos aquellos puntos en S que estén de alguna manera aislados con respecto a la mayor parte de S.

El mé todo del Determinante Mínimo de la Covarianza consiste en encontrar un subconjunto $S' \subset S$ de tamaño h de tal forma que el determinante de la matriz de covarianza de S' sea lo menor posible respecto a todos los subconjuntos de tamaño h de S. Varios comentarios en este momento son pertinentes:

- La cantidad h es una escala, nos dice cuál será la definición de outlier, de hecho es posible hablar de h-outliers en este sentido.

- Supongamos que nuestros puntos S siguen un aley de distribución normal. Sea h fija e ideal para nuestro problema, más adelante hablaremos del concepto de entropía, por el momento solo diremos que la entropía de un conjunto S es la cantidad de información que contiene. A mayor entropía será más difícil comunicar mediante un mensaje corto aquello que está en S, por otro lado si la entropía es pequeña significa que la cantidad de información de S también lo es. Dicho lo anterior, podemos pensar en una anomalía como un elemento que agrega sustancialmente mayor información a nuestro conjunto S. Es posible demostrar que bajo la hipótesis de normalidad hecha en este inciso, que calcular el determinante de la matriz de covarianza de S es muy cercano a calcular la entropía de S (la entroía diferencial para ser precisos), por eso es que buscamos minimizar esa (ambas) cantidad.

- Existen dos complicaciones fundamentales para la correcta implementación del método del DMC. La primera es elegir correctamente a h. Afortunadamente existen buenas razones para elegir a $h = \frac{(N+d+1)}{2}$ (o el entero más cercano), una de ellas es la Maldición de la dimensión. La segunda complicación de este método es que minimizar el determinante de la matriz de covarianza puede ser (y en muchos casos lo será) un problema no convexo de optimización y por tanto demasiado difícil de llevar a cabo.

3.2.2 Análisis topológico de Datos

El análisis topológico de datos busca encontrar ciclos (esta vez no serán co-ciclos) en la representación geométrica de una base de datos. Ello supone una hipótesis fundamental sobre las bases de datos $S = \{(x_i, y_i)_{i \leq N}$ (supongamos por simplicidad que $x_i \in \mathbb{R}, y_i \in \mathbb{R}$): existe una función f de tal forma que $f(x_i) = y_i$. El espacio topológico que buscamos estudiar es la gráfica de f, $\Gamma(f) \subseteq \mathbb{R}^2$, en general estos espacios serán parecidos a una curva.

Los ciclos en este contexto se definen de la siguiente manera: supongamos que tenemos dos puntos $p = (x, f(x)), q = (y, f(y)) \in \Gamma(f)$. Lo que buscamos saber sobre la curva $\Gamma(f)$ es si ella es conexa, es decir si es posible viajar de manera continua entre todo par de puntos dentro de ella. Si eso fuera posible entonces habría una identificación entre la curva que une a p y a q con el intervalo (x, y). Una manera como ello no sería posible es cuando los puntos entre p y q se identifican con dos intervalos disconexos $(x, z_1), (z_2, y)$ con $z_1 < z_2$. La diferencia combinatoria entre ambos casos es lo siguiente: $x - y = 0$ si consideramos a p y q suficientemente cercanos sin embargo $x - z_1 + z_2 - y \neq$ sin importar qué tan cerca estén p y q pues $z_1 - z_2 > 0$ siempre bajo la hipótesis de disconexidad.

Lo que describimos anteriormente es lo que se conoce como homología del espacio $\Gamma(f)$, sin embargo recordemos que en un problema de Ciencia de Datos uno no puede suponer que tiene acceso a la función f y por tanto es imposible estudiar a $\Gamma(f)$. La idea especial de TDA es construir una filtración de espacios topológicos Γ_n correspondientes a subconjuntos de tamaño n, $S_n \subseteq S$. Los ciclos se definirán como ciclos que persisten en los espacios topológicos Γ_n a medida que n crece.

Una última observación para concluir esta sección es que esta idea de ciclo es contradictoria al estudio de outliers pues es difícil imaginar outliers cuya relevancia en el espacio topológico persista a medida que consideramos muestreos más grandes.